SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 1, January 1983

0363-0129/83/2101-0001 $01.25/0

STRICTLY OBSERVABLE LINEAR SYSTEMS*

JACOB HAMMER† AND MICHAEL HEYMANN‡

Abstract. A theory of strictly observable linear systems is developed in a module theoretic framework which is consistent with the classical algebraic theory of linear time invariant realization. The theory incorporates in a unified framework the reduction of linear systems through precompensation, through state feedback, and through dynamic output feedback.

1. Introduction. In Hautus and Heymann [1978] and in Hammer and Heymann [1981], the foundations for an algebraic theory of linear systems were formulated, using the linear realization theory of Kalman [1965] (see also Kalman, Falb, and Arbib [1969], Chapt. 10) as the starting point. In Hautus and Heymann [1978] emphasis has been placed on the input/state behavior and on static state feedback using the theory of $K[z]$-modules ($K[z]$ being the ring of polynomials in z over a field K). In Hammer and Heymann [1981] the theory has been extended to investigate the structure of dynamic as well as static output feedback. It has been shown there that an important role in the theory of output feedback is played by the *latency* structure and the *latency kernel* of the system. The latency structure is characterized by the class of system inputs whose corresponding outputs are identically zero prior to the time $t=0$. This structure is algebraically expressed by modules over the ring $K[[z^{-1}]]$ of power series (in z^{-1} over the field K) and led to a rich structure theory as evidenced in Hammer and Heymann [1981].

In the present paper we focus on a "dual" class of inputs, namely, those that generate outputs terminating at $t=0$. This leads to a $K[z]$-module structure and in particular to the concept of *strict observability* which is the main theme of the paper. The basic definition of strict observability in our framework is that in the above mentioned class of inputs all elements are polynomial (i.e., terminating at or before $t=0$).

The concept of strict observability is closely related to various concepts that have been studied (from various different points of view) in the literature. Probably the first time the concept appeared was in Basile and Marro [1969] and in the paper by Nikolskii [1970] who defined a linear system to be *ideally observable* if its state can be observed from knowledge of the output alone (without knowledge of the corresponding input). Nikolskii showed that ideal observability holds if and only if the observability is maintained under every static state feedback law. The same concept was introduced independently in Rappaport and Silverman [1971], who called it *perfect observability* (see also Payne and Silverman [1973]). In Heymann [1972] the concept of *feedback irreducibility* was introduced and a system was called *feedback irreducible* if its observability is invariant (i.e., indestructible) under state feedback. Irreducibility was also studied in Morse [1975], where a system was defined to be *irreducible* if the subspace v^*, i.e., the largest (A, B)-invariant subspace in the kernel of C, is zero

* Received by the editor February 24, 1981, and in revised form December 18, 1981.

† Center for Mathematical System Theory, University of Florida, Gainesville, Florida 32611. The work of this author was done while he was with the Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel, and supported in part by the U.S. Army Research Office grant DAAG29-80-C0050, and the U.S. Air Force under grant AFOSR76-3034D through the Center for Mathematical System Theory, University of Florida, Gainesville.

‡ Department of Electrical Engineering, Technion, Israel Institute of Technology, Haifa, Israel. The research of this author was supported in part by the Technion Fund for Promotion of Research.

(see also Morse [1973]). Morse showed that his definition of irreducibility is equivalent to feedback irreducibility and obtained various other results on irreducible systems. The equivalence of irreducibility and strong observability was shown in Molinary [1976], and more recently the equivalence between the various concepts mentioned above was discussed in Hautus [1979], where also an important characterizing rank condition was given. Other recent related papers are Fuhrmann and Willems [1979], [1981] and Khargonekar and Emre [1980].

In this paper we study the effects of bicausal precompensation (i.e., cascade control) and of state as well as output feedback on the structural properties of linear systems, with strict observability playing a central role in the theory.

In § 2 we give some mathematical preliminaries reviewing the algebraic setup. In § 3 strict observability is formally defined and some basic consequences that follow immediately from the definition are discussed. In particular, the structure of injective precompensation orbits is investigated. Theorem 3.2 states that every injective input-output (i/o) map can be made strictly observable by bicausal precompensation, and Theorem 3.3 states the fact that an injective i/o map can be rendered strictly observable, also, by static state feedback in *every possible* realization. Theorem 3.4 summarizes the properties of strictly observable i/o maps. In § 4 the structure of bounded Ω^+K-modules is discussed. In § 5 the structure of precompensation orbits of injective i/o maps is investigated in detail. Reduced reachability indices are defined. In Theorem 5.1 a Wiener–Hopf type factorization is proved for injective i/o maps. A characterization of injective precompensation orbits based on the reduced reachability indices is given in Theorem 5.3, and a "dynamics assignment" theorem (by precompensators) is given in Theorem 5.5. Section 6 is devoted to an investigation of the effect of dynamic output feedback. Theorem 6.5 states that an injective i/o map can be made strictly observable, also, by output feedback. Theorem 6.6 gives an "index assignment" result, i.e., it states that by output feedback any admissible set of reachability indices can be attained for systems defined over infinite fields. In § 7 contact is made between the present theory and the geometric control theory of Wonham and Morse and supremal (A, B)-invariant subspaces in $\ker C$ are characterized. Finally, in § 8 generalization to noninjective i/o maps is discussed.

2. Preliminaries on the mathematical setup. The reader is assumed to be familiar with the mathematical setup and terminology of Hautus and Heymann [1978] as well as Hammer and Heymann [1981], which we review briefly.

For a field K and a K-linear space S, we denote by ΛS the set of all formal Laurent series in z^{-1} of the form

$$s = \sum_{t=t_0}^{\infty} s_t z^{-t}, \qquad s_t \in S. \tag{2.1}$$

It can then be seen that, with coefficientwise addition and convolution multiplication, the set ΛK forms a field, and under similar operations the set ΛS becomes a ΛK-linear space. When S is finite dimensional, then so is also ΛS (as a ΛK-linear space) and $\dim_{\Lambda K} \Lambda S = \dim_K S$.

The set ΛS contains as subsets the set Ω^+S of (polynomial) elements of the form $\sum_{t\leqq 0} s_t z^{-t}$, and the set Ω^-S of (power series) elements of the form $\sum_{t\geqq 0} s_t z^{-t}$. In particular, Ω^+K and Ω^-K form principal ideal domains under the operations defined in ΛK. Furthermore, Ω^+S and Ω^-S are free modules over Ω^+K and Ω^-K, respectively, and in case the K-linear space S is finite dimensional, both of these modules are of rank equal to $\dim_K S$.

Let U and Y be K-linear spaces. A ΛK-linear map $\bar{f}: \Lambda U \to \Lambda Y$ represents a linear time invariant system with *input value space* U and *output value space* Y (Wyman [1972]). The *order* of the ΛK-linear map $\bar{f}$ is defined as $\operatorname{ord} \bar{f} := \inf \{\operatorname{ord} \bar{f}u - \operatorname{ord} u \mid 0 \neq u \in \Lambda U\}$, and, in case U and Y are finite dimensional, $\operatorname{ord} \bar{f} > -\infty$. Below we shall always assume that U and Y are finite dimensional and we denote

$$m := \dim_K U \quad \text{and} \quad p := \dim_K Y.$$

Further, let L denote the K-linear space of K-linear maps $U \to Y$. With every ΛK-linear map $\bar{f}: \Lambda U \to \Lambda Y$ one associates an element $\mathcal{T}_f = \sum T_t z^{-t}$ in ΛL, called the *transfer function* of $\bar{f}$. The coefficients T_k of the transfer function are given by $T_k := p_k \cdot \bar{f} \cdot i_u$, where the K-linear maps p_k and i_u are defined as

$$i_u : U \to \Lambda U : u \mapsto u \qquad \text{(injection)}$$

$$p_k : \Lambda Y \to Y : \sum y_t z^{-t} \mapsto y_k.$$

It can then be readily seen that the action of $\bar{f}$ on an element $u = \sum u_t z^{-t} \in \Lambda U$ is given by the convolution formula

$$\bar{f}u = \sum_t \left(\sum_k T_k U_{t-k} \right) z^{-t}. \tag{2.2}$$

For conciseness, we shall frequently identify ΛK-linear maps with their transfer functions.

Next, we define some terminology. Let s be an element in ΛS. Then, s is called (i) *polynomial* if $s \in \Omega^+ S$, (ii) *strictly polynomial* if $s \in z\Omega^+ S$, (iii) *causal* if $s \in \Omega^- S$, (iv) *strictly causal* if $s \in z^{-1}\Omega^- S$, (v) *static* if $s \in S$, and (vi) *rational* if there exists a nonzero polynomial $\psi \in \Omega^+ K$ such that ψs is polynomial. We denote by $\Lambda_r S$ the set of all rational elements in ΛS, so that $\Lambda_r K$ is the field of polynomial quotients, and $\Lambda_r S$ is a $\Lambda_r K$-linear space.

The above terminology also applies to ΛK-linear maps $\bar{f}: \Lambda U \to \Lambda Y$ through the respective properties of their transfer functions as elements of ΛL. Upon applying the convolution formula (2.2), it is easy to verify that $\bar{f}$ is: (i) polynomial if and only if $\bar{f}[\Omega^+ U] \subset \Omega^+ Y$, (ii) strictly polynomial if and only if $\bar{f}[\Omega^+ U] \subset z\Omega^+ Y$, (iii) causal if and only if $\bar{f}[\Omega^- U] \subset \Omega^- Y$, (iv) strictly causal if and only if $\bar{f}[\Omega^- U] \subset z^{-1}\Omega^- Y$, (v) static if and only if $\bar{f}[U] \subset Y$, and (vi) rational if and only if $\bar{f}[\Lambda_r U] \subset \Lambda_r Y$. A strictly causal and rational ΛK-linear map $\bar{f}: \Lambda U \to \Lambda Y$ is called a *linear i/o (input-output) map*. A ΛK-linear map $\bar{l}: \Lambda U \to \Lambda U$ is called *bicausal* if it is causal and has a causal inverse.

We associate with a linear i/o map $\bar{f}$ a number of related constructs. First, we define the two $\Omega^+ K$-homomorphisms

$$j^+ : \Omega^+ U \to \Lambda U \qquad \text{(natural injection)},$$

$$\pi^+ : \Lambda Y \to \Lambda Y / \Omega^+ Y (=: \Gamma^+ Y) \qquad \text{(canonical projection)}.$$

Then we associate with every linear i/o map $\bar{f}: \Lambda U \to \Lambda Y$ the $\Omega^+ K$-homomorphism

$$\tilde{f} := \pi^+ \cdot \bar{f} \cdot j^+$$

called the *restricted linear i/o map*. We also associate with $\bar{f}$ its *output value map* defined as

$$f := p_1 \cdot \bar{f} \cdot j^+ : \Omega^+ U \to Y.$$

The output value map gives the output value at (time) $t = 1$, and is, in general, (only) a K-linear map. In certain cases, there exists an Ω^+K-module structure on Y, compatible with its K-linear structure, such that the output value map f is an Ω^+K-homomorphism as well. If this is the case then $\bar{f}$ is called a *linear i/s (input-state) map.*

By a realization of a restricted linear i/o map $\tilde{f}: \Omega^+U \to \Gamma^+Y$, we refer to a triple (X, g, h), where X is an Ω^+K-module, and where $g: \Omega^+U \to X$ and $h: X \to \Gamma^+Y$ are Ω^+K-homomorphisms satisfying $\tilde{f} = h \cdot g$. The module X is called the *state space.* The realization (X, g, h) is called *reachable* if g is surjective and *observable* if h is injective. Clearly, the condition $\tilde{f} = h \cdot g$ implies that $\ker g \subset \ker \tilde{f}$. Conversely, if $\ker g \subset \ker \tilde{f}$, there exists an Ω^+K-homomorphism $h: X \to \Gamma^+Y$ such that (X, g, h) is a realization of $\tilde{f}$.

Given a realization (X, g, h), the map $g: \Omega^+U \to X$ can be viewed as the output response map of a linear i/o map $\bar{g}: \Lambda U \to \Lambda X$, which, in fact, is a linear i/s map. We say that $\bar{g}$ is *reachable* if g is surjective. Finally, if (X, g, h) is a realization of $\bar{f}$, there exists a (static) map $H: X \to Y$ such that $\bar{f} = H \cdot \bar{g}$. The last formula is called a *state representation* of $\bar{f}$.

In the present paper we shall be particularly interested in the following type of Ω^+K-modules. An Ω^+K-submodule $\Delta \subset \Lambda S$ is called *bounded* if there exists an integer $k < \infty$ such that $\operatorname{ord} \delta \leqq k$ for every nonzero element $\delta \in \Delta$. If Δ is a nonzero and bounded module, then the least integer k satisfying this order inequality is called the (*order*) *bound* of Δ. Clearly, Ω^+S itself and all its Ω^+K-submodules are examples of bounded modules. A more detailed examination of the structure of bounded modules is given in § 4 below.

3. Strict observability: Basic properties. Let $\bar{f}: \Lambda U \to \Lambda Y$ be a linear i/o map and, as before, let $\pi^+: \Lambda Y \to \Gamma^+Y$ denote the canonical projection. We introduce the following:

DEFINITION 3.1. A linear i/o map $\bar{f}: \Lambda U \to \Lambda Y$ is called *strictly observable* if $\ker \pi^+ \bar{f} \subset \Omega^+U$.

It follows immediately from the definition that if $\bar{f}$ is strictly observable then $\ker \pi^+ \bar{f}$ is bounded and the only ΛK-linear space contained in it is the null space. Since, obviously, $\ker \bar{f} \subset \ker \pi^+ \bar{f}$, it follows that if $\bar{f}$ is strictly observable then $\bar{f}$ is *injective*, (i.e., $\ker \bar{f} = 0$). In Hammer and Heymann [1981, Lemma 5.11] it was shown that *every injective linear i/s map is strictly observable.*

Let ΛU be a fixed ΛK-linear space and consider the class of all rational bicausal ΛK-linear maps $\Lambda U \to \Lambda U$. Clearly, this class forms a (noncommutative) group under the operation of composition. Under the action of this group (with elements acting as bicausal precompensators), the class of linear i/o maps $\Lambda U \to \Lambda Y$ is partitioned into (mutually exclusive) equivalence classes called (*bicausal*) *precompensation orbits.* We next investigate these orbits.

First observe that if a linear i/o map is injective, then so is every element in its precompensation orbit. Thus, an orbit is either *injective* or *noninjective.* Since, as we have seen, strict observability implies injectivity, it follows that if a precompensation orbit contains strictly observable elements it is injective. The theorem below, the proof of which is postponed to § 5 (see Proof 5.2), states that the converse of the above statement is also true, namely that every injective orbit contains strictly observable elements.

THEOREM 3.2. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective linear i/o map. Then, there exists a bicausal precompensator $\bar{l}: \Lambda U \to \Lambda U$ such that $\bar{f}\bar{l}$ is strictly observable.*

Consider a reachable realization (X, g, h) of a linear i/o map $\bar{f}: \Lambda U \to \Lambda Y$; let $\bar{g}: \Lambda U \to \Lambda X$ be the i/s map associated with g, and let $\bar{f} = H \cdot \bar{g}$ be the corresponding

state representation for $\bar{f}$. Also let $\bar{l}: \Lambda U \to \Lambda U$ be a bicausal precompensator for $\bar{f}$. We say that $\bar{l}$ has a *static state feedback representation* in the realization (X, g, h) if there exists a pair of static ΛK-linear maps $F: \Lambda X \to \Lambda U$ and $G: \Lambda U \to \Lambda U$, with G invertible, such that $\bar{l} = (I + F\bar{g})^{-1} G$.

In Hautus and Heymann [1978 Thm. 5.7], it was shown that $\bar{l}$ has a static state feedback representation in a reachable realization (X, g, h) if and only if $\bar{l}^{-1}[\ker \tilde{g}] \subset \Omega^+ U$.

Suppose now that $\bar{f}: \Lambda U \to \Lambda Y$ is an injective linear i/o map and that $\bar{l}: \Lambda U \to \Lambda U$ is a bicausal precompensator for $\bar{f}$ such that $\bar{f}\bar{l}$ is strictly observable, that is, $\ker \pi^+ \bar{f}\bar{l} \subset \Omega^+ U$. Let (X, g, h) be any reachable realization of $\bar{f}$ and let $\tilde{g}$ be the restricted i/s map associated with g. Then $\ker \tilde{g} = \ker g \subset \ker \tilde{f}$ (the equality following from the i/s property), and it follows that

$$\bar{l}^{-1}[\ker \tilde{g}] \subset \bar{l}^{-1}[\ker \tilde{f}] \subset \bar{l}^{-1}[\ker \pi^+ \bar{f}] = \ker \pi^+ \bar{f}\bar{l} \subset \Omega^+ U.$$

By the previous paragraph, we conclude that $\bar{l}$ has a static state feedback representation over $\bar{g}$ and we just proved:

THEOREM 3.3. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective linear i/o map and let $\bar{l}: \Lambda U \to \Lambda U$ be a bicausal precompensator such that $\bar{f}\bar{l}$ is strictly observable. Then $\bar{l}$ has a static state feedback representation in every reachable realization of $\bar{f}$.*

In Heymann [1972], a transfer matrix was called *feedback irreducible* if under the application of static state feedback in a canonical realization, the resultant closed loop system is necessarily also canonical, that is, the observability property is preserved. We shall see that strict observability is equivalent to feedback irreducibility so that Theorem 3.2 combined with Theorem 3.3 is equivalent to Theorem 6.64 in Heymann [1972].

Let $\bar{f}: \Lambda U \to \Lambda Y$ be a strictly observable linear i/o map and let $\delta(\bar{f})$ denote its McMillan degree. If $\bar{f}'$ is any other i/o map in the bicausal precompensation orbit of $\bar{f}$, then by Theorem 3.3, $\bar{f}$ can be obtained from $\bar{f}'$ by static state feedback in any reachable realization of $\bar{f}'$. It follows, therefore, that $\delta(\bar{f}) \leqq \delta(\bar{f}')$, where $\delta(\bar{f}')$ is the McMillan degree of $\bar{f}'$. Thus, all strictly observable linear i/o maps in a given (injective) bicausal precomposation orbit have the same McMillan degree δ, which is the minimal degree among all McMillan degrees of elements in the orbit. Furthermore, strict observability implies feedback irreducibility. Conversely, suppose that $\bar{f}$ is a feedback irreducible linear i/o map and let $\bar{f} = H \cdot \bar{g}$ be a canonical state representation of $\bar{f}$. By Theorem 3.2, there exists a bicausal precompensator $\bar{l}$ such that $\bar{f}' := \bar{f}\bar{l}$ is strictly observable, and by Theorem 3.3, $\bar{l}$ has a static state feedback representation over $\bar{g}$. It then follows (see Hautus and Heymann [1978], Cor. 5.9) that $\bar{g}' := \bar{g}\bar{l}$ is also a reachable linear i/s map and $\ker \tilde{g} = \bar{l}^{-1}[\ker \tilde{g}]$. From the feedback irreducibility of $\bar{f}$ it follows that the state representation $\bar{f}' = H\bar{g}'$ is also canonical, whence $\ker \tilde{f}' = \ker \tilde{g}'$. Consequently,

$$\begin{aligned} \ker \pi^+ \bar{f} = \bar{l}[\ker \pi^+ \bar{f}\bar{l}] = \bar{l}[\ker \tilde{f}'] = \bar{l}[\ker \tilde{g}'] \\ = \bar{l}\bar{l}^{-1}[\ker \tilde{g}] = \ker \tilde{g} \subset \Omega^+ U, \end{aligned}$$

so that $\bar{f}$ is strictly observable. Our preceding discussion is summarized in the following:

THEOREM 3.4. *Consider the class of linear i/o maps in a fixed injective precomposition orbit. Let δ be the minimal McMillan degree of elements in the orbit. Then the following statements are equivalent:*

(i) *$\bar{f}$ is strictly observable.*

(ii) *$\bar{f}$ is feedback irreducible.*

(iii) *$\bar{f}$ has McMillan degree δ.*

(iv) *Every i/o map $\bar{f}'$ in the orbit can be transformed into $\bar{f}$ by static state feedback in any reachable realization.*

Consider now two linear i/o maps $\bar{f}_1 : \Lambda U \to \Lambda Y$ and $\bar{f}_2 : \Lambda U \to \Lambda W$, and assume that there exists a polynomial map $P : \Lambda Y \to \Lambda W$ such that $\bar{f}_2 = P \cdot \bar{f}_1$. Then if $u \in \ker \pi^+ \bar{f}_1$ (i.e., if $\bar{f}_1(u) \in \Omega^+ Y$) it follows also that $\bar{f}_2(u) = P \cdot \bar{f}_1(u) \in \Omega^+ W$, that is, $u \in \ker \pi^+ \bar{f}_2$. We conclude then, that the existence of a polynomial map P such that $\bar{f}_2 = P \cdot \bar{f}_1$ implies that $\ker \pi^+ \bar{f}_1 \subset \ker \pi^+ \bar{f}_2$. That the converse of the above statement is also true will be shown in the ensuing discussion. First, we need the following auxiliary result (proof omitted), which is a consequence of the Smith–McMillan canonical form theorem:

LEMMA 3.5. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be a rational ΛK-linear map, let $r := \dim_{\Lambda K} \operatorname{Im} \bar{f}$ and let $Y_0 \subset Y$ be any r-dimensional subspace. Then there exists a polynomial unimodular map $M : \Lambda Y \to \Lambda Y$ such that $\operatorname{Im} M \cdot \bar{f} = \Lambda Y_0$.*

We also require the following result (compare Hammer and Heymann [1981, Lem. 5.1]).

LEMMA 3.6. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be a ΛK-linear map. If $\mathcal{R} \subset \ker \pi^+ \bar{f}$ is a ΛK-linear subspace, then $\mathcal{R} \subset \ker \bar{f}$.*

With the above lemmas we can now state and prove the polynomial factorization theorem:

THEOREM 3.7. *Let $\bar{f}_1 : \Lambda U \to \Lambda Y$ and $\bar{f}_2 : \Lambda U \to \Lambda W$ be rational ΛK-linear maps. There exists a polynomial ΛK-linear map $P : \Lambda Y \to \Lambda W$ such that $\bar{f}_2 = P \cdot \bar{f}_1$ if and only if* $\ker \pi^+ \bar{f}_1 \subset \ker \pi^+ \bar{f}_2$.

Proof. That the condition of the theorem is necessary was seen in the discussion preceding Lemma 3.5. To prove sufficiency, assume that $\ker \pi^+ \bar{f}_1 \subset \ker \pi^+ \bar{f}_2$. Let $r := \dim_{\Lambda K} \operatorname{Im} \bar{f}_1$ and let Y_0 be any r-dimensional subspace of Y. By Lemma 3.5 there exists a unimodular polynomial map $M : \Lambda Y \to \Lambda Y$ such that $\operatorname{Im} M \cdot \bar{f}_1 = \Lambda Y_0$. If we denote $\bar{f}_0 := M \bar{f}_1$, it follows immediately from the necessity condition above combined with the fact that both M and M^{-1} are polynomial maps, that $\ker \pi^+ \bar{f}_0 = \ker \pi^+ \bar{f}_1$, whence $\ker \pi^+ \bar{f}_0 \subset \ker \pi^+ \bar{f}_2$. Lemma 3.6 then implies that $\ker \bar{f}_0 \subset \ker \bar{f}_2$ so that there exists a ΛK-linear map $P_0 : \Lambda Y \to \Lambda W$ such that $P_0 \cdot \bar{f}_0 = \bar{f}_2$. Let $Y_1 \subset Y$ be a direct summand of Y_0 in Y, that is, $Y = Y_0 \oplus Y_1$. Also, let $\bar{q} : \Lambda Y \to \Lambda Y$ denote the projection onto ΛY_0 along ΛY_1, i.e., if $y = y_0 + y_1 \in \Lambda Y$ is the decomposition of y into its components $y_0 \in \Lambda Y_0$ and $y_1 \in \Lambda Y_1$, then $\bar{q}(y) = y_0$. We now define the map $P := P_0 \cdot \bar{q} \cdot M$ and for each $u \in \Lambda U$ we have

$$P \cdot \bar{f}_1(u) = P_0 \cdot \bar{q} \cdot M \bar{f}_1(u) = P_0 \bar{q} \bar{f}_0(u) = P_0 \bar{f}_0(u) = \bar{f}_2(u),$$

whence $P \cdot \bar{f}_1 = \bar{f}_2$. To conclude the proof we need to show that P is polynomial, and since by definition M is polynomial, it suffices to prove that so also is $P_0 \cdot \bar{q}$. To this end, we first note that every element $y \in \Omega^+ Y$ decomposes uniquely as $y = y_0 + y_1$ with $y_0 \in \Omega^+ Y_0$ and $y_1 \in \Omega^+ Y_1$. Thus

$$P_0 \bar{q}(y) = P_0(y_0) = P_0 \bar{q}(y_0) = P_0 \bar{q} M \bar{f}_1(u) = P \cdot \bar{f}_1(u) = \bar{f}_2(u)$$

for some $u \in \ker \pi^+ \bar{f}_1$. Since by hypothesis $\ker \pi^+ \bar{f}_1 \subset \ker \pi^+ \bar{f}_2$, it follows that $\bar{f}_2(u) = P_0 \bar{q}(y) \in \Omega^+ W$, and the proof is complete. □

COROLLARY 3.8. *Let $\bar{f}_1, \bar{f}_2 : \Lambda U \to \Lambda Y$ be two rational ΛK-linear maps. There exists a unimodular polynomial map $M : \Lambda Y \to \Lambda Y$ such that $\bar{f}_2 = M \bar{f}_1$ if and only if* $\ker \pi^+ \bar{f}_1 = \ker \pi^+ \bar{f}_2$.

Proof. Necessity follows immediately from Theorem 3.7. To prove sufficiency, assume that $\ker \pi^+ \bar{f}_1 = \ker \pi^+ \bar{f}_2$. Then by Lemma 3.6, $\ker \bar{f}_1 = \ker \bar{f}_2$ so that

$\dim \operatorname{Im} \bar{f}_1 = \dim \operatorname{Im} \bar{f}_2 := r$. Let $Y_0 \subset Y$ be an r-dimensional K-linear subspace and let $M_1, M_2 : \Lambda Y \to \Lambda Y$ be unimodular polynomial maps such that $\operatorname{Im} M_1 \bar{f}_1 = \operatorname{Im} M_2 \bar{f}_2 = \Lambda Y_0$ (see Lemma 3.5). Denoting $\bar{f}_{10} := M_1 \bar{f}$ and $\bar{f}_{20} := M_2 \bar{f}_2$, we obviously also have $\ker \pi^+ \bar{f}_{10} = \ker \pi^+ \bar{f}_{20}$. By Theorem 3.7, there exist then polynomial maps $P_{10}, P_{20} : \Lambda Y \to \Lambda Y$ such that $\bar{f}_{20} = P_{10} \bar{f}_{10}$ and $\bar{f}_{10} = P_{20} \bar{f}_{20}$. Let $Y_1 \subset Y$ be a direct summand of Y_0 in Y and let $\bar{q} : \Lambda Y \to \Lambda Y$ be the projection defined in the proof of Theorem 3.7. Now define the polynomial maps $P_1 = \bar{q}(P_{10} - I)\bar{q} + I$ and $P_2 = \bar{q}(P_{20} - I)\bar{q} + I$ where I is the identity map in ΛY. Clearly then also $\bar{f}_{20} = P_1 \cdot \bar{f}_{10}$ and $\bar{f}_{10} = P_2 \cdot \bar{f}_{20}$, and also $P_2 \cdot P_1 = P_1 \cdot P_2 = I$. (The reader can verify these facts by direct computation.) It follows that P_1 is unimodular and the unimodular map $M := M_2^{-1} P_1 M_1$ satisfies the condition of the corollary. □

We conclude the section with an additional characterization of strict observability.

COROLLARY 3.9. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be a linear i/o map. Then $\bar{f}$ is strictly observable if and only if it has a polynomial left inverse.*

Proof. First observe that if $I : \Lambda U \to \Lambda U$ is the identity map, then $\ker \pi^+ I = \Omega^+ U$. Thus, be definition, $\bar{f}$ is strictly observable if and only if $\ker \pi^+ \bar{f} \subset \ker \pi^+ I$. By Theorem 3.7 this kernel inclusion holds if and only if there is a polynomial map $P : \Lambda Y \to \Lambda U$ such that $P \cdot \bar{f} = I$, concluding the proof. □

4. Bounded Ω^+K-modules. Let $f : \Lambda U \to \Lambda Y$ be an injective linear i/o map, say of order k. It is then readily seen that $\ker \pi^+ \bar{f}$ is a bounded Ω^+K-submodule of ΛU and its order bound is less than or equal to $(-)k$. Indeed, if $u \neq 0$ has order greater than $(-)k$, then $\operatorname{ord} \bar{f}(u) \geqq \operatorname{ord} \bar{f} + \operatorname{ord} u > 0$, whence $\bar{f}(u) \in z^{-1} \Omega^- Y$, and since $\bar{f}(u) \neq 0$ it follows that $u \notin \ker \pi^+ \bar{f}$.

In the present section we shall study the structure of bounded Ω^+K-submodules of ΛU. We emphasize again that U is a finite dimensional K-linear space. The structure of bounded Ω^+K-submodules of ΛU is essentially the same as that of submodules of $\Omega^+ U$, which was discussed in some detail in Hautus and Heymann [1978, § 6] and also in Forney [1975].

Let $\Delta \subset \Lambda U$ be a bounded Ω^+K-submodule with order bound k^Δ, and for each integer j, let $S_j \subset U$ be the K-linear space spanned by the leading coefficients $\hat{u} \in U$ of all elements $u \in \Delta$ which satisfy $\operatorname{ord} u \geqq j$. In this way, we obtain a chain of K-linear spaces

$$U \supset \cdots \supset S_{j-1} \supset S_j \supset \cdots \supset S_{k^\Delta - 1} \supset S_{k^\Delta} \supset S_{k^\Delta + 1} = 0 \tag{4.1}$$

Now, by the finite dimensionality of U, there exists an integer $k_\Delta (\leqq k^\Delta)$ such that $S_{k_\Delta} \neq S_{k_\Delta + 1}$ and $S_{k_\Delta - j} = S_{k_\Delta}$ for all $j > 0$. We call the chain $\{S_j\}$ the *order chain* of Δ and the nonincreasing 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, as the *reduced reachability chain* and the *reduced reachability list of* $\bar{f}$.

PROPOSITION 4.2. *Let $\Delta, \Delta' \subset \Lambda U$ be bounded Ω^+K-submodules with order chains $\{S_j\}$ and $\{S'_j\}$ and order lists $\{\mu_j\}$ and $\{\mu'_j\}$, respectively.*

(i) *If $\Delta' \subset \Delta$ then for each integer j, $S'_j \subset S_j$ and $\mu'_j \leqq \mu_j$.*

(ii) *If $\Delta' \subset \Delta$ and for each integer j, $\mu'_j = \mu_j$, then $\Delta' = \Delta$.*

Proof. (i) This is an immediate consequence of the preceding discussion.

(ii) If $\Delta' \subset \Delta$, then the equalities $\mu'_j = \mu_j$ imply that $S'_j = S_j$ for all j, and if $u \in \Delta$ is any element, there exists an element $u' \in \Delta'$ such that $\operatorname{ord}(u - u') > \operatorname{ord} u$. Further, $u - u' \in \Delta$ so that by the same argument, there is an element $u'' \in \Delta'$ such that $\operatorname{ord}(u - u' - u'') > \operatorname{ord}(u - u')$. Proceeding stepwise, we finally find elements $u', u'', \cdots, u^r \in \Delta'$

such that $\operatorname{ord}(u-u'-u''-\cdots-u^r)>k^\Delta$, where k^Δ is the order bound of Δ. Since $u-u'-\cdots-u^r\in\Delta$, we conclude that $u-u'-\cdots-u^r=0$, whence $u=u'+u''+\cdots+u^r\in\Delta$, so that also $\Delta\subset\Delta'$, and we conclude that $\Delta=\Delta'$, as claimed. □

We turn now to a brief review of some results on proper bases for ΛK-linear spaces and Ω^+K-modules. A set of elements $u_1,\cdots,u_k\in\Lambda U$ is called *properly independent* if and only if their leading coefficients $\hat{u}_1,\cdots,\hat{u}_k\in U$ are K-linearly independent. A basis for a subspace $\mathcal{R}\subset\Lambda U$ is called a *proper basis* if it consists of properly independent elements. If $u_1,\cdots,u_k\in\Lambda U$ is a properly independent set of vectors, then it is also ΛK-linearly independent, as was shown e.g. in Hammer and Heymann [1981, Lem. 4.2], where also the following characterization of proper independence was proved. (See also Forney [1975].)

LEMMA 4.3. *A set of nonzero elements $u_1,\cdots,u_k\in\Lambda U$ is properly independent if and only if for every set of scalars $\alpha_1,\cdots,\alpha_k\in\Lambda K$, or alternatively, if and only if for every set of scalars $\alpha_1,\cdots,\alpha_k\in\Omega^+K$, the following holds*:

$$\operatorname{ord}\sum_{i=1}^{k}\alpha_iu_i=\min\{\operatorname{ord}\alpha_iu_i\,|\,i=1,\cdots,k\}.$$

Proper bases play a role in the theory of causal ΛK-linear maps analogous to the role of bases in general in the theory of linear maps. In particular, let $\bar{f}:\Lambda U\to\Lambda Y$ be a ΛK-linear map and let $u_1,\cdots,u_m$ be a basis for ΛU. If $\bar{f}$ acts causally on every element u_i, that is $\operatorname{ord}f(u_i)\geqq\operatorname{ord}u_i$, it is not necessarily implied that $\bar{f}$ is a causal map. Yet, if $u_1,\cdots,u_m$ is a proper basis, the causality of $\bar{f}$ is assured. This is shown in the following proposition (see also Wolovich [1974]):

PROPOSITION 4.4. *Let $u_1,\cdots,u_m$ be a proper basis for the ΛK-linear space ΛU and let $\bar{f}:\Lambda U\to\Lambda Y$ be a ΛK-linear map. Then $\bar{f}$ is causal if and only if* $\operatorname{ord}\bar{f}(u_i)\geqq\operatorname{ord}u_i$ *for all $i=1,\cdots,m$.*

Proof. The "only if" part is true by definiton. To prove the "if" part, assume $\operatorname{ord}\bar{f}(u_i)\geqq\operatorname{ord}u_i$, $i=1,\cdots,m$, let $0\neq u\in\Lambda U$ be any element and write $u=\sum_{i=1}^m\alpha_iu_i$ for appropriate scalars $\alpha_i\in\Lambda K$, $i=1,\cdots,m$. Then,

$$\begin{aligned}\operatorname{ord}\bar{f}(u)=\operatorname{ord}\sum_{i=1}^{m}\alpha_i\bar{f}(u_i)&\geqq\min\{\operatorname{ord}\alpha_i\bar{f}(u_i)|i=1,\cdots,m\}\\&\geqq\min\{\operatorname{ord}\alpha_iu_i|i=1,\cdots,m\}=\operatorname{ord}u,\end{aligned}$$

where the last step is by Lemma 4.3. Thus $\bar{f}$ is causal. □

Through a similar application of Lemma 4.3, we also have the following:

COROLLARY 4.5. *Let $u_1,\cdots,u_m$ be a proper basis for ΛU and let $l:\Lambda U\to\Lambda U$ be a ΛK-linear map. Then $\bar{l}$ is bicausal if and only if the following conditions both hold*:

(i) $\operatorname{ord}\bar{l}(u_i)=\operatorname{ord}u_i$, $i=1,\cdots,m$, *and*

(ii) *The set $\bar{l}(u_1),\cdots,\bar{l}(u_m)$ is a proper basis for ΛU.*

A basis $u_1,\cdots,u_m$ of an Ω^+K-module $\Delta\subset\Lambda U$ is called *proper* if $u_1,\cdots,u_m$ are properly independent, and it will be called *ordered* if $\operatorname{ord}u_i\geqq\operatorname{ord}u_{i+1}$ for all $i=1,2,\cdots,m-1$.

THEOREM 4.6. *Let $\Delta\subset\Lambda U$ be a bounded Ω^+K-submodule with order chain $\{S_j\}$ and order list $\{\mu_j\}$. Then* (i) *there exists an ordered proper basis for Δ.* (ii) *If $u_1,\cdots,u_m$ is an order proper basis for Δ, then the following hold*:

$$\operatorname{ord}u_j=i\ \text{for}\ \mu_{i+1}<j\leqq\mu_i\ \text{and}\ i\leqq k^\Delta;\tag{4.7}$$

For each integer $i\leqq k^\Delta$, the set of leading coefficients
$\hat{u}_1,\cdots,\hat{u}_{\mu_i}$ *forms a basis for S_i.* (4.8)

Proof. The proof is essentially the same as that of Theorem 6.11 in Hammer and Heymann [1981], which deals with proper bases for $\Omega^- K$-submodules of ΛU. We shall therefore give only an outline. In view of the chain property of the S_i, there exists a set $u_1^0, \cdots, u_m^0$ of vectors in U (where $m = \mu_{k_\Delta} = \text{rank } \Delta$) such that for each $k_\Delta \leqq i \leqq k^\Delta$, $u_1^0, \cdots, u_{\mu_i}^0$ is a basis for S_i. Then, for each $k_\Delta \leqq i \leqq k^\Delta$ and each $\mu_{i+1} < j \leqq \mu_i$, there is an element $u_j \in \Delta$ having order i and leading coefficient $\hat{u}_j = u_j^0$. Obviously, the set $u_1, \cdots, u_m$ is properly independent and the $\Omega^+ K$-module Δ' generated by $u_1, \cdots, u_m$ satisfies $\Delta' \subset \Delta$. That actually $\Delta' = \Delta$ follows upon application of Proposition 4.2 (ii). Hence $u_1, \cdots, u_m$ is an ordered proper basis for Δ and satisfies conditions (4.7) and (4.8) by construction . Finally, that each ordered proper basis has these properties follows from the fact that for each integer j, every ordered proper basis $u_1, \cdots, u_m$ of Δ has precisely μ_j elements whose order is greater than or equal to j and $\text{span}_K \{\hat{u}_1, \cdots, \hat{u}_k\} = S_j$. □

Let $\Delta, \Delta' \subset \Lambda U$ be $\Omega^+ K$-submodules. An $\Omega^+ K$-homomorphism $q : \Delta \to \Delta'$ is called *order preserving* if $\text{ord } q(u) = \text{ord } u$ for each $0 \neq u \in \Delta$. If an order preserving q is surjective it is obviously an isomorphism, and we call it in this case an *order (preserving) isomorphism.* The submodules Δ and Δ' are then called *order isomorphic* (compare with the polynomial case in Hautus and Heymann [1978]).

PROPOSITION 4.9. *Let $\Delta, \Delta' \subset \Lambda U$ be bounded $\Omega^+ K$-submodules. Then Δ and Δ' are order isomorphic if and only if they have the same order lists.*

Proof. If Δ and Δ' are order isomorphic, then it follows directly from Theorem 4.6 and Corollary 4.5 that they have the same order lists. Conversely, assume that the bounded modules Δ and Δ' are nonzero and have the same order lists. Then, by Theorem 4.6, the following hold: (i) Δ and Δ' have ordered proper bases $u_1, \cdots, u_m$ and $u_1', \cdots, u_m'$, respectively, (ii) $m' = m$, and (iii) $\text{ord } u_i' = \text{ord } u_i$ for all $i = 1, \cdots, m$. By Hammer and Heymann [1981, Thm. 4.4], there exist then elements $u_{m+1}, \cdots, u_n$ and $u_{m+1}', \cdots, u_n'$ such that both of the sets $u_1, \cdots, u_n$ and $u_1', \cdots, u_n'$ form proper bases of ΛU, and $\text{ord } u_i' = \text{ord } u_i$ for all $i = 1, \cdots, n$. But then, the ΛK-linear map $\bar{l} : \Lambda U \to \Lambda U$ defined through its values as $\bar{l} u_i = u_i'$, $i = 1, \cdots, n$, is bicausal by Corollary 4.5, and, since evidently $\bar{l}[\Delta] = \Delta$, our proof is complete. □

It will be convenient in the sequel to define for a bounded $\Omega^+ K$-module $\Delta \subset \Lambda U$ of rank m, a set of integers $\{\nu_1, \cdots, \nu_m\}$ called the *degree indices* of Δ, as follows. Let $u_1, \cdots, u_m$ be an ordered proper basis of Δ and for each $i = 1, \cdots, m$ define $\nu_i = -\text{ord } u_i$. The relationship between the degree indices and the order list of Δ is established by Theorem 4.6 through (4.7) and (4.8), as follows:

$$\nu_j = -i \quad \text{for } \mu_{i+1} < j \leqq \mu_i, \quad i \leqq k^\Delta. \tag{4.10}$$

An $\Omega^+ K$-submodule $\Delta \subset \Lambda U$ is called *full* if it contains a basis for ΛU. In case Δ is a bounded module, then, clearly, Δ is full if and only if $\text{rank}_{\Omega^+ K} \Delta = \dim U$.

5. The precompensation orbit of injective i/o maps. In the present section we shall study the structure of the $\Omega^+ K$-module $\ker \pi^+ \bar{f}$ for injective linear i/o maps. We shall also investigate the structural invariants of bicausal precompensation orbits.

It is well known from linear realization theory (see e.g. Fuhrmann [1976]) that in view of the rationality of $\bar{f}$, $\ker \pi^+ \bar{f} j^+$ is a full submodule of ΛU. It follows then immediately, since $\ker \pi^+ \bar{f} j^+ \subset \ker \pi^+ \bar{f}$, that $\ker \pi^+ \bar{f}$ is also full.

Let $\bar{f} : \Lambda U \to \Lambda Y$ be an injective linear i/o map. We define the *reduced reachability indices* $\{\nu_1, \cdots, \nu_m\}$ of $\bar{f}$ as the degree indices of $\ker \pi^+ \bar{f}$. We observe that in view of the strict causality of $\bar{f}$, the ν_i are all positive integers. Indeed, if $0 \neq u \in \ker \pi^+ \bar{f}$, then $0 \neq \bar{f}(u) \in \Omega^+ Y$ and $\text{ord } u < \text{ord } \bar{f}(u) \leqq 0$.

Consider now an injective linear i/o map $\bar{f}: \Lambda U \to \Lambda Y$ and let $\bar{l}: \Lambda U \to \Lambda U$ be a bicausal precompensator for $\bar{f}$. Clearly $\bar{l}$ is also an order preserving Ω^+K-isomorphism on ΛU, and since $\bar{l}[\ker \pi^+ \bar{f}\bar{l}] = \ker \pi^+ \bar{f}$ we see (in view of Proposition 4.9) that the reduced reachability list or, equivalently, the set of reduced reachability indices, is an orbital invariant of bicausal precompensation. Combining this fact with Corollary 3.8, we obtain the following central factorization result:

THEOREM 5.1. *Let $f, f': \Lambda U \to \Lambda U$ be injective linear i/o maps with reduced reachability indices $\{\nu_1, \cdots, \nu_m\}$ and $\{\nu'_1, \cdots, \nu'_m\}$, respectively. Then $\nu_i = \nu'_i$, $i = 1, \cdots, m$ if and only if there exists a polynomial unimodular map $M: \Lambda Y \to \Lambda Y$ and a bicausal precompensator $\bar{l}: \Lambda U \to \Lambda U$ such that $\bar{f}' = M\bar{f}\bar{l}$.*

Proof. If $\bar{f}' M \bar{f} \bar{l}$ with M polynomial unimodular and $\bar{l}$ bicausal, then, by Corollary 3.8, $\ker \pi^+ \bar{f}' = \ker \pi^+ \bar{f}\bar{l}$, whence $\bar{l}[\ker \pi^+ \bar{f}'] = \ker \pi^+ \bar{f}$, and by Proposition 4.9, $\ker \pi^+ \bar{f}'$ and $\ker \pi^+ \bar{f}$ have the same order lists (or equivalently the same reduced reachability indices). Conversely, if $\nu_i = \nu'_i$ for $i = 1, \cdots, m$, then $\ker \pi^+ \bar{f}$ and $\ker \pi^+ \bar{f}'$ have the same order lists, and by Proposition 4.9 are order isomorphic. Thus, there exists a bicausal isomorphism on ΛU, say $\bar{l}^{-1}$, such that $\ker \pi^+ \bar{f}' = \bar{l}^{-1} \ker \pi^+ \bar{f} = \ker \pi^+ \bar{f}\bar{l}$. By Corollary 3.8 there exists then a unimodular polynomial map $M: \Lambda Y \to \Lambda Y$ such that $\bar{f}' = M\bar{f}\bar{l}$, concluding the proof. □

A factorization of the type obtained in Theorem 5.1 is sometimes called in the literature a Wiener–Hopf factorization (compare Fuhrmann and Willems [1979]).

Before proceeding with our discussion, we turn to the proof of Theorem 3.2, which is an immediate consequence of Theorem 5.1.

Proof 5.2. *Proof of Theorem* 3.2. Assume that $\bar{f}$ has reduced reachability indices $\{\nu_1, \cdots, \nu_m\}$. The injectivity of $\bar{f}$ implies that $r := \dim Y \geqq m (= \dim U)$. Let $\bar{f}': \Lambda U \to \Lambda Y$ be the ΛK-linear map whose transfer matrix is given by

$$\mathscr{T}_{\bar{f}'} = \begin{bmatrix} z^{-\nu_1} & & 0 \\ & \ddots & \\ 0 & & z^{-\nu_m} \\ \hline & 0 & \end{bmatrix}.$$

Clearly $\bar{f}'$ is strictly observable and has the same reduced reachability indices $\{\nu_1, \cdots, \nu_m\}$ as $\bar{f}$. Theorem 5.1 then implies that $\bar{f} = M\bar{f}'\bar{l}$ for some polynomial unimodular map M and a bicausal ΛK-linear map $\bar{l}$. Then the map $\bar{l}^{-1}$ is a bicausal precompensator for $\bar{f}$ and the map $\bar{f}'' = \bar{f}\bar{l}^{-1}$ is strictly observable since $\bar{f}'' = M\bar{f}'$, and by Corollary 3.8 $\ker \pi^+ \bar{f}'' = \ker \pi^+ \bar{f}' (\subset \Omega^+ U)$. □

We conclude this section with a discussion of the problem of "dynamics assignment" through bicausal precompensation. That is, we ask to what extent it is possible to modify a system's essential dynamic characteristics through the application of bicausal precompensator.

We first recall some classical concepts. Let $\bar{f}: \Lambda U \to \Lambda Y$ be a linear i/o map and let $\tilde{f} := \pi^+ \cdot \bar{f} \cdot j^+$ be the restricted i/o map associated with $\bar{f}$. The Ω^+K-submodule $\Delta := \ker \tilde{f} = \ker \pi^+ \bar{f} \cap \Omega^+ U$, called the *realization kernel* (*or realization module*) of $\bar{f}$, uniquely defines the class of all canonical realizations of $\bar{f}$ (see e.g. Hautus and Heymann [1978]). In particular, let $X_\Delta := \Omega^+ U / \Delta$, let $g_\Delta := \Omega^+ U \to X_\Delta$ be the canonical projection and let $h_\Delta: X_\Delta \to \Lambda Y / \Omega^+ Y$ denote the (unique) Ω^+K-homomorphism such that $\tilde{f} = g_\Delta \cdot h_\Delta$. Then $(X_\Delta, g_\Delta, h_\Delta)$ is a canonical realization of $\tilde{f}$. Thus, the realization kernel Δ characterizes the essential dynamical properties of $\bar{f}$ and its *reachability indices* are the degree indices of the realization kernel Δ. (The reachability indices are, of course, the well known *Kronecker invariants* of canonical realizations of $\bar{f}$—see also Hautus and Heymann [1978], Kalman [1971] and Kailath [1980]).

Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective i/o map with reachability indices $\{\sigma_1, \cdots, \sigma_m\}$ and reduced reachability indices $\{\nu_1, \cdots, \nu_m\}$. Since, clearly, $\ker \bar{f} = \ker \pi^+ \bar{f} j^+ \subset \ker \pi^+ \bar{f}$, it follows upon application of Proposition 4.2 (i) and formula (4.10), that $\sigma_i \geqq \nu_i$ for $i = 1, \cdots, m$. We have seen previously that the reduced reachability indices are orbital invariants for injective orbits of precompensation and they are shared by all i/o maps in the orbit. This, in particular, holds also for the strictly observable i/o maps. If $\bar{f}$ is strictly observable, then $\ker \pi^+ \bar{f} \subset \Omega^+ U$, whence $\ker \pi^+ \bar{f} = \ker \pi^+ \bar{f} j^+$, implying that the reduced reachability indices of $\bar{f}$ coincide with its reachability indices, that is, $\sigma_i = \nu_i$, $i = 1, \cdots, m$. Conversely, suppose an i/o map $\bar{f}$ in the precompensation orbit satisfies $\sigma_i = \nu_i$, $i = 1, \cdots, m$. Then $\ker \pi^+ \bar{f} = \ker \pi^+ \bar{f} j^+$ (see Proposition 4.2 (ii)), and it follows that $\ker \pi^+ \bar{f} \subset \Omega^+ U$, implying that $\bar{f}$ is strictly observable. We just proved the following:

THEOREM 5.3. *Consider a fixed injective bicausal precompensation orbit O and let $\{\nu_1, \cdots, \nu_m\}$ denote the reduced reachability indices of elements in O. Consider an i/o map $\bar{f} \in O$ with reachability indices $\{\sigma_1, \cdots, \sigma_m\}$. Then* (i) $\sigma_i \geqq \nu_i$, $i = 1, \cdots, m$; (ii) $\sigma_i = \nu_i$, $i = 1, \cdots, m$ *if and only if $\bar{f}$ is strictly observable.*

In § 3 we saw that the McMillan degrees of i/o maps in an injective bicausal precompensation orbit are bounded below by the McMillan degree of the strictly observable i/o maps in the orbit. Since the McMillan degree of an i/o map is equal to the sum of its reachability indices, this result is of course contained in Theorem 5.3, which gives a much stronger minimality result.

Before we proceed further, we wish to make a few remarks on the explicit construction of $\ker \pi^+ \bar{f}$ and the computation of the reduced reachability indices. Suppose $\bar{f}$ is an injective linear i/o map with transfer matrix $\mathscr{T} = \mathscr{T}(z^{-1})$. Then $\operatorname{rank} \mathscr{T} = m$ and we let $\psi = \psi(z)$ denote the least common denominator of the entries of $\mathscr{T}$. Then $\psi \cdot \mathscr{T}$ is a polynomial matrix and there exists a unimodular polynomial matrix M, such that $M(\psi \cdot \mathscr{T}) = \begin{bmatrix} D \\ 0 \end{bmatrix}$, where D is a nonsingular polynomial matrix. Hence $M \cdot \mathscr{T} = \begin{bmatrix} \psi^{-1} D \\ 0 \end{bmatrix}$, and we claim that $\ker \pi^+ \bar{f} = \psi \cdot D^{-1} \Omega^+ U$. Indeed, $u \in \ker \pi^+ \bar{f}$ if and only if $\bar{f}(u) = \mathscr{T} \cdot u \in \Omega^+ Y$ (where we do not distinguish sharply between the map and its associated transfer matrix). But, since M is a unimodular polynomial matrix, $\mathscr{T} \cdot u \in \Omega^+ Y$ if and only if $M \cdot \mathscr{T} \cdot u \in \Omega^+ Y$, which in turn holds if and only if $\psi^{-1} D u \in \Omega^+ Y$. Now, $\ker \pi^+ \bar{f}$ is a full bounded submodule of ΛU and hence has an ordered proper basis $\{d_1, \cdots, d_m\}$. The reduced reachability indices of $\bar{f}$ are then $\{\nu_1, \cdots, \nu_m\}$ where $\nu_i = -\operatorname{ord} d_i$. Finally, we note that upon defining the matrix $D_+ := [d_1, \cdots, d_m]$, we can also write $\ker \pi^+ \bar{f} = D^+ \Omega^+ U$, whence there exists a unimodular polynomial matrix N such that $(\psi^{-1} D) N = D_+$.

LEMMA 5.4. *Let K be an infinite field and let $\Delta \subset \Omega^+ U$ be a full $\Omega^+ K$-submodule with order indices $\{\sigma_1, \cdots, \sigma_m\}$, $(\sigma_i \leqq \sigma_{i+1})$. Further, let $\{\nu_1, \cdots, \nu_m\}$, $(\nu_i \leqq \nu_{i+1})$, be a set of positive integers such that $\nu_i \leqq \sigma_i$, $i = 1, \cdots, m$. Let Y be a K-linear space such that $\dim Y = r \geqq m$. Then there exists an injective linear i/o map $\bar{f}: \Lambda U \to \Lambda Y$ with reduced reachability indices $\{\nu_1, \cdots, \nu_m\}$ and $\ker \bar{f} = \Delta$.*

Proof. Let $d_1, \cdots, d_m$ be an ordered proper basis for Δ and define the matrix $D := [d_1, \cdots, d_m]$. Let $\alpha \in K$ be any element which is not a zero of $\det D$. (Such an α exists since K is infinite.) For each $i = 1, \cdots, m$ let $\delta_i := \sigma_i - \nu_i$ and define the $(m \times m)$-matrix $D_0 := \operatorname{diag}((z-\alpha)^{\delta_1}, \cdots, (z-\alpha)^{\delta_m})$. We now let $\bar{f}: \Lambda U \to \Lambda Y$ be the K-linear map whose transfer matrix is defined by

$$\mathscr{T} = \begin{bmatrix} D_0 & D^{-1} \\ & 0 \end{bmatrix},$$

where the zero submatrix is $(m-r) \times m$ and may be empty. To see that $\bar{f}$ has the

desired properties, note first that D_0 and D are right coprime and $\ker \tilde{f} = D\Omega^+U$ (see also Hautus and Heymann [1978]). Furthermore, $\ker \pi^+ \bar{f} = DD_0^{-1}\Omega^+U$, whence it follows that the set $(z-\alpha)^{-\delta_1} \cdot d_1, \cdots, (z-\alpha)^{-\delta_m} \cdot d_m$ forms a proper basis for $\ker \pi^+ \bar{f}$, and since $\operatorname{ord}[(z-\alpha)^{-\delta_i} d_i] = -\nu_i$, the proof is complete. □

THEOREM 5.5. *Let K be an infinite field and consider an injective bicausal precompensation orbit O with reduced reachability indices $\{\nu_1, \cdots, \nu_m\}$. Let $\Delta \subset \Omega^+U$ be a full Ω^+K-submodule with order indices $\{\sigma_1, \cdots, \sigma_m\}$. There exists an i/o map $\bar{f} \in O$ such that $\ker \tilde{f} = \Delta$ if and only if $\sigma_i \geqq \nu_i$, $i = 1, \cdots, m$.*

Proof. Necessity follows from Theorem 5.3 (i). To see the sufficiency, let $\sigma_i \geqq \nu_i$, $i = 1, \cdots, m$. By Lemma 5.4 there exists an injective linear i/o map f_0, (not necessarily in O), which has $\{\nu_1, \cdots, \nu_m\}$ as reduced reachability indices and $\ker \tilde{f}_0 = \Delta$. Let $\bar{f}_1$ be any i/o map in O. By Theorem 5.1 there exist then a unimodular polynomial map M and a bicausal ΛK-linear map $\bar{l}$ such that $f_1 = M\bar{f}_0\bar{l}$. Thus, $\bar{f} := \bar{f}_1\bar{l}^{-1} = M\bar{f}_0$, where $\bar{f} \in O$, and by Corollary 3.8 $\ker \pi^+ \bar{f} = \ker \pi^+ \bar{f}_0$, concluding the proof. □

It is noteworthy that the requirement of infinite fields in Theorem 5.5 and Lemma 5.4 is an essential one. To demonstrate this fact, consider the following elementary example. Let K be the field of integers modulo 2. Let $\dim Y = \dim U = 1$ and consider as realization kernel the module $\Delta = z(z+1)\Omega^+U$. The degree index of Δ is $\sigma = 2$, but Δ cannot be (canonically) associated with precompensation orbits whose reduced reachability index is $\nu = 1$.

6. Strict observability and output feedback. In § 3 we have seen that every injective i/o map can be rendered strictly observable by static state feedback. The main result of the present section is that every injective i/o map can be rendered strictly observable also by application of (dynamic) causal output feedback.

We begin with some preliminaries. Let $\bar{f} : \Lambda U \to \Lambda Y$ be a linear i/o map and let $\bar{l} : \Lambda U \to \Lambda U$ be a bicausal precompensator for $\bar{f}$. We shall say that $\bar{l}$ is *$\bar{f}$-causal* if there exist a causal ΛK-linear (ouput feedback) map $\bar{g} : \Lambda Y \to \Lambda U$, and an invertible static map $V : \Lambda U \to \Lambda U$ such that $\bar{l} = (I + \bar{g}\bar{f})^{-1}V$. Similarly, we shall say that $\bar{l}$ is *$\bar{f}$-polynomial* if there exist a polynomial ΛK-linear map $\bar{g} : \Lambda Y \to \Lambda U$ and an invertible static map $V : \Lambda U \to \Lambda U$ such that $\bar{g} \cdot \bar{f}$ is strictly causal and $\bar{l} = (I + \bar{g}\bar{f})^{-1}V$. Denoting $\bar{f}' := \bar{f}\bar{l}$, we obtain through a simple calculation that if $\bar{l}$ is $\bar{f}$-polynomial, then $\bar{l}^{-1}$ is $\bar{f}'$ polynomial, and if $\bar{l}$ is $\bar{f}$-causal, then $\bar{l}^{-1}$ is $\bar{f}'$-causal.

While it is always true that $\ker \pi^+ \bar{f} = \bar{l}[\ker \pi^+ \bar{f}']$, it is in general not true that a similar formula relates $\ker \tilde{f}'$ ($\subset \ker \pi^+ \bar{f}'$) with $\ker \tilde{f}$ ($\subset \ker \pi^+ \bar{f}$). An exception to this general situation is given in the following:

LEMMA 6.1. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be a linear i/o map, let $\bar{l} : \Lambda U \to \Lambda U$ be a bicausal precompensator and let $\bar{f}' := \bar{f}\bar{l}$. If $\bar{l}$ is $\bar{f}$-polynomial, then $\ker \tilde{f} = \bar{l}[\ker \tilde{f}']$.*

Proof. If $u \in \ker \tilde{f}$ then $u \in \Omega^+U$ and $\bar{f}(u) \in \Omega^+Y$. Since $\bar{g}$ is a polynomial map, it then also follows that $\bar{g}\bar{f}(u) \in \Omega^+U$. Thus, $\bar{l}^{-1}(u) = V^{-1}(I + \bar{g}\bar{f})u \in \Omega^+U$. Moreover $\bar{f}'\bar{l}^{-1}(u) = \bar{f}\bar{l}\bar{l}^{-1}(u) = \bar{f}(u) \in \Omega^+Y$. Hence $\bar{l}^{-1}(u) \in \ker \tilde{f}'$ (or $u \in \bar{l}[\ker \tilde{f}']$) and consequently $\ker \tilde{f} \subset \bar{l}[\ker \tilde{f}']$. The inverse inclusion follows similarly from the fact that $\bar{l}^{-1}$ is $\bar{f}'$-polynomial, and the lemma follows. □

Combining Lemma 6.1 with Proposition 4.9, we obtain

THEOREM 6.2. *Let $\bar{f} : \Lambda U \to \Lambda Y$ be a linear i/o map, let $\bar{l} : \Lambda U \to \Lambda U$ be a bicausal precompensator and write $\bar{f}' = \bar{f} \cdot \bar{l}$. If $\bar{l}$ is $\bar{f}$-polynomial then $\bar{f}$ and $\bar{f}'$ have the same sets of reachability indices.*

Consider now an injective linear i/o map $\bar{f} : \Lambda U \to \Lambda Y$, and let $\bar{l} : \Lambda U \to \Lambda U$ be a bicausal precompensator for $\bar{f}$. Clearly, in view of the injectivity of $\bar{f}$, there exist a ΛK-linear map $\bar{g} : \Lambda Y \to \Lambda U$ and an invertible static map $V : \Lambda U \to \Lambda U$ such that

$\bar{l} = (I + \bar{g}\bar{f})^{-1}V$ and $\bar{g}\bar{f}$ is strictly causal. Next, let $\bar{g} = \bar{g}^- + \bar{g}^+$ where $\bar{g}^-$ is causal and $\bar{g}^+$ is (strictly) polynomial. Then $\bar{g}^- \cdot \bar{f}$ is obviously strictly causal, and so also is $\bar{g}^+ \cdot \bar{f}$, being the difference of two strictly causal maps. Thus, we have the following:

$$\begin{aligned}\bar{l} &= (I + \bar{g}\bar{f})^{-1}V = (I + \bar{g}^-\bar{f} + \bar{g}^+\bar{f})^{-1}V \\ &= (I + \bar{g}^-\bar{f})^{-1}[I + \bar{g}^+\bar{f}(I + \bar{g}^-\bar{f})^{-1}]^{-1}V = \bar{l}^- \cdot \bar{l}^+,\end{aligned} \tag{6.3}$$

where $\bar{l}^- := (I + \bar{g}^-\bar{f})^{-1}$ is a bicausal precompensator for $\bar{f}$ and is $\bar{f}$-causal, and where $\bar{l}^+ := [I + \bar{g}^+\bar{f}(I + \bar{g}^-\bar{f})^{-1}]^{-1}V = (I + \bar{g}^+(\bar{f}\bar{l}^-))^{-1}V$ is a bicausal precompensator for $\bar{f}\bar{l}^-$ and is $(\bar{f}\bar{l}^-)$-polynomial. If we now apply Theorem 6.2, we conclude that the maps $\bar{f}\bar{l}$ and $\bar{f}\bar{l}^-$ have the same sets of reachability indices, the important fact being that $\bar{l}^-$ represents a (dynamic) causal output feedback around $\bar{f}$. This proves the following:

THEOREM 6.4. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective linear i/o map and assume that $\bar{l}: \Lambda U \to \Lambda U$ is a bicausal precompensator for $\bar{f}$ such that $\bar{f}\bar{l}$ is rational and has reachability indices $\sigma_1, \cdots, \sigma_m$. Then there exists a causal ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda U$ such that $\bar{f}(I + \bar{g}\bar{f})^{-1}$ also has reachability indices $\sigma_1, \cdots, \sigma_m$.*

As an immediate consequence of Theorems 3.2 and 6.4, we have the following result:

THEOREM 6.5. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective linear i/o map. Then $\bar{f}$ can be transformed into a strictly observable map by application of causal (dynamic) output feedback.*

Finally, upon application of Theorem 6.4 to Theorem 5.5, we obtain

THEOREM 6.6. *Let K be an infinite field and let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective i/o map with reduced reachability indices $\nu_1, \cdots, \nu_m$. For every set of integers $\sigma_1, \cdots, \sigma_m$ satisfying $\sigma_i \geqq \nu_i$, $i = 1, \cdots, m$, there exists a causal ΛK-linear map $\bar{g}: \Lambda Y \to \Lambda U$ such that $\bar{f}(I + \bar{g}\bar{f})^{-1}$ has $\sigma_1, \cdots, \sigma_m$ as reachability indices.*

7. Some further properties of $\ker \pi^+\bar{f}$. In the present section, we wish to make formal contact between the present theory and some concepts that appeared in the linear system theory literature. In particular, we wish to make contact with concepts from the geometric theory as expounded by Wonham and Morse (see e.g. Wonham [1979]). It will be assumed that the reader is familiar with the basic concepts of that theory, and with the basic algebraic framework of linear realization theory (as presented, e.g., in Hautus and Heymann [1978]).

Let $\bar{f}: \Lambda U \to \Lambda Y$ be a linear i/o map and let (X, g, h) be a reachable realization of $\bar{f}$ (i.e., $\pi^+\bar{f}j^+ = h \cdot g$, and $g: \Omega^+U \to X$ is surjective). The *unobservable subspace* (submodule) of (X, g, h) is defined as $\ker h \subset X$ and we say that (X, g, h) is *observable* if $\ker h = 0$, i.e., if h is injective. Let $\bar{g}: \Lambda U \to \Lambda X$ denote the (extended) i/s map associated with g and let $\bar{f} = H\bar{g}$ be the corresponding state representation (i.e., $H = p_1 \cdot h$).

We recall that a subspace $S \subset X$ is called *weakly invariant* if the controlled trajectory for every $x \in S$ can be maintained in S by choice of control action. Weakly invariant subspaces coincide with the well known (A, B)-invariant spaces of geometric linear system theory (see in particular Hautus [1979] for comparison of the various concepts). Of particular interest is the maximal weakly invariant subspace contained in $\ker H$, which is frequently denoted in the literature by v^*. We shall show below that v^* is related to $\ker \pi^+\bar{f}$ and, in particular, that $v^* = p_1\bar{g}[\ker \pi^+\bar{f}]$.

The Ω^+K-module $\ker \pi^+\bar{f}$ consists of the class of all inputs for which the corresponding output is identically zero for all $t \geqq 1$. Let $u \in \ker \pi^+\bar{f}$ be any control and write $u = u^+ + u^-$, where $u^+ \in \Omega^+U$ and $u^- \in z^{-1}\Omega^-U$. Then $0 = p_1\bar{f}u = p_1H\bar{g}u =$

$Hp_1\bar{g}u = Hp_1\bar{g}u^+ + Hp_1\bar{g}u^-$ and, in view of the strict causality of $\bar{g}$, $p_1\bar{g}u^- = 0$ and we have $p_1\bar{g}u = p_1\bar{g}u^+ \in \ker H$. The state $p_1\bar{g}u^+ = gu^+ \in X$ is the state at time $t = 1$ generated by the control u^+. This state is maintained in $\ker H$ by application (after $t = 0$) of the control u^-, and hence it is clear that $p_1\bar{g}u \in v^*$, so that $p_1\bar{g}[\ker \pi^+\bar{f}] \subset v^*$. To see that the inverse inclusion also holds, let $x \in v^*$ be any state. In view of the reachability of (X, g, h), there exists $u^+ \in \Omega^+U$ such that $x = gu^+ = p_1\bar{g}u^+$. Further, there exists $u^- \in z^{-1}\Omega^-U$ such that the corresponding state trajectory (starting at x) remains in $\ker H$, i.e., the output trajectory is identically zero. Thus, $p_k\bar{f}(u^+ + u^-) = 0$ for all $k \geqq 1$, whence $u = u^+ + u^- \in \ker \pi^+\bar{f}$. We summarize the above discussion in the following:

THEOREM 7.1. *Let (X, g, h) be a reachable realization of a liner i/o map $\bar{f}: \Lambda U \to \Lambda Y$ and let $\bar{f} = H \cdot \bar{g}$ be the corresponding (reachable) state representation. Then the maximal weakly invariant subspace in* $\ker H$ *is given by*

$$v^* = p_1\bar{g}[\ker \pi^+\bar{f}]. \tag{7.2}$$

We shall next investigate several properties of v^* and its relation to unobservability and feedback. First, the following can be readily verified.

LEMMA 7.3. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be a linear i/o map and let $\bar{f} = H\bar{g}$ be a state representation for $\bar{f}$. Then,*

(i) $p_1\bar{g}[\ker \pi^+\bar{f}] \subset \ker H$.

(ii) *If $\Delta \subset \Lambda U$ is an Ω^+K-module satisfying $p_1\bar{g}[\Delta] \subset \ker H$, then $\Delta \subset \ker \pi^+\bar{f}$.*

Consider now the reachable realization (X, g, h) and let $\bar{f} = H \cdot \bar{g}$ be the associated state representation. Clearly, the unobservable subspace $S = \ker h$ satisfies $S \subset \ker H$, and it is easily seen that, in fact, S is the *maximal* Ω^+K-module contained in $\ker H$. Let us apply static state feedback $F: X \to U$ in the reachable realization (X, g, h) (see Hautus and Heymann [1978] for details). Then the reachable extended linear i/s map $\bar{g}: \Lambda U \to \Lambda X$ is transformed into the reachable i/s map $\bar{g}_F := \bar{g}(I + F\bar{g})^{-1}$, and the i/o map $\bar{f}$ is transformed into $\bar{f}_F := \bar{f}(I + F\bar{g})^{-1}$ (so that $\bar{f}_F = H\bar{g}_F$). Let $g_F := p_1 \cdot \bar{g}_F \cdot j^+$ be the output response map associated with $\bar{g}_F$. Then there is a reachable realization (X_F, g_F, h_F) of $\bar{f}_F$, and we denote the unobservable subspace of this realization by $S_F := \ker h_F$. We then have the following theorem, which gives a sharp insight into the nature of the subspace $v^*(= \bar{p}_1\bar{g}[\ker \pi^+\bar{f}])$.

THEOREM 7.4. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be an injective linear i/o map, let (X, g, h) be a reachable realization and let $\bar{f} = H \cdot \bar{g}$ be the associated state representation. Then the following hold:*

(i) *For every static state feedback $F: X \to U$, $S_F \subset p_1\bar{g}[\ker \pi^+\bar{f}]$.*

(ii) *There exists a static state feedback $F_0: X \to U$ (for which $\bar{f}_{F_0}$ is strictly observable) such that $S_{F_0} = p_1\bar{g}[\ker \pi^+\bar{f}]$.*

Proof. (i) The reachability of (X, g, h) implies that, for each feedback F, the realization (X_F, g_F, h_F) is also reachable (see e.g. Hautus and Heymann [1978]). Hence g_F is surjective and there is an Ω^+K-module $\Delta \subset \Omega^+U$ such that $S_F = g_F[\Delta] = p_1\bar{g}_F[\Delta]$ and, since $S_F \subset \ker H$, it follows by Lemma 7.3 that $\Delta \subset \ker \pi^+\bar{f}_F$. Thus, denoting $\bar{l} := (I + F\bar{g})^{-1}$, we obtain

$$S_F = p_1\bar{g}_F[\Delta] \subset p_1\bar{g}_F[\ker \pi^+\bar{f}_F] = p_1\bar{g}\bar{l}[\ker \pi^+\bar{f}\bar{l}] = p_1\bar{g}\bar{l}\bar{l}^{-1}[\ker \pi^+\bar{f}] = p_1\bar{g}[\ker \pi^+\bar{f}],$$

as claimed.

(ii) By Theorem 3.3, there exists an F_0 such that $\bar{f}_{F_0}$ is strictly observable (i.e., $\ker \pi^+\bar{f}_{F_0} \subset \Omega^+U$), so that $g_{F_0}[\ker \pi^+\bar{f}_{F_0}] \subset X$ is an Ω^+K-module. Then since $g_F[\ker \pi^+\bar{f}_{F_0}] = p_1\bar{g}_F[\ker \pi^+\bar{f}_{F_0}] = p_1\bar{g}[\ker \pi^+\bar{f}]$, it follows that $p_1\bar{g}[\ker \pi^+\bar{f}]$ $(= p_1\bar{g}_{F_0}[\ker \pi^+\bar{f}_{F_0}])$ is an Ω^+K-module in $\ker H$, so that $p_1\bar{g}[\ker \pi^+\bar{f}] \subset S_{F_0}$. Combining this with (i) above, we have that $p_1\bar{g}[\ker \pi^+\bar{f}] = S_{F_0}$. □

In the special case when $\bar{f}$ is a strictly observable i/o map, we have that $p_1\bar{g}[\ker \pi^+\bar{f}] = g[\ker g] = 0$, implying that for every static state feedback $F: X \to U$, $S_F = 0$. Thus, the observability is preserved under state feedback, in agreement with Theorem 3.4.

8. Remarks on noninjective i/o maps. We turn now to some observations and comments on noninjective i/o maps. If $\bar{f}: \Lambda U \to \Lambda Y$ is a linear i/o map, we say that $\bar{f}$ has a *static kernel* if there exists a K-linear subspace $U_0 \subset U$ such that $\ker \bar{f} = \Lambda U_0$. If $\ker \bar{f}$ is static, $\bar{f}$ can be made injective by simple restriction of the input value space. The noninjectivity of $\bar{f}$ then stems from the fact that its input value space was chosen to be too large. We proceed now to extend the framework of our theory to noninjective i/o maps.

In Hammer and Heymann [1981, Prop. 5.6], it was shown that a linear i/s map always has a static kernel. Consider now a linear i/s map $\bar{f}: \Lambda U \to \Lambda Y$ and assume that $\ker \bar{f} = \Lambda U_0$ for a subspace $U_0 \subset U$. Choose a direct sum complement $U_1 \subset U$ for U_0 such that $U = U_0 \oplus U_1$ and let $P_1: U \to U_1$ denote the projection of U onto U_1 along U_0. There evidently exists then an injective i/s map $\bar{f}_1: \Lambda U_1 \to \Lambda Y$ such that

$$\bar{f} = \bar{f}_1 P_1. \tag{8.1}$$

The above restriction procedure, and the fact that injective i/s maps are always strictly observable, motivate us in extending the concept of strict observability to noninjective i/o maps as follows:

DEFINITION 8.2. A linear i/o map $\bar{f}: \Lambda U \to \Lambda Y$ is called *extended strictly observable* if the following conditions hold:

(i) $\bar{f}$ has a static kernel $\Lambda U_0 \subset \Lambda U$.

(ii) There exists a subspace $U_1 \subset U$ such that $U_1 \oplus U_0 = U$ and a strictly observable i/o map $f_1: \Lambda U_1 \to \Lambda Y$ such that $\bar{f} = \bar{f}_1 \cdot P_1$, where $P_1: U \to U_1$ is the projection onto U_1 along U_0.

The following theorem generalizes Theorem 3.2 to noninjective linear i/o maps.

THEOREM 8.3. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be a linear i/o map. There exists a bicausal precompensator $\bar{l}: \Lambda U \to \Lambda U$ such that $\bar{f}\bar{l}$ is extended strictly observable.*

The proof of Theorem 8.3 depends on the following:

LEMMA 8.4. *Let $\bar{f}: \Lambda U \to \Lambda Y$ be a linear i/o map. There exists a bicausal precompensator $\bar{l}: \Lambda U \to \Lambda U$ such that $\bar{f}\bar{l}$ has a static kernel.*

The proof of Lemma 8.4 depends on (and is an easy consequence of) the existence of proper bases for ΛK-linear spaces as discussed in Hammer and Heymann [1981]. The details of the proof are omitted.

Proof 8.5. *Outline of proof of Theorem* 8.3. By Lemma 8.4, there exists a bicausal precompensator $\bar{l}: \Lambda U \to \Lambda U$ such that the map $\bar{f}' := \bar{f}\bar{l}$ has a static kernel ΛU_0. There exists then a direct sum complement ΛU_1 to ΛU_0 and an injective i/o map $\bar{f}'': \Lambda U_1 \to \Lambda Y$ such that $\bar{f}' = \bar{f}'' \cdot P_1$, where $P_1: U \to U_1$ is the projection onto U_1 along U_0. By Theorem 3.2 there exists a bicausal precompensator $\bar{l}_2: \Lambda U_1 \to \Lambda U_1$ such that $\bar{f}_1 := \bar{f}'' \cdot \bar{l}_2$ is strictly observable. Finally, it can be shown that $\bar{l}_2$ can be extended to a bicausal ΛK-linear map $\bar{l}_3: \Lambda U \to \Lambda U$ such that $\bar{l}_2 P_1 = P_1 \bar{l}_3$, and we have $\bar{f}_1 \cdot P_1 = \bar{f}'' \bar{l}_2 P_1 = \bar{f}'' P_1 \bar{l}_3 = \bar{f}' \bar{l}_3 = \bar{f}(\bar{l}_1 \bar{l}_3)$, concluding the proof. □

REFERENCES

G. BASILE AND G. MARRO [1969], *On the observability of linear systems with unknown inputs*, J. Optim. Theory Appl., 3, pp. 410–415.

E. J. DAVISON AND S. H. WANG [1974], *Properties and calculation of transmission zeros of linear multivariable systems*, Automatica, 10, pp. 643–658.

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.

P. A. FUHRMANN AND J. C. WILLEMS [1979], *Factorization indices at infinity for rational matrix functions*, Integral Equations Oper. Theory, 2, pp. 287–301.

——— [1981], *A study of (A, B)-invariant subspaces via polynomial models*, Int. J. Control, 31, pp. 467–494.

J. HAMMER AND M. HEYMANN [1981], *Causal factorization and linear feedback*, SIAM J. Control Optim., 19, pp. 445–468.

M. L. J. HAUTUS [1979], *(A, B)-invariant subspaces and stabilizability subspaces: some properties and applications*, Memorandum COSOR 79-17, Eindhoven University of Technology, the Netherlands.

M. L. J. HAUTUS AND M. HEYMANN [1978], *Linear feedback—an algebraic approach*, SIAM J. Control Optim., 16, pp. 83–105.

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.

T. KAILATH [1980], *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ.

R. E. KALMAN [1965], *Algebraic structure of linear dynamical systems: I. The module of Σ"*, Proc. Nat. Acad. Sci., 54, pp. 1503–1508.

——— [1971], *Kronecker invariants and feedback*, in Ordinary Differential Equations, 1971 NRL-MRC Conference, L. Weiss, ed., Academic Press, New York, pp. 459–471.

R. E. KALMAN, P. L. FALB AND M. A. ARBIB [1969], *Topics in Mathematical System Theory*, McGraw-Hill, New York.

P. P. KHARGONEKAR AND E. EMRE [1980], *Further results on polynomial characterizations of (F, G)-invariant subspaces*, Preprint, Center for Mathematical Systems Theory, Univ. of Florida, Gainesville.

B. P. MOLINARY [1976], *A strong controllability and observability in linear multivariable systems*, IEEE Trans. Automat. Control, AC-33, pp. 761–764.

A. S. MORSE [1973], *Structural invariants of linear multivariable systems*, SIAM J. Control, 11, pp. 446–465.

——— [1975], *System invariants under feedback and cascade control*, in Mathematical System Theory, Udine 1975, Lecture Notes in Economics and Mathematical Systems 131, Springer-Verlag, New York, pp. 61–74.

M. S. NIKOLSKII [1970], *Ideally observable systems*, Soviet Math. Dokl., 11, pp. 527–530.

H. J. PAYNE AND L. M. SILVERMAN [1973], *On the discrete time algebraic Riccati equation*, IEEE Trans. Automat. Control, AC-18, pp. 226–234.

D. RAPPAPORT AND L. M. SILVERMAN [1971], *Structure and stability of discrete-time optimal systems*, IEEE Trans. Automat. Control, AC-16, pp. 227–233.

W. A. WOLOVICH [1974], *Linear Multivariable Systems*, Springer-Verlag, New York.

W. M. WONHAM [1979], *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York.

B. F. WYMAN [1972], *Linear systems over commutative rings*, Lecture notes, Stanford Univ., Stanford, CA.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 1, January 1983

0363-0129/83/2101-0002 $01.25/0

DISCRETE APPROXIMATION OF CONTINUOUS TIME STOCHASTIC CONTROL SYSTEMS*

NORBERT CHRISTOPEIT†

Abstract. In this paper it is shown how a continuous time stochastic control problem can be solved by discretization in time. A condition is given under which the optimal strategies for the discrete time problems—obtained by dynamic programming—approximate the solution of the original continuous time problem.

Key words. stochastic control, discrete approximation.

1. Introduction. We consider a system whose dynamics are described by the stochastic differential equation

$$dX = f(t, X(t), U(t))\,dt + dB, \qquad 0 \leqq t \leqq 1, \tag{1.1}$$

with initial condition

$$X(0) = x. \tag{1.2}$$

Here B is a standard Brownian motion, X is the state and U the control process. The objective is to minimize the expected final loss

$$J(U) = Eg(X(1)) \tag{1.3}$$

in some class of admissible controls. The class of admissible controls to be considered will be fairly broad, including randomized controls as well as feedback controls depending on the whole past of the state process.

The approach to be taken is to approximate the continuous time system (1.1)–(1.2) by a sequence of discrete time systems described by stochastic difference equations, with control action taking place only at the beginning of each decision interval. For the discrete time problems, we have then the powerful machinery of dynamic programming at our disposal. This will lead to an analytic expression or at least to a numerical approximation for the discrete time strategies, which turn out to be—generally nonstationary—Markovian. These ideas are explained in § 2. The main step is then to obtain the optimal control law for the original continuous time problem as the limit in a certain weak sense of the discrete time optimal controls as the decision intervals get smaller and smaller. This is done in § 5. For the proof, some estimates on transition probability densities related to solutions of the discrete time systems are needed, which are developed in §§ 3 and 4. These estimates are basically discrete time analogues of corresponding estimates known in the theory of diffusion processes (cf. [15]).

The method of discretizing the system equation is not new. In [16], Yamada treats the control of a system by stationary Markov controls over an infinite time horizon. However, the approximating "discrete" decision problems are discrete only as far as the controls are concerned; the state of the approximating systems still evolve in continuous time according to a stochastic differential equation. Hence, the dynamic programming algorithm cannot be immediately applied to these problems, but some intermediate considerations are necessary (cf. § 5). Our approach, in the general

* Received by the editors September 8, 1980, and in revised form November 15, 1981. This work was supported by the SFB 21 at the University of Bonn.

† Institute for Econometrics and Operations Research, University of Bonn, Adenauerallee 24–42, 53 Bonn, West Germany.

nonstationary finite horizon case, leads directly to discrete time systems governed by stochastic difference equations and hence to problems that are directly amenable to the dynamic programming technique. In addition, we would like to stress one important point. Of course, discretization techniques can be and have been used to derive existence results for continuous time stochastic control problems (cf. [13]). Such results can be obtained as a by-product of our approach, as indicated at the end of § 2. Our main point of interest is, however, the possibility of actually identifying the optimal control of the continuous time problem by looking at the optimal control laws for the discrete time problems, which will generally be easier to obtain. Finally, let us remark that we confine ourselves to the one-dimensional case in order to keep notation as simple as possible. The general multidimensional case can be treated along the same lines and does not offer any new difficulties.

2. Formulation of the continuous and discrete time problems. Throughout the paper let $\mathcal{U}$ denote some compact and convex set of control points. The following assumptions about f will be made.

(A1) There exists a constant L_1 such that for all (s, x, u), $(s', x', u') \in [0, 1] \times \mathbb{R} \times \mathcal{U}$ the inequality

$$|f(s, x, u) - f(s', x', u')| \leqq L_1(|s - s'| + |x - x'| + |u - u'|)$$

is valid.

Looking through the proofs the reader will notice that this global Lipschitz condition can be weakened to a local one by means of the usual truncation procedures (cf. [8]). But, in order not to obscure the basic ideas of the analysis, which will become rather involved anyway, we desist from working out these additional technical details. Secondly, we shall require that f satisfy the linear growth condition

(A2) There exists a constant C_f such that

$$|f(t, x, u)|^2 \leqq C_f(1 + |x|^2)$$

for all $(t, x, u) \in [0, 1] \times \mathbb{R} \times \mathcal{U}$.

Let us now make precise what we understand by an admissible control. An admissible control will be any measurable process $U(t)$, $0 \leqq t \leqq 1$, defined on some probability space $(\Omega, \mathcal{F}, P)$, taking on values in $\mathcal{U}$ for almost all (t, ω) and for which there exists on $(\Omega, \mathcal{F}, P)$ a standard Wiener process $(B(t), \mathcal{F}_t)$, $0 \leqq t \leqq 1$, such that $U = (U(t))$ is adapted to $(\mathcal{F}_t)$ and the equation

$$X(t) = x + \int_0^t f(s, X(s), U(s))\, ds + B(t), \qquad 0 \leqq t \leqq 1, \tag{2.1}$$

possesses a unique solution $X = (X(t))$ with continuous sample paths. X will be called the solution corresponding to the control U. Actually, the statement about existence and uniqueness of a solution to (2.1) is something of a hoax, as will be seen in a moment, and should rather be viewed as a way of identifying X.

By virtue of (A1), the random function $F(t, x, \omega)$ defined by

$$F(t, x, \omega) = f(t, x, U(t, \omega))$$

is measurable and for fixed t and x $\mathcal{F}_t$-measurable, and satisfies

$$|F(t, x, \omega) - F(t, y, \omega)| \leqq L_1|x - y|$$

for all t and all ω. But then it is known that the equation

$$X(t) = x + \int_0^t F(s, X(s))\, ds + B(t)$$

possesses a unique (in the sense of sample path uniqueness) solution with continuous sample paths, which is adapted to $(\mathcal{F}_t)$ (cf. [8]). And this solution apparently solves (2.1), the two equations being identical.

Let us discuss one further point. Of course we want to admit feedback controls $u(t, X)$, where $u : [0, 1] \times C \to \mathcal{U}$ is a measurable function adapted to $(\mathcal{C}_t)$ (C = space of continuous functions on $[0, 1]$ with the sup-norm, $(\mathcal{C}_t)$ = canonical filtration on C), or, as a special case, Markov controls $u(t, X(t))$, where now u is a measurable function on $[0, 1] \times \mathbb{R}$. How do these controls fit in our framework? The answer is simple. Under the assumptions made we may start with an arbitrary Brownian motion $B(t)$ on some probability space $(\Omega, \mathcal{F}, P)$ and define a new probability measure P^u on $(\Omega, \mathcal{F})$ by the Girsanow measure transformation formula, i.e.

$$dP^u = \zeta(f^u)\, dP,$$

where, with $X(t) = x + B(t)$,

$$\zeta(f^u) = \exp\left[\int_0^1 f(t, X(t), u(t, X))\, dX - \frac{1}{2}\int_0^1 |f(t, X(t), u(t, X))|^2\, dt\right].$$

Then, under measure P^u, X solves (2.1) with the new Wiener process $(B^u(t), \mathcal{F}_t^X)$ given by

$$dB^u = dX - f(t, X(t), u(t, X))\, dt$$

(cf. [1]). So, if we set $U(t) = u(t, X)$, U is adapted to $(\mathcal{F}_t^X)$ and X is the unique solution of

$$X(t) = x + \int_0^t f(s, X(s), U(s))\, ds + B^u(t).$$

So our class of admissible controls contains indeed all feedback control laws.

We shall use the notation $J(U) = Eg(X(1))$ if X is the (unique) solution corresponding to the admissible control U. We shall also write $J(u)$ for the cost resulting from the use of the feedback control law u. This notation is justified since under the assumptions made the weak solution of the stochastic differential equation $dX = f(t, X(t), u(t, X))\, dt + dB$ is unique in law and $J(u) = J(U)$ if $U(t) = u(t, X)$.

Remark. It should be noted that, though the solution to (2.1) is unique for a given probability space with a given Brownian motion, corresponding to one and the same process U there may nevertheless be different probability measures P and different Wiener processes $(B(t), \mathcal{F}_t)$ fulfilling the requirements in the definition of admissibility, which may lead to different distributions of the solution process X and hence to different values for $J(U)$. So, to be rigorous, we should write $J[U; P, B]$ or $J[U; P, X]$ to take account of this ambiguity. We shall, however, continue to use the simpler notation $J(U)$, interpreting U as a shorthand symbol for the triple (U, P, X) if necessary. There will be no confusion, since in the sequel there will always be a

definite measure P and a definite Brownian motion B associated with each U. Moreover, for Markov controls u, the distribution of X is unique.

Let us now turn to the cost functional g. We shall impose the following growth condition.

(A3) g is continuous, and $g(x) = O(\exp(\kappa|x|)$ for some $\kappa > 0$.

This condition will ensure that the loss functions $g(X^N(1))$ of the discrete problems to be defined below are uniformly integrable. Note that it incurs no loss of generality that we did not include cost functionals of integral type, since such functionals may be transformed into endpoint criteria by the usual trick of introducing an additional component to the system equation.

The basic idea of our approach is to discretize the system equation (1.1). The corresponding discrete time systems will be described by the stochastic difference equations

$$X_{n+1} = X_n + f_n^N(X_n, U_n)\Delta^N + \varepsilon_n^N, \qquad n = 0, 1, \cdots, \Delta^{-N} - 1, \tag{2.2}$$

with initial condition

$$X_0 = x. \tag{2.3}$$

Here we have put $f_n^N(x, u) = f(n\Delta^N, x, u)$. Δ^{-1} is some natural number bigger than 2 (any other partition of the unit interval with mesh tending to 0 would also do). The ε_n^N are independent normal random variables with mean 0 and variance Δ^N, which are defined on a probability space $(\Omega, \mathcal{F}, P)$ (clearly this may be taken to be the same for all N, n). As controls U_n we admit all nonanticipating random variables on $(\Omega, \mathcal{F}, P)$ with values in $\mathcal{U}$, i.e., all random variables U_n which are independent of $\varepsilon_n^N, \cdots, \varepsilon_{\Delta^{-N}-1}^N$. We shall call an Δ^{-N}-tuple $\pi = (U_0, \cdots, U_{\Delta^{-N}-1})$ an admissible strategy for the Nth system and denote the class of admissible strategies by $\mathcal{S}^N$. The objective is then to choose a strategy π in such a way as to minimize

$$J^N(\pi) = Eg(X_{\Delta^{-N}}), \tag{2.4}$$

$X = (X_0, \cdots, X_{\Delta^{-N}})$ being the solution of (2.2)–(2.3) corresponding to π. This problem will be referred to as the Nth stage discrete problem.

We shall imbed the discrete time problems in a more general sequential decision problem. To this end, define for $\pi = (U_0, \cdots, U_{\Delta^{-N}-1}) \in \mathcal{S}^N$ and any one-dimensional Borel set B

$$p_n^N(B|xu) = P\{X_n + f_n^N(X_n, U_n)\Delta^N + \varepsilon_n^N \in B | X_n = x, U_n = u\}$$

$$= P\{\varepsilon_n^N \in B - x - f_n^N(x, u)\Delta^N\}.$$

Then p_n^N is a transition function which is independent of the particular choice of U_n.

Let us recall the definition of a randomized strategy in dynamic programming. We shall understand by it a Δ^{-N}-tuple $\pi = (\pi_0, \cdots, \pi_{\Delta^{-N}-1})$, where $\pi_n(\Gamma|h_n)$ is a transition function from the histories $h_n = xu_0x_1u_1x_2 \cdots u_{n-1}x_n$ to $\mathcal{U}$. Any $(U_0, \cdots, U_{\Delta^{-N}-1}) \in \mathcal{S}^N$ may then be viewed as a randomized strategy $\pi = (\pi_0, \cdots, \pi_{\Delta^{-N}-1})$ if we take π_n to be a regular version of the conditional distribution

$$\pi_n(\Gamma|h_n) = P\{U_n \in \Gamma | X_0U_0X_1 \cdots U_{n-1}X_n = h_n\}.$$

Then (2.4) may be written

$$\begin{aligned} Eg(X_{\Delta^{-N}}) = \int_{\mathcal{U}} \pi_0(du_0|x) \int p_0^N(dx_1|xu_0) \int_{\mathcal{U}} \pi_1(du_1|xu_0x_1) \\ \times \cdots \times \int p_{\Delta^{-N}-2}^N(dx_{\Delta^{-N}-1}|x_{\Delta^{-N}-2}u_{\Delta^{-N}-2}) \\ \times \int_{\mathcal{U}} \pi_{\Delta^{-N}-1}(du_{\Delta^{-N}-1}|xu_0x_1 \cdots u_{\Delta^{-N}-2}x_{\Delta^{-N}-1}) \\ \times \int p_{\Delta^{-N}-1}^N(dx_{\Delta^{-N}}|x_{\Delta^{-N}-1}u_{\Delta^{-N}-1})g(x_{\Delta^{-N}}) \\ =: E^{\pi}g. \end{aligned}$$

An important role is played by the Markov strategies, which are Δ^{-N}-tuples $u = (u_0, \cdots, u_{\Delta^{-N}-1})$ of measurable functions $u_n : \mathbb{R} \to \mathcal{U}$. For such strategies,

$$\pi_n(\Gamma|h_n) = \delta_\Gamma(u_n(x_n)),$$

($\delta_\Gamma(a)$ = Dirac measure concentrated at point a), where x_n is the last state of the history h_n. It is then well known from dynamic programming that, under assumptions (A1)–(A3), the problem of minimizing $E^{\pi}g$ in the class of all randomized strategies possesses a solution in the subclass of Markov strategies (cf. [7], the lower semicontinuous case). Hence, in solving our Nth stage problem, we may restrict ourselves to Markov strategies, and the optimal strategy can be found by running through the recurrence relations of dynamic programming. Henceforth, let $u^N = (u_0^N, \cdots, u_{\Delta^{-N}-1}^N)$ be the optimal (Markov) strategy found this way, i.e.,

$$E^{u^N}g = Eg(X_{\Delta^{-N}}^N) = J^N(u^N) \leqq J^N(\pi)$$

for all $\pi \in \mathcal{S}^N$, where $X^N = (X_0^N, \cdots, X_{\Delta^{-N}}^N)$ is the solution of the stochastic difference equation

$$X_{n+1}^N = X_n^N + f_n^N(X_n^N, u_n^N(X_n^N))\Delta^N + \varepsilon_n^N, \qquad n = 0, 1, \cdots, \Delta^{-N} - 1, \tag{2.5}$$

$$X_0^N = x. \tag{2.6}$$

In many cases, the solutions u^N obtained for the discrete problems either analytically or, approximately, by machine calculation, suggest that there exists a limit control $u(t, x)$ such that the u^N converge to u in some sense. A simple example is given in [4], where the stationary control laws $u^N(x)$ converge almost everywhere to $u(x) = -\text{sign}(x)$, which indeed turns out to be the optimal control for the continuous time problem. Actually, pointwise (a.e.) convergence can be expected to be the easiest one to handle, but to be realized only in singular cases. On the other hand, the kind of convergence which is likely to occur most often, but will at the same time be the most tough one to handle, should be some sort of weak convergence. So this is what we are going to require.

(A4) There exist a subsequence $(N') \subset (N)$ and a measurable function $u : [0, 1] \times \mathbb{R} \to \mathcal{U}$ such that for every measurable bounded function $\phi(t, x)$ with compact support

$$\begin{aligned} \lim_{N' \to \infty} \sum_{i=0}^{\Delta^{-N'}-1} \int_{i\Delta^{N'}}^{(i+1)\Delta^{N'}} ds \int \phi(s, x) f(i\Delta^{N'}, x, u_i^{N'}(x))\, dx \\ = \int_0^1 ds \int \phi(s, x) f(s, x, u(s, x))\, dx. \end{aligned}$$

This kind of convergence of control laws is a particularly appropriate one since under Roxin's condition of convex velocity sets $f(t, x, \mathcal{U})$—the usual condition guaranteeing existence of an optimal feedback control (cf. [1])—such a limit u will exist (cf. [16, Cor. 1]).

Without restricting generality, we shall assume henceforth that $(N') = (N)$. The main result to be proved is then the

THEOREM. *Under assumptions* (A1)–(A4), *the control* $u(t, x)$ *is optimal for the continuous time problem, and the values of the discrete time problems converge to the value of the continuous time problem*: $J^N(u^N) \to J(u)$.

3. Some facts about discrete exponentials. The preceding section showed that we shall have to deal with difference equations of the form

$$X_{n+1} = X_n + f_n \Delta t + \varepsilon_n, \qquad n = 0, 1, \cdots, M-1, \tag{3.1}$$

$$X_0 = x, \tag{3.2}$$

where $\Delta t = \Delta^N$, $M = \Delta^{-N}$ (for notational simplicity, since N will be fixed in this section), the ε_n are independent normal random variables with mean 0 and variance Δt, and the random variables f_n are independent of $\varepsilon_n, \cdots, \varepsilon_{M-1}$.

Let $(\mathcal{F}_n)$, $n = 0, 1, \cdots, M$, denote an increasing sequence of σ-algebras such that X_n and f_n are measurable with respect to $\mathcal{F}_n$, while ε_n is measurable with respect to $\mathcal{F}_{n+1}$ and independent of $\mathcal{F}_n$. For example, we may take $\mathcal{F}_n = \sigma\{X_0, \cdots, X_n; f_0, \cdots, f_n; \varepsilon_0, \cdots, \varepsilon_{n-1}\}$, where we agree to put $f_N \equiv \varepsilon_N \equiv 0$. Let now $\phi = (\phi_0, \cdots, \phi_{M-1})$ be any random M-vector adapted to $(\mathcal{F}_n)$. For $k = 0, 1, \cdots, M$, $n = 0, 1, \cdots, M-k$, we introduce the exponential

$$\begin{aligned} \zeta_k^{k+n}(\phi) &= \exp\left[\sum_{i=k}^{k+n-1} \phi_i \varepsilon_i - \frac{1}{2} \sum_{i=k}^{k+n-1} \phi_i^2 \Delta t\right], \\ \zeta_k^k(\phi) &= 1. \end{aligned} \tag{3.3}$$

Note that ζ_k^{k+n} is measurable with respect to $\mathcal{F}_{k+n}$. Let us list some useful properties of exponentials, which the reader will easily recognize as the discrete time analogues of well known properties of continuous time exponentials (cf. [9]).

Property 1. If $|\phi_i| \leqq C$ for all $i = 0, 1, \cdots, M-1$, then $E[\zeta_k^{k+n}(\phi)]^\alpha < \infty$ for all $\alpha > 0$ and all k, n.

This follows at once from

$$E \exp\left[\alpha \phi_i \varepsilon_i - \frac{\alpha}{2} \phi_i^2 \Delta t\right] \leqq E \exp\left[\alpha C |\varepsilon_i|\right] < \infty.$$

Property 2. Suppose that all expectations are finite. Then, for $\alpha \geqq 1$,

$$\begin{aligned} &E\{[\zeta_k^{k+n}(\phi)]^\alpha | \mathcal{F}_{k+n-1}\} \\ &\quad = [\zeta_k^{k+n-1}(\phi)]^\alpha E\left\{\exp\left[\alpha \phi_{k+n-1} \varepsilon_{k+n-1} - \frac{\alpha}{2} \phi_{k+n-1}^2 \Delta t\right] | \mathcal{F}_{k+n-1}\right\} \\ &\quad = [\zeta_k^{k+n-1}(\phi)]^\alpha e^{(\alpha(\alpha-1)/2)\phi_{k+n-1}^2 \Delta t} \\ &\quad \geqq [\zeta_k^{k+n-1}(\phi)]^\alpha, \end{aligned} \tag{3.4}$$

which means that, for all k, $([\zeta_k^{k+n}(\phi)]^\alpha, \mathcal{F}_{k+n})$, $n = 0, 1, \cdots, M-k$, is a submartingale. In particular, for $\alpha = 1$, $(\zeta_k^{k+n}(\phi), \mathcal{F}_{k+n})$ is a martingale and

$$E\{\zeta_k^{k+n}(\phi) | \mathcal{F}_k\} = 1 \tag{3.5}$$

for all admissible k, n.

Property 3. If $|\phi_i| \leqq C$ for all $i = 0, 1, \cdots, M-1$, then it follows from (3.4) that

$$E\{[\zeta_k^{k+n}(\phi)]^\alpha | \mathcal{F}_{k+n-1}\} \leqq [\zeta_k^{k+n-1}(\phi)]^\alpha e^{(\alpha(\alpha-1)/2)C^2\Delta t},$$

from which by induction

$$E\{[\zeta_k^{k+n}(\phi)]^\alpha | \mathcal{F}_k\} \leqq e^{(\alpha(\alpha-1)/2)C^2 n\Delta t} \leqq e^{(\alpha(\alpha-1)/2)C^2}. \tag{3.6}$$

Property 4. For unbounded ϕ_i, define $\phi_i^C = \phi_i \chi_{[|\phi_i| \leqq C]}$ (χ_A = indicator of A). Then $\zeta_k^{k+n}(\phi^C) \to \zeta_k^{k+n}(\phi)$ as $C \to \infty$, and, by Fatou's lemma,

$$E\{\zeta_k^{k+n}(\phi) | \mathcal{F}_k\} \leqq \liminf_{C\to\infty} E\{\zeta_k^{k+n}(\phi^C) | \mathcal{F}_k\} \leqq 1.$$

Actually, we then know that the results of Property 2 for $\alpha = 1$ are valid and (3.5) holds. In particular,

$$E\{\zeta_0^n(\phi)\} = 1 \tag{3.7}$$

for all $n = 0, 1, \cdots, M$.

On the underlying probability space $(\Omega, \mathcal{F}, P)$ consider now the measure $\tilde{P}$ defined by

$$d\tilde{P} = \zeta_0^M(\phi) dP. \tag{3.8}$$

By virtue of (3.7), $\tilde{P}$ is actually a probability measure. Let $\tilde{E}$ denote expectation with respect to $\tilde{P}$. We then have the following:

LEMMA 1. *Let the random variable g be measurable with respect to $\mathcal{F}_{n+k}$ and integrable with respect to $\tilde{P}$. Then*

$$\tilde{E}\{g | \mathcal{F}_n\} = E\{g\zeta_n^{n+k}(\phi) | \mathcal{F}_n\}.$$

Proof. For $A \in \mathcal{F}_n$

$$\begin{aligned}\int_A E\{g\zeta_n^{n+k}(\phi)|\mathcal{F}_n\}\, d\tilde{P} &= \int_A \zeta_0^n(\phi) E\{g\zeta_n^{n+k}(\phi)|\mathcal{F}_n\}\, dP \\ &= \int_A E\{g\zeta_0^{n+k}(\phi)|\mathcal{F}_n\}\, dP = \int_A g\zeta_0^{n+k}(\phi)\, dP \\ &= \int_A g\zeta_0^M(\phi)\, dP = \int_A g\, d\tilde{P}.\end{aligned}$$

□

Define random variables

$$\tilde{\varepsilon}_n = \varepsilon_n - \phi_n \Delta t, \qquad n = 0, 1, \cdots, M-1.$$

Note that $\tilde{\varepsilon}_n$ is measurable with respect to $\mathcal{F}_{n+1}$. We have the following important result.

PROPOSITION 1. *Under the measure $\tilde{P}$ the $\tilde{\varepsilon}_n$ are independent normal random variables with mean 0 and variance Δt.*

Proof. According to Lemma 1, for every real number z

$$\begin{aligned}\tilde{E}\{e^{iz\tilde{\varepsilon}_n}|\mathcal{F}_n\} &= E\{e^{iz\tilde{\varepsilon}_n}\zeta_n^{n+1}(\phi)|\mathcal{F}_n\} \\ &= E\{e^{iz(\varepsilon_n - \phi_n\Delta t)} e^{\phi_n\varepsilon_n - \phi_n^2\Delta t/2}|\mathcal{F}_n\} \\ &= e^{-iz\phi_n\Delta t - \phi_n^2\Delta t/2} E\{e^{iz\varepsilon_n + \phi_n\varepsilon_n}|\mathcal{F}_n\}.\end{aligned}$$

Since ϕ_n is $\mathcal{F}_n$-measurable and ε_n is independent of $\mathcal{F}_n$, the last conditional expectation can be calculated to equal $\exp[(iz + \phi_n)^2\Delta t/2]$. Hence the conditional characteristic

function of $\tilde{\varepsilon}_n$ under $\tilde{P}$ is

$$\tilde{E}\{e^{iz\tilde{\varepsilon}_n}|\mathcal{F}_n\} = e^{-\Delta t z^2/2}.$$

This proves the assertion. □

This result shows that under the measure $\tilde{P}$ the random vector $X = (X_0, \cdots, X_M)$ solves the difference equation

$$X_{n+1} = X_n + (f_n + \phi_n)\Delta t + \tilde{\varepsilon}_n, \qquad n = 0, 1, \cdots, M-1,$$
$$X_0 = x,$$

with independent $N(0, \Delta t)$-distributed disturbances. This is a discrete version of the Girsanov measure transformation theorem (cf. [1], [9]).

Let us now come back to the difference equations (2.5)–(2.6) of § 2. For fixed N and $x^N = (x_0^N, \cdots, x_M^N) \in \mathbb{R}^{M+1}$ (still $M = \Delta^{-N}$, $\Delta t = \Delta^N$) denote by $f^N(x^N)$ the M-dimensional vector whose components are $f_n^N(x_n^N, u_n^N(x_n^N))$ and let $Y^N = (Y_0^N, \cdots, Y_M^N)$ be the solution of the difference equation

$$Y_{n+1}^N = Y_n^N + \varepsilon_n^N, \qquad n = 0, 1, \cdots, M-1, \tag{3.9}$$

$$Y_0^N = x \tag{3.10}$$

under the measure P and $X^N = (X_0^N, \cdots, X_M^N)$ the solution of (2.5)–(2.6). Let $\mu^N = \mu_{Y^N}$ and μ_{X^N} denote the $(M+1)$-dimensional distribution of Y^N and X^N under P, respectively, and $\tilde{\mu}^N = \tilde{\mu}_{Y^N}$ the distribution of Y^N under $\tilde{P}^N$, where now

$$d\tilde{P}^N = \zeta_0^M(f^N(Y^N))\, dP.$$

Then, by the argument just given, Y^N solves (2.5)–(2.6) under $\tilde{P}^N$ with the ε_n^N replaced by

$$\tilde{\varepsilon}_n^N = \varepsilon_n^N - f_n^N(Y_n^N, u_n^N(Y_n^N))\,\Delta t,$$

and

$$\mu_{X^N} = \tilde{\mu}^N$$

since the solution of (2.5)–(2.6) is uniquely defined in distribution by the joint distribution of the disturbances. Moreover, $\tilde{\mu}^N$ is absolutely continuous with respect to μ^N with Radon–Nikodým derivative given by

$$\begin{aligned}\frac{d\tilde{\mu}^N}{d\mu^N}(x^N) &= E\{\zeta_0^M(f^N(Y^N))|Y^N = x^N\} \\ &= \exp\left[\sum_{i=0}^{M-1} f_i^N(x_i^N, u_i^N(x_i^N))\Delta x_i^N - \frac{1}{2}\sum_{i=0}^{M-1} f_i^N(x_i^N, u_i^N(x_i^N))^2 \Delta t\right], \\ &= \zeta_0^M(f^N(x^N)),\end{aligned} \tag{3.11}$$

where $\Delta x_i^N = x_{i+1}^N - x_i^N$. The way of writing used in the last equation is justified since under the measure μ^N the Δx_i^N, $i = 0, 1, \cdots, M-1$, are independent $N(0, \Delta t)$-distributed random variables on $\Omega = \mathbb{R}^{M+1}$ and $f_i^N(x_i^N, u_i^N(x_i^N))$ is independent of $\Delta x_i^N, \cdots, \Delta x_{M-1}^N$. Hence $\zeta_k^{k+n}(f^N(x^N))$ is indeed an exponential on the probability space $(\mathbb{R}^{M+1}, \mathcal{B}^{M+1}, \mu^N)$ and with respect to the natural filtration $(\mathcal{B}_n^{M+1})$ (where $\mathcal{B}_n^{M+1}$ is the σ-field generated by the first n components of x^N). We shall use this fact later on without further mentioning, in particular, the validity of (3.5), (3.6) and Lemma 1.

In the following sections, we shall work with the probability measures P and $\tilde{P}^N$ on some reference space $(\Omega, \mathcal{F})$ and with the probability measures μ^N and $\tilde{\mu}^N$ on $(\mathbb{R}^{M+1}, \mathcal{B}^{M+1})$. We shall denote expectation with respect to μ^N and $\tilde{\mu}^N$ by E^N and $\tilde{E}^N$, respectively, while the symbol E will be reserved to denote expectation with respect to P (there will be no need to denote expectation with respect to $\tilde{P}^N$).

4. Some estimates on transition probabilities. Denote by $p^N(k, x, i, \xi)$ the density of the transition probability

$$P[Y_i^N \in B | Y_k^N = x] = \mu^N[x_i^N \in B | x_k^N = x]$$

and by $\tilde{p}^N(k, x, i, \xi)$ the density of the transition probability

$$P[X_i^N \in B | X_k^N = x] = \tilde{P}^N[Y_i^N \in B | Y_k^N = x] = \tilde{\mu}^N[x_i^N \in B | x_k^N = x]$$

for $i > k$. The existence of p^N is clear, since the distribution of $(Y_1^N, \cdots, Y_{\Delta^{-N}}^N)$ given $Y_0^N = x$ is nondegenerate multivariate Gaussian. The transition density $\tilde{p}^N$ may then be calculated from p^N using formula (3.11); its actual form is, however, not very useful for our purposes.

In the next section we shall embed the discrete time problems in a continuous time framework by defining random functions $X^N(t)$, $0 \leqq t \leqq 1$, with continuous sample paths through

$$X^N(i\Delta t) = X_i^N \quad \text{for } i = 0, 1, \cdots, \Delta^{-N}, \tag{4.1}$$

interpolating linearly between neighboring grid points. In accordance with this, it will also turn out useful to have the transition densities defined for continuous time. To this end, committing a little abuse of notation, we shall use $p^N(s, x, t, \xi)$ to denote the step function whose value for $k\Delta^N \leqq s < (k+1)\Delta^N$, $i\Delta^N \leqq t < (i+1)\Delta^N$ is given by $p^N(k, x, i, \xi)$, and similarly for $\tilde{p}^N(s, x, t, \xi)$. Or, if we introduce the useful notation

$$[s]_N = k\Delta^N \quad \text{if } k\Delta^N \leqq s < (k+1)\Delta^N,$$

the above convention means that we agree to use

$$p^N(s, x, t, \xi) = p^N([s]_N, x, [t]_N, \xi)$$

and

$$p^N([s]_N\Delta^{-N}, x, [t]_N\Delta^{-N}, \xi)$$

to denote the same thing. It will always be clear from the context which interpretation is to be given to the time variables. Similarly, we shall use

$$\zeta_s^1(\phi) = \zeta_{[s]_N}^1(\phi)$$

to denote the same thing as

$$\zeta_{[s]_N\Delta^{-N}}^{\Delta^{-N}}(\phi).$$

Accordingly, we shall often switch from discrete time sample space to continuous sample functions. Therefore, with each $(\Delta^{-N}+1)$-vector $(x_0^N, \cdots, x_{\Delta^{-N}}^N)$ we associate a continuous function $x^N(t)$, $0 \leqq t \leqq 1$, defined in accordance with (4.1). In particular,

$$x^N([t]_N) = x_{[t]_N\Delta^{-N}}^N.$$

Moreover, we shall use the shorthand notation

$$E_{s,x}^N[g(x^N)] = E_{[s]_N,x}^N[g(x^N)] = E^N[g(x^N) | x^N([s]_N) = x]$$

and

$$\tilde{E}^N_{s,x}[g(x^N)] = \tilde{E}^N_{[s]_N,x}[g(x^N)] = \tilde{E}^N[g(x^N)|x^N([s]_N) = x]$$

for integrable functions $g(x^N)$.

Remark. In the sequel, conditional expectations $E^N_{s,x}$ and $\tilde{E}^N_{s,x}$ with $k\Delta^N \leqq s < (k+1)\Delta^N$ will be taken only of such functions $g(x^N)$ that depend only on the present and the future components of x^N, i.e., $g(x^N) = g(x^N_k, \cdots, x^N_{\Delta^{-N}})$. In this case,

$$E^N_{s,x}[g(x^N)] = \int p^N(k, x, k+1, x_{k+1})\, dx_{k+1} \int p^N(k+1, x_{k+1}, k+2, x_{k+2})\, dx_{k+2}$$

$$\times \cdots \times \int p^N(\Delta^{-N}-1, x_{\Delta^{-N}-1}, \Delta^{-N}, x_{\Delta^{-N}}) g(x, x_{k+1}, \cdots, x_{\Delta^{-N}})\, dx_{\Delta^{-N}}$$

$$= \int g(x^N)\, d\mu^N_{s,x},$$

where $\mu^N_{s,x}$ is the probability measure defined by the iterated integral over indicator functions $\chi_{[(x,x_{k+1},\cdots,x_{\Delta^{-N}})\in A]}$, and a simlar formula holds for $\tilde{E}^N_{s,x}$. Henceforth we shall always mean these versions of the conditional expectations.

The first thing we need is a powerful estimate for the solutions of difference equations.

LEMMA 2 (discrete Gronwall–Bellman inequality). *Let α_i, ϕ_i and L be nonnegative numbers such that*

$$\phi_{n+1} \leqq \alpha_n + L \sum_{i=0}^{n} \phi_i, \qquad n = 0, 1, \cdots, N-1.$$

Then

$$\phi_{n+1} \leqq (1+L)^n (\phi_0 L + \|\alpha\|_n)$$

for all $n = 0, 1, \cdots, N-1$ (with $\|\alpha\|_n = \max\{\alpha_0, \cdots, \alpha_n\}$).

The proof is easily done by induction.

Next we derive a uniform L_p-bound for discrete exponentials corresponding to admissible drifts.

LEMMA 3. *Let $f(t, x, u)$ satisfy the linear growth condition* (A2). *Then there exist numbers $p > 1$, $\alpha > 0$ and $K_f > 0$ such that for all $N = 0, 1, 2, \cdots$, and every admissible simple strategy $u = (u_0, \cdots, u_{M-1}) \in \mathcal{S}^N$ (with $M = \Delta^{-N}$ as in* § 3)

$$E^N_{s,x}\{[\zeta^1_s(f^u(x^N))]^p\} \leqq K_f e^{\alpha|x|^2} \quad \text{for all } 0 \leqq s \leqq 1,$$

where $f^u(x^N)$ is the M-vector with components $f^u_i(x^N) = f^N_i(x^N_i, u_i(x^N_i))$, $i = 0, 1, \cdots, M-1$.

Proof. The proof is similar to the one in [1] for the continuous time case. For $1 < p \leqq 2$, put $\gamma = C_f(p^2-p)/2$ and let $[s]_N = k\Delta^N$. Denote by z^N_n, $n = k, \cdots, M$, the process determined by

$$\begin{aligned} \Delta z^N_n &= \Delta x^N_n - p f^u_n(x^N)\Delta t, \qquad n = k, \cdots, M-1, \\ z^N_k &= x, \end{aligned} \tag{4.2}$$

with $\Delta z^N_n = z^N_{n+1} - z^N_n$ and $\Delta t = \Delta^N$. Then, under the probability measure $\mu^{N,p}_{s,x}$ on $(R^{M+1}, \mathcal{B}^{M+1})$ defined by

$$d\mu^{N,p}_{s,x} = \zeta^1_s(p f^u(x^N)) d\mu^N_{s,x}$$

the Δz_n^N are independent $N(0, \Delta t)$-distributed random variables (cf. Proposition 1). We then have

$$E_{s,x}^N\{[\zeta_s^1(f^u(x^N))]^p\}$$
$$= E_{s,x}^N\left\{\zeta_s^1(pf^u(x^N))\exp\left[\frac{p^2-p}{2}\sum_{n=k}^{M-1} f_n^N(x_n^N, u_n(x_n^N))^2\Delta t\right]\right\}$$
$$\leqq e^{\gamma}E_{s,x}^N\left\{\zeta_s^1(pf^u(x^N))\exp\left[\gamma\sum_{n=k}^{M-1}|x_n^N|^2\Delta t\right]\right\}$$

by virtue of the growth condition. But, from (4.2),

$$x_n^N = z_n^N + p\sum_{i=k}^{n-1} f_i^u(x^N)\Delta t, \qquad n = k+1, \cdots, M,$$

whence

$$|x_n^N|^2 \leqq 2|z_n^N|^2 + 2p^2\sum_{i=k}^{n-1}|f_i^u(x^N)|^2\Delta t$$
$$\leqq 2(|z_n^N|^2 + 4C_f) + 8C_f\Delta t\sum_{i=k}^{n-1}|x_i^N|^2,$$

again by the growth condition and since $p^2 \leqq 4$. By Lemma 2,

$$|x_n^N|^2 \leqq (1+8C_f\Delta t)^{n-k-1}[8C_f\Delta t|x|^2 + 8C_f + 2\max_{i=k,\cdots,n-1}|z_i^N|^2]$$

for $n = k+1, \cdots, M$. But then

$$\sum_{n=k}^{M-1}|x_n^N|^2\Delta t \leqq \kappa[1+|x|^2 + \max_{n=k,\cdots,M}|z_n^N|^2]$$

for some constant κ independent of N. Hence

$$E_{s,x}^N\{[\zeta_s^1(f^u(x^N))]^p\}$$
$$\leqq e^{\gamma}e^{\gamma\kappa(1+|x|^2)}E_{s,x}^{N,p}\{\exp\gamma\kappa\max_{n=k,\cdots,M}|z_n^N|^2\}$$
$$\leqq e^{\gamma(1+\kappa)}e^{3\gamma\kappa|x|^2}E_{s,x}^{N,p}\{\exp[2\gamma\kappa\max_{n=k,\cdots,M}|z_n^N - x|^2]\},$$

where $E_{s,x}^{N,p}$ denotes expectation with respect to $\mu_{s,x}^{N,p}$. Since, under $\mu_{s,x}^{N,p}$, each $(z_n^N - x)$ is the sum of $(n-k)$ independent $N(0, \Delta t)$-distributed increments Δz_i^N, the last expectation is bounded by some finite constant A independent of N, provided $p > 1$ is chosen sufficiently small to make γ near 0. Put $\alpha = 3\gamma\kappa$ and $K_f = Ae^{\gamma(1+\kappa)}$ to get the assertion. □

We shall now derive some estimates on transition densities.

PROPOSITION 2. *For every $0 \leqq k < i \leqq M$ and every positive real number R*

$$\int_{|\xi-x|>R} p^N(k, x, i, \xi)\, d\xi \leqq K_1 e^{-R},$$

and, with p, α chosen as in Lemma 3, $1/p + 1/q = 1$,

$$\int_{|\xi-x|>R} \tilde{p}^N(k, x, i, \xi)\, d\xi \leqq K_2 e^{\alpha|x|^2/p}e^{-R/q},$$

where the constants K_1 and K_2 are independent of N, k, i and x.

Proof. Note that the conditional distribution of $Y_i^N - x$ given $Y_k^N = x$ under P is the same as the distribution of $\sum_{j=k}^{i} \varepsilon_j^N$. Hence the term to be estimated is

$$P[|Y_i^N - x| > R | Y_k^N = x] = 2P\left[\sum_{j=k}^{i} \varepsilon_j^N > R\right].$$

But

$$P\left[\sum_{j=k}^{i} \varepsilon_j^N > R\right] = P\left[\exp\left(\sum_{j=k}^{i} \varepsilon_j^N - \frac{i-k}{2}\Delta^N\right) > \exp\left(R - \frac{i-k}{2}\Delta^N\right)\right]$$
$$= P[\zeta_k^{i+1}(1) > e^{R-(i-k)\Delta^N/2}] \leqq e^{1/2} e^{-R}.$$

by Chebyshev's L^1-inequality and (3.5). Repeating with $-\sum_{j=k}^{i} \varepsilon_j^N$ yields the first assertion with $K_1 = 2e^{1/2}$. By Lemmas 1 and 3

$$\tilde{P}^N[|Y_i^N - x| > R | Y_k^N = x] = E[\zeta_k^M(f^N(Y^N))\chi_{[|Y_i^N - x| > R]} | Y_k^N = x]$$
$$\leqq [K_f e^{\alpha|x|^2}]^{1/p} \cdot P[|Y_i^N - x| > R | Y_k^N = x]^{1/q}. \qquad \square$$

In the following lemmas we shall make use of the continuous time notation introduced at the beginning of this section.

LEMMA 4. *For $t > s$, there exists a function $\Phi_{t-s}(|h|)$ such that*

$$\int [p^N(s, x, t, \xi) - p^N(s, x, t, \xi + h)]^2 \, d\xi \leqq \Phi_{t-s}(|h|)$$

for all x and all N satisfying[1] *$\Delta^N < (t-s)/2$, and $\Phi_{t-s}(|h|) \to 0$ as $h \to 0$.*

Proof. With $[s]_N = k\Delta^N$, $[t]_N = i\Delta^N$, the term to be estimated becomes

$$\int [p^N(k, x, i, \xi) - p^N(k, x, i, \xi + h)]^2 \, d\xi$$

$$= [2\pi(i-k)\Delta^N]^{-1} \int \left\{\exp\left[-\frac{(\xi - x)^2}{2(i-k)\Delta^N}\right] - \exp\left[-\frac{(\xi + h - x)^2}{2(i-k)\Delta^N}\right]\right\}^2 d\xi$$

$$= \frac{1}{2}[\pi(i-k)\Delta^N]^{-1/2}$$

$$\times \left\{[2\pi(i-k)\Delta^N/2]^{-1/2} \int \left\{\exp\left[-\frac{(\xi - x)^2}{2(i-k)\Delta^N/2}\right] + \exp\left[-\frac{(\xi + h - x)^2}{2(i-k)\Delta^N/2}\right]\right\} d\xi\right.$$

$$\left. - 2[2\pi(i-k)\Delta^N/2]^{-1/2} \right.$$

$$\left. \times \int \exp\left[-\frac{(\xi - x)^2}{2(i-k)\Delta^N/2}\right] \exp\left[-\frac{h(\xi - x)}{(i-k)\Delta^N}\right] d\xi \exp\left[-\frac{h^2}{2(i-k)\Delta^N}\right]\right\}$$

$$= [\pi(i-k)\Delta^N]^{-1/2}\left(1 - \exp\left[-\frac{h^2}{4(i-k)\Delta^N}\right]\right).$$

Since, for $\Delta^N < (t-s)/2$, $(i-k)\Delta^N > (t-s)/2$, the last expression is smaller than

$$\Phi_{t-s}(h) = \sqrt{2}[\pi(t-s)]^{-1/2}(1 - e^{-h^2/(2(t-s))}). \qquad (4.3) \qquad \square$$

[1] A condition like $\Delta^N < t - s$ must be imposed to avoid $[s]_N = [t]_N$ and hence singularity of p^N. We choose the above for esthetical reasons: it makes the normal density appear in the bounds.

LEMMA 5. *For $t>s$, there exists a function $\Psi_{t-s}(|h|)$ such that*

$$\int |p^N(s,x,t,\xi)-p^N(s,x,t,\xi+h)|\,d\xi \leqq \Psi_{t-s}(|h|)$$

for all x and all N satisfying $\Delta^N<(t-s)/2$, and $\Psi_{t-s}(|h|)\to 0$ as $h\to 0$.

Proof. For an arbitrary real number R we have

$$\int |\Delta p^N(\xi)|\,d\xi = \int_{|\xi-x|\leqq R} |\Delta p^N(\xi)|\,d\xi + \int_{|\xi-x|>R} |\Delta p^N(\xi)|\,d\xi = I_1+I_2,$$

where we have put $\Delta p^N(\xi)=p^N(s,x,t,\xi)-p^N(s,x,t,\xi+h)$ for a moment. By Proposition 2,

$$I_2 \leqq K_1(e^{-R}+e^{-R}e^{|h|}) \leqq K_3 e^{-R}$$

for $|h|<\min(R,1)$. As to the first term,

$$I_1 \leqq \sqrt{2R}\left(\int |\Delta p^N(\xi)|^2\,d\xi\right)^{1/2}.$$

With the special choice $R=1/\sqrt{|h|}$, it follows from Lemma 4 and formula (4.3) that for $\Delta^N<(t-s)/2$

$$I_1+I_2 \leqq K_3 e^{-1/\sqrt{|h|}} + \left(\frac{2\sqrt{2}}{\sqrt{\pi(t-s)}}\,\frac{1-e^{-h^2/2(t-s)}}{\sqrt{|h|}}\right)^{1/2} =: \Psi_{t-s}(|h|). \tag{4.4}$$

□

LEMMA 6. *There exists a function $\psi(|h|)$ such that*

$$\int_{(s+2\Delta^N)\wedge 1}^{1} du \int |p^N(s,x,u,\xi)-p^N(s,x,u,\xi+h)|\,d\xi \leqq \psi(|h|)$$

uniformly in N, x and s, and $\psi(|h|)\to 0$ as $h\to 0$.

Proof. From Lemma 5 and formula (4.4),

$$\begin{aligned}
&\int_{(s+2\Delta^N)\wedge 1}^{1} du \int |p^N(s,x,u,\xi)-p^N(s,x,u,\xi+h)|\,d\xi \\
&\quad \leqq \int_{(s+2\Delta^N)\wedge 1}^{1} du\left\{K_3 e^{-1/\sqrt{|h|}} + \left(\frac{2\sqrt{2}}{\sqrt{\pi(u-s)}}\,\frac{1-e^{-h^2/2(u-s)}}{\sqrt{|h|}}\right)^{1/2}\right\} \\
&\quad \leqq K_3 e^{-1/\sqrt{|h|}} + K_4 \int_0^1 dv \left(\frac{1}{\sqrt{v}}\,\frac{1-e^{-h^2/2v}}{\sqrt{|h|}}\right)^{1/2}.
\end{aligned}$$

Substituting $w=1/\sqrt{v}$ we obtain

$$\int_0^1 dv\,\frac{1}{\sqrt{v}}\,\frac{1-e^{-h^2/2v}}{\sqrt{|h|}} \leqq 2\int_0^\infty dw\,\frac{1}{w^2}\,\frac{1-e^{-h^2w^2/2}}{\sqrt{|h|}} = \frac{2}{\sqrt{|h|}}|h|\sqrt{\pi/2} = 2\sqrt{|h|}\sqrt{\pi/2}$$

(cf. [10]). Hence the function

$$\psi(|h|) = K_3 e^{-1/\sqrt{|h|}} g + K_4[2\sqrt{|h|}\sqrt{\pi/2}]^{1/2} \tag{4.5}$$

will meet our requirements. □

The following lemma is an easy consequence of Lemma 6.

LEMMA 7. *Let g be any bounded measurable function on $\mathbb{R}^2$ and put $g_h(s,\xi)=g(s,\xi-h)$. Then, for every s and x,*

$$\int_s^1 du\,|E^N_{s,x}[g(u,x^N([u]_N))-g_h(u,x^N([u]_N))]|\leqq\|g\|(\psi(|h|)+4\Delta^N)$$

uniformly in N (with $\|\cdot\|=$ ess-sup).

The next lemma shows that a similar estimate is true for the tilded measures.

LEMMA 8. *There exists a function $\tilde\psi(|h|)$ such that for any bounded measurable function g and every (s,x)*

$$\left|\int_s^1 du\tilde E^N_{s,x}[g(u,x^N([u]_N))-g_h(u,x^N([u]_N))]\right|\leqq K_5 e^{\alpha|x|^2/p}\|g\|(\tilde\psi(|h|)+\Delta^{N/q})$$

uniformly in N, and $\tilde\psi(|h|)\to 0$ as $h\to 0$. Here α and the conjugate pair of numbers (p,q) are determined as in Lemma 3.

Proof. Remember our continuous time notation for the discrete exponentials and put $\Delta g^N(u)=g(u,x^N([u]_N))-g_h(u,x^N([u]_N))$. Then, using a (p,q)-Hölder estimate, we obtain from Lemmas 1 and 3 that

$$\begin{aligned}\left|\int_s^1 du\tilde E^N_{s,x}[\Delta g^N(u)]\right|&=\left|\int_s^1 du E^N_{s,x}[\zeta^1_s(f^N(x^N))\Delta g^N(u)]\right|\\&=\left|E^N_{s,x}[\zeta^1_s(f^N(x^N))\int_s^1\Delta g^N(u)\,du]\right|\\&\leqq[K_f e^{\alpha|x|^2}]^{1/p}\left\{E^N_{s,x}\left|\int_s^1\Delta g^N(u)\,du\right|^q\right\}^{1/q}\end{aligned}$$

and (noting that $q>2$)

$$E^N_{s,x}\left|\int_s^1\Delta g^N(u)\,du\right|^q\leqq(2\|g\|)^{q-2}E^N_{s,x}\left[\int_s^1\Delta g^N(u)\,du\right]^2.$$

Since, by partial integration, $[\int_a^b h(u)\,du]^2=2\int_a^b h(u)\,du\int_u^b h(v)\,dv$, we obtain

$$\begin{aligned}&E^N_{s,x}\left[\int_s^1\Delta g^N(u)\,du\right]^2\\&\quad=2E^N_{s,x}\left(\int_s^1 du E^N_{s,x}\left\{\Delta g^N(u)\int_u^1\Delta g^N(v)\,dv|x^N_0,\cdots,x^N([u]_N)\right\}\right)\\&\quad=2E^N_{s,x}\left(\int_s^1\Delta g^N(u)\,du E^N_{u,x^N([u]_N)}\left[\int_u^1\Delta g^N(v)\,dv\right]\right)\\&\quad\leqq 2\|g\|(\psi(|h|)+4\Delta^N)\cdot E^N_{s,x}\int_0^1|\Delta g^N(u)|\,du\\&\quad\leqq 4\|g\|^2(\psi(|h|)+4\Delta^N),\end{aligned}$$

since $\Delta g^N(u)$ is measurable with respect to $\sigma\{x^N_0,\cdots,x^N([u]_N)\}$, and the coordinate mapping $x^N\to x^N_i$, $i=0,\cdots,\Delta^{-N}$, is a Markov process under μ^N, and upon using the estimate of Lemma 7 for $E^N_{u,x}[\int_u^1\Delta g^N(v)\,dv]$. Thus the assertion holds with

$$\tilde\psi(|h|)=\psi(|h|)^{1/q}\quad\text{and}\quad K_5=2^q K_f^{1/p}.$$ □

From this result we finally arrive at

PROPOSITION 3. *For every s and x,*

$$\int_{(s+2\Delta^N)\wedge 1}^{1} du \int |\tilde{p}^N(s,x,u,\xi)-\tilde{p}^N(s,x,u,\xi+h)|\,d\xi \leqq K_6 e^{\alpha|x|^2/p}[\tilde{\psi}(|h|)+\Delta^{N/q}]$$

uniformly in N.

Proof. From Lemma 8 we find that for any bounded measurable g

$$\left|\int_{(s+2\Delta^N)\wedge 1}^{1} du\, \tilde{E}^N_{s,x}[g(u,x^N([u]_N))-g_h(u,x^N([u]_N))]\right|$$

$$=\left|\int_{(s+2\Delta^N)\wedge 1}^{1} du \int [\tilde{p}^N(s,x,u,\xi)-\tilde{p}^N(s,x,u,\xi+h)]g(u,\xi)\,d\xi\right|$$

$$\leqq K_5 e^{\alpha|x|^2/p}\|g\|[\tilde{\psi}(|h|)+\Delta^{N/q}]+4\|g\|\Delta^N.$$

For N fixed, choose $g(u,\xi)$ to be the indicator of the set $[\tilde{p}^N(s,x,u,\xi)-\tilde{p}^N(s,x,u,\xi+h)\geqq 0]$. Then, from the relation above,

$$\int_{(s+2\Delta^N)\wedge 1}^{1} du \int [\tilde{p}^N(s,x,u,\xi)-\tilde{p}^N(s,x,u,\xi+h)]^+\,d\xi$$

$$\leqq e^{\alpha|x|^2/p}[K_5\tilde{\psi}(|h|)+(K_5+4)\Delta^{N/q}].$$

Repeating with g equal to the indicator of the set $[\tilde{p}^N(s,x,u,\xi)-\tilde{p}^N(s,x,u,\xi+h)<0]$ gives the desired result. □

5. Proof of the theorem. Coming back to our difference equation (2.5)–(2.6), we write the solution in the form

$$X^N_{n+1}=x+\sum_{i=0}^{n} f^N_i(X^N_i,u^N_i(X^N_i))\Delta^N+\sum_{i=0}^{n}\varepsilon^N_i$$

and obtain from Lemma 2, for f satisfying (A2), the estimate

$$|X^N_{n+1}|\leqq\left(|x|\sqrt{C_f}+|x|+\sqrt{C_f}+\max_{i=0,\cdots,n}\left|\sum_{j=0}^{i}\varepsilon^N_j\right|\right)(1+\sqrt{C_f}\Delta^N)^n$$

from which

$$E|X^N_{n+1}|^4\leqq K_7[1+(n\Delta^N)^2](1+\sqrt{C_f}\,\Delta^N)^{4n}$$

since $E[\max_{i=0,\cdots,n}|\sum_{j=0}^{i}\varepsilon^N_j|]^4=E[\max_{i=0,\cdots,n}|\sum_{j=0}^{i}\varepsilon^N_j|^4]\leqq(\frac{4}{3})^4E|\sum_{j=0}^{n}\varepsilon^N_j|^4$. Since the constant K_7 is independent of n and N, finally

$$E|X^N_n|^4\leqq K_8 \tag{5.1}$$

uniformly in $N=0,1,\cdots,n=0,1,\cdots,\Delta^{-N}$.

In accordance with (4.1), define processes $F^N(t)$ and $B^N(t)$, $0\leqq t\leqq 1$, with continuous sample paths by

$$F^N(t)=\int_0^t f([s]_N,X^N([s]_N),u^N(s,X^N([s]_N)))\,ds$$

and

$$B^N(t)=\sum_{i=0}^{[t]_N\Delta^{-N}-1}\varepsilon^N_i+(t-[t]_N)\Delta^{-N}\varepsilon^N_{[t]_N\Delta^{-N}}.$$

Here $u^N(s,\xi)$ denotes the step function (in s) taking the value $u_i^N(\xi)$ for $i\Delta^N \leqq s < (i+1)\Delta^N$. Then the difference equations (2.5)–(2.6) may be written in the compact form

$$X^N(t) = x + F^N(t) + B^N(t). \tag{5.2}$$

From a little exercise in Hölder estimates it then follows that

$$\begin{aligned} E|F^N(t)-F^N(s)|^4 &\leqq (t-s)^3 \int_s^t E|f([r]_N, X^N([r]_N), u^N(r, X^N([r]_N)))|^4\, dr \\ &\leqq (t-s)^3 \int_s^t 3C_f^2 E(1+|X^N([r]_N)|^4)\, dr \\ &\leqq 3C_f^2(1+K_8)(t-s)^4. \end{aligned} \tag{5.3}$$

As to the process $B^N(t)$, it follows from a simple geometrical consideration that for $t<s$

$$\begin{aligned} |B^N(s)-B^N([t]_N+\Delta^N)| \leqq \max\{&|B^N([s]_N)-B^N([t]_N+\Delta^N)|, \\ &|B^N([s]_N+\Delta^N)-B^N([t]_N+\Delta^N)|\}. \end{aligned}$$

Hence, for fixed t and $h>0$, denoting $k=[t]_N\Delta^{-N}$ and $l=([t+h]_N+\Delta^N)\Delta^{-N}$ for a moment,

$$\sup_{t\leqq s\leqq t+h} |B^N(s)-B^N(t)|^4 \leqq 2^4\left[\max_{k<i\leqq l}\left|\sum_{j=k+1}^{i}\varepsilon_j^N\right|^4 + |\varepsilon_k^N|^4\right].$$

Since $\sum_{j=k+1}^{i}\varepsilon_j^N$, $i=k+1,\cdots,l$, is a martingale,

$$E\left\{\max_{k<i\leqq l}\left|\sum_{j=k+1}^{i}\varepsilon_j^N\right|^4\right\} \leqq \left(\frac{4}{3}\right)^4 E\left|\sum_{j=k+1}^{l}\varepsilon_j^N\right|^4.$$

Calculating the last expectation we obtain

$$E\sup_{t\leqq s\leqq t+h}|B^N(s)-B^N(t)|^4 \leqq K_9[(l-k)^2+1]\Delta^{2N} \leqq K_9(h^2+4h\Delta^N+5\Delta^{2N}).$$

Hence, given $\varepsilon>0$, $\eta>0$, we may first choose h so small and then N_0 so large that

$$\varepsilon^{-4}K_9(h+4\Delta^N+5\Delta^{2N}/h) \leqq \eta \quad \text{for all } N \geqq N_0,$$

thus achieving

$$P\left[\sup_{t\leqq s\leqq t+h}|B^N(s)-B^N(t)| \geqq \varepsilon\right] \leqq \eta h$$

for all $N\geqq N_0$ and all t. Together with (5.3) and (5.2) this implies that the sequence of random functions (X^N, F^N, B^N), $N=1,2,\cdots$, is tight in C^3 (C = space of continuous functions on [0, 1]) (cf. [2], [13]). Hence, by Skorokhod's embedding theorem we may assume (after passing to a subsequence and changing the probability space) that there exists a random function (X,F,B) with values in C^3 such that

$$(X^N, F^N, B^N) \to (X, F, B) \quad \text{a.e.} \tag{5.4}$$

in the topology of uniform convergence (for a more detailed description of this technique see [3], [12], [13]; the use of N to denote the convergent subsequence will be justified shortly). Moreover,

$$X(t) = x + F(t) + B(t) \tag{5.5}$$

holds with probability one for all t, and $B(t)$ is easily recognized as Brownian motion. Also, by Fatou's lemma, the estimates (5.1) and (5.3) carry over to the limits $X(t)$ and $F(t)$:

$$E|X(t)|^4 \leqq K_8 \quad \text{and} \quad E|F(t)-F(s)|^4 \leqq 3C_f^2(1+K_8)(t-s)^4.$$

Further, $F(t)$ is absolutely continuous:

$$F(t,\omega) = \int_0^t \bar{f}(s,\omega)\,ds$$

for some integrable function $\bar{f}(s,\omega)$, and $\mathscr{F}_s = \mathscr{F}_s^X \vee \mathscr{F}_s^F \vee \mathscr{F}_s^B$ is independent of the increments $B(t)-B(s)$ for $s<t$. To see this, take any two bounded continuous functions g and h. Then

$$\begin{aligned}
&E\{g[X(s),F(s),B(s)]h[B(t)-B(s)]\}\\
&\quad = \lim_{N\to\infty} E\{g[X^N([s]_N),F^N([s]_N),B^N([s]_N)]h[B^N([t]_N)-B^N([s]_N+\Delta^N)]\}\\
&\quad = \lim_{N\to\infty} E\{g[X^N([s]_N),F^N([s]_N),B^N([s]_N)]\}\cdot E\{h[B^N([t]_N)-B^N([s]_N+\Delta^N)]\}\\
&\quad = E\{g[X(s),F(s),B(s)]\}\cdot E\{h[B(t)-B(s)]\}.
\end{aligned}$$

This argument immediately extends to functions $g[X(s_1),\cdots,X(s_n),F(s_1),\cdots,F(s_n),B(s_1),\cdots,B(s_n)]$, $s_1<s_2<\cdots<s_n\leqq s$, thus proving the assertion.

As a consequence, we may take $\bar{f}$ adapted to $(\mathscr{F}_t)$ and X satisfying the stochastic differential equation

$$dX = \bar{f}\,dt + dB, \tag{5.6}$$

is an Ito process (cf. [14]). Moreover, since

$$|\bar{f}(t,\omega)|^4 = \lim_{h\downarrow 0} h^{-4}|F(t+h,\omega)-F(t,\omega)|^4$$

for all (t,ω) in a set A, $\lambda\times P(A)=1$, it follows again by Fatou's lemma that

$$E\int_0^1 |\bar{f}(t,\omega)|^4\,dt \leqq 3C_f^2(1+K_8),$$

so that X satisfying (5.6) is indeed an Ito process whose distribution measure is absolutely continuous with respect to Wiener measure (cf. [14, Thm. 7.2]).
The crucial step in the analysis is now to show that for all times $s\leqq t$

$$\begin{aligned}
&\int_s^t f([r]_N, X^N([r]_N), u^N(r, X^N([r]_N)))\,dr\\
&\qquad \to \int_s^t f(r,X(r),u(r,X(r)))\,dr \quad \text{as } N\to\infty
\end{aligned} \tag{5.7}$$

in probability. Here $u(r,\xi)$ is the function appearing in assumption (A4). Then it will follow from (5.4) and (5.5) that

$$F(t) = \int_0^t f(r,X(r),u(r,X(r)))\,dr$$

for all t with probability one and hence

$$dX = f(t, X(t), u(t, X(t)))\, dt + dB. \tag{5.8}$$

This is the desired result, saying that the limit process $X(t)$ is a solution of our stochastic differential equation corresponding to the limit control u determined by (A4).

Remark. Starting from $X^N = (X_0^N, \cdots, X_{\Delta^{-N}}^N)$ defined by (2.5)–(2.6) (on any probability space) and defining the $X^N(t)$ by linear interpolation, the above reasoning shows that every weakly convergent subsequence of the $X^N(\cdot)$ has the same limit, namely the unique distribution measure (on C) of the solution to (5.8). Since, by tightness, every subsequence possesses a further subsequence which converges weakly to this limit, it follows readily that the sequence $X^N(\cdot)$ itself converges in distribution to the solution of (5.8).

We now come to the

Proof of (5.7). For shortness, let us introduce the notation $g(r, \xi) = f(r, \xi, u(r, \xi))$ and $g^N(r, \xi) = f([r]_N, \xi, u^N(r, \xi))$. Take a C^∞-function ϕ with compact support such that $0 \leqq \phi \leqq 1$, $\phi(-\xi) = \phi(\xi)$ and $\int \phi(\xi)\, d\xi = 1$. For $\varepsilon > 0$, define $\phi_\varepsilon(\xi) = \varepsilon^{-1}\phi(\xi/\varepsilon)$ and set

$$g_\varepsilon^N(r, \xi) = \int \phi_\varepsilon(\xi - \eta) g^N(r, \eta)\, d\eta$$

and

$$g_\varepsilon(r, \xi) = \int \phi_\varepsilon(\xi - \eta) g(r, \eta)\, d\eta.$$

Observing that the growth condition (A2) carries over to the functions $g^N(s, x)$, $g_\varepsilon^N(s, X)$ and $g_\varepsilon(s, x)$ with a constant independent of N and ε (for ε small enough), we can see that the random variables $g^N(s, X^N([s]_N))$, $g_\varepsilon^N(s, X^N([s]_N))$ and $g_\varepsilon(s, X(s))$ have uniformly (in N and ε) bounded fourth moments with respect to the product measure $ds \times dP$. Then it follows from assumption (A4) that for each $\varepsilon > 0$ and each continuous function ψ having compact support

$$\int \psi(\xi) \left(\int_s^t g_\varepsilon^N(r, \xi)\, dr \right) d\xi = \int_s^t dr \int \psi_\varepsilon(\xi) g^N(r, \xi)\, d\xi$$

$$\to \int_s^t dr \int \psi_\varepsilon(\xi) g(r, \xi)\, d\xi = \int \psi(\xi) \left(\int_s^t g_\varepsilon(r, \xi)\, dr \right) d\xi,$$

where $\psi_\varepsilon = \phi_\varepsilon * \psi$. From this it follows, arguing with the Arzela–Ascoli theorem as in [15, proof of Lem. 11.4.1], that for all $s \leqq t$ and fixed $\varepsilon > 0$

$$\int_s^t g_\varepsilon^N(r, \xi)\, dr \to \int_s^t g_\varepsilon(r, \xi)\, dr$$

boundedly and uniformly in ξ on compact sets. For fixed sample paths $x^N(t) = X^N(t, \omega)$, $x(t) = X(t, \omega)$ and $h > 0$ we can find a partition $s = \tau_1 < \cdots < \tau_{k(h)} = t$ such that

$$\sup_{i = 1, \cdots, k(h)-1} \sup_{r, r' \in [\tau_i, \tau_{i+1}]} |x^N([r]_N) - x(r')| \leqq h$$

for all $N \geqq N(h)$. This follows from the uniform convergence of $x^N(\cdot)$ to $x(\cdot)$ together with the uniform continuity of $x(\cdot)$. Moreover, $|g_\varepsilon^N(r, \xi) - g_\varepsilon^N(r, \xi + h)| \leqq K_\varepsilon |h|$

uniformly in r and ξ on compact sets (with the constants K_ε depending on the set), and a similar estimate holds for $g_\varepsilon(r, \xi)$. Hence

$$\begin{aligned}
&\left|\int_s^t g_\varepsilon^N(r, x^N([r]_N))\,dr - \int_s^t g_\varepsilon(r, x(r))\,dr\right| \\
&\quad \leqq \sum_{i=1}^{k(h)-1} \int_{\tau_i}^{\tau_{i+1}} |g_\varepsilon^N(r, x^N([r]_N)) - g_\varepsilon^N(r, x(\tau_i))|\,dr \\
&\quad\quad + \sum_{i=1}^{k(h)-1} \left| \int_{\tau_i}^{\tau_{i+1}} g_\varepsilon^N(r, x(\tau_i))\,dr - \int_{\tau_i}^{\tau_{i+1}} g_\varepsilon(r, x(\tau_i))\,dr \right| \\
&\quad\quad + \sum_{i=1}^{k(h)-1} \int_{\tau_1}^{\tau_{i+1}} |g_\varepsilon(r, x(\tau_i)) - g_\varepsilon(r, x(r))|\,dr \\
&\quad \leqq 3(t-s)K_\varepsilon h + \text{second term.}
\end{aligned}$$

But the second term converges to 0 as $N \to \infty$ for each $h > 0$. Taking $h \downarrow 0$ yields

$$\lim_{N\to\infty} \int_s^t g_\varepsilon^N(r, X^N([r]_N, \omega))\,dr = \int_s^t g_\varepsilon(r, X(r, \omega))\,dr$$

for almost every ω.

Hence, in order to prove (5.6), it remains to show that

$$\int_s^t g_\varepsilon(r, X(r))\,dr \to \int_s^t g(r, X(r))\,dr \quad \text{as } \varepsilon \downarrow 0 \tag{5.9}$$

in probability and

$$\limsup_{\varepsilon\downarrow 0} \limsup_{N\to\infty} P\left[\left|\int_s^t g^N(r, X^N([r]_N))\,dr - \int_s^t g_\varepsilon^N(r, X^N([r]_N))\,dr\right| \geqq \delta\right] = 0 \tag{5.10}$$

for all $\delta > 0$.

In order to show (5.9), note that, for each r, $g_\varepsilon(r, \xi) \to g(r, \xi)$ for all ξ not in a set Γ_r of Lebesgue measure 0 (cf. [11]). Consider the set S of points (r, ω) where $g_\varepsilon(r, X(r, \omega)) \not\to g(r, X(r, \omega))$. S is measurable and its r-sections S_r are contained in $X(r, \cdot)^{-1}(\Gamma_r)$. Since X is an Ito process satisfying (5.6), its finite dimensional distributions are absolutely continuous with respect to Lebesgue measure; hence $P(S_r) = 0$ for all r. It follows that $g_\varepsilon(r, X(r, \omega)) \to g(r, X(r, \omega))$ for almost all (r, ω), whence (5.9) by bounded convergence.

The estimate (5.10) is more delicate. Put $A_{N,K} = \{(r, \xi) : |g^N(r, \xi)| \leqq K\}$, $g^{N,K}(r, \xi) = g^N(r, \xi)\chi_{A_{N,K}}(r, \xi)$ and $g_\varepsilon^{N,K}(r, \xi) = (g^{N,K})_\varepsilon(r, \xi) = \phi_\varepsilon * g^{N,K}(r, \cdot)(\xi)$. Then

$$\begin{aligned}
\Delta(N, \varepsilon) &= E\left\{\int_s^t [g^N(r, X^N([r]_N)) - g_\varepsilon^N(r, X^N([r]_N))]\,dr\right\}^2 \\
&\leqq 9E\left\{\int_s^t [g^{N,K}(r, X^N([r]_N)) - g_\varepsilon^{N,K}(r, X^N([r]_N))]\,dr\right\}^2 \\
&\quad + 9E\left\{\int_s^t [g^N(r, X^N([r]_N)) - g^{N,K}(r, X^N([r]_N))]\,dr\right\}^2 \\
&\quad + 9E\left\{\int_s^t [g_\varepsilon^{N,K}(r, X^N([r]_N)) - g_\varepsilon^N(r, X^N([r]_N))]\,dr\right\}^2 \\
&= 9(J_1 + J_2 + J_3).
\end{aligned}$$

Putting $\nabla_\varepsilon^{N,K} g(r,\xi) = g^{N,K}(r,\xi) - g_\varepsilon^{N,K}(r,\xi)$ for a moment and changing to sample space notation, we write (with x being the fixed initial value (1.2))

$$
\begin{aligned}
J_1 &= \tilde{E}^N \Big\{ \int_s^t \nabla_\varepsilon^{N,K} g(r, x^N([r]_N))\, dr \Big\}^2 \\
&= 2\tilde{E}^N \Big\{ \int_s^t \nabla_\varepsilon^{N,K} g(r, x^N([r]_N))\, dr\, \tilde{E}^N_{r,x^N([r]_N)} \int_r^t \nabla_\varepsilon^{N,K} g(u, x^N([u]_N))\, du \Big\} \\
&\leqq 4K\tilde{E}^N \int_s^t dr \Big\{ \chi_{[|z-x|>K]}(x^N([r]_N)) |\tilde{E}^N_{r,x^N([r]_N)} \int_r^t \nabla_\varepsilon^{N,K} g(u, x^N([u]_N))\, du \\
&\qquad + \chi_{[|z-x|\leqq K]}(x^N([r]_N)) |\tilde{E}^N_{r,x^N([r]_N)} \int_r^t \nabla_\varepsilon^{N,K} g(u, x^N([u]_N))\, du \Big\} \\
&\leqq 8K^2 \int_s^t dr\, \tilde{\mu}^N([|x^N([r]_N) - x| > K]) \\
&\quad + 4K(t-s) \cdot \sup_{r, |z-x| \leqq K} \Big| \int_r^t du\, \tilde{E}^N_{r,z} [\nabla_\varepsilon^{N,K} g(u, x^N([u]_N))] \Big| \\
&\leqq 8K_2 e^{\alpha|x|^2/p} \cdot K^2 e^{-K/q} + 16K^2(t-s)\Delta^N \\
&\quad + 4K(t-s) \sup_{r, |z-x| \leqq K} \Big| \int_{r+2\Delta^N}^t du \int \tilde{p}^N(r, z, u, \xi)[g^{N,K}(u,\xi) - g_\varepsilon^{N,K}(u,\xi)]\, d\xi \Big|,
\end{aligned}
$$

where we have again used the partial integration formula, Proposition 2 and split the integral over $[r, t]$ into one over $[r, r+2\Delta^N]$ and one over $[r+2\Delta^N, t]$. Using Proposition 3, the term under the sup may be further estimated by

$$
\begin{aligned}
&\int \phi_\varepsilon(\zeta)\, d\zeta \Big| \int_{r+2\Delta^N}^t du \int \tilde{p}^N(r, z, u, \xi)[g^{N,K}(u,\xi) - g^{N,K}(u, \xi+\zeta)]\, d\xi \Big| \\
&\qquad \leqq \int \phi_\varepsilon(\zeta)\, d\zeta \int_{r+2\Delta^N}^t du \int |\tilde{p}^N(r, z, u, \xi) - \tilde{p}^N(r, z, u, \xi-\zeta)| |g^{N,K}(u,\xi)|\, d\xi \\
&\qquad \leqq K_6 K e^{\alpha|z|^2/p} \cdot \delta(N, \varepsilon)
\end{aligned}
$$

with

$$
\delta(N, \varepsilon) = \int \phi(\zeta) \tilde{\psi}(\varepsilon|\zeta|)\, d\zeta + \Delta^{N/q}.
$$

Since $\tilde{\psi}(|h|)$ is continuous in h and converges to 0 as $h \to 0$ (cf. Lemma 8 and formula (4.5)),

$$
\limsup_{\varepsilon \downarrow 0} \limsup_{N \to \infty} \delta(N, \varepsilon) = 0.
$$

So, finally, the sup may be estimated by $K_6 K e^{4\alpha|x|^2/p} e^{4\alpha K^2/p} \delta(N, \varepsilon)$, thus leading to the estimate

$$
J_1 \leqq M_1 K^2 (e^{-K/q} + \Delta^N + e^{4\alpha K^2/p} \delta(N, \varepsilon))
$$

with some suitably chosen constant M_1.

Since $g^N(r,\xi)-g^{N,K}(r,\xi)=g^N(r,\xi)\chi_{\bar{A}_{N,K}}(r,\xi)$ ($\bar{A}$ denoting complement), we have for J_2 the estimate

$$J_2=E\left\{\int_s^t g^N(r,X^N([r]_N))\chi_{\bar{A}_{N,K}}(r,X^N([r]_N))\,dr\right\}^2$$

$$\leqq\int_s^t dr\,E[g^N(r,X^N([r]_N))]^2\cdot P[|g^N(r,X^N([r]_N))|>K]$$

$$\leqq K^{-2}\int_s^t\{E[g^N(r,X^N([r]_N))]^2\}^2\,dr$$

$$\leqq M_2K^{-2}.$$

Finally turning to J_3, note first that $g_\varepsilon^N(r,\xi)-g_\varepsilon^{N,K}(r,\xi)=\int\phi_\varepsilon(\zeta)h^{N,K}(r,\xi+\zeta)\,d\zeta=h_\varepsilon^{N,K}(r,\xi)$ with $h^{N,K}(r,\xi)=g^N(r,\xi)\chi_{\bar{A}_{N,K}}(r,\xi)$. But then

$$|h_\varepsilon^{N,K}(r,\xi)|^2\leqq\left\{\sqrt{C_f}\int\phi_\varepsilon^{1/2}(\xi-\zeta)(1+|\zeta|)\phi_\varepsilon^{1/2}(\xi-\zeta)\chi_{[\sqrt{C_f}(1+|z|)>K]}(\zeta)\,d\zeta\right\}^2$$

$$\leqq C_f\int\phi_\varepsilon(\xi-\zeta)(1+|\zeta|)^2\,d\zeta\int\phi_\varepsilon(\xi-\zeta)\chi_{[\sqrt{C_f}(1+|z|)>K]}(\zeta)\,d\zeta.$$

But the last integral may be estimated by

$$K^{-2}C_f\int(1+|\zeta|)^2\phi_\varepsilon(\xi-\zeta)\,d\zeta.$$

Hence

$$|h_\varepsilon^{N,K}(r,\xi)|^2\leqq K^{-2}C_f^2\left\{\int\phi_\varepsilon(\xi-\zeta)(1+|\zeta|)^2\,d\zeta\right\}^2$$

$$\leqq 16K^{-2}C_f^2\left\{1+4|\xi|^2+4\varepsilon^2\int\phi(z)|z|^2\,dz\right\}^2$$

$$\leqq M_3K^{-2}(1+|\xi|^4)$$

with some constant M_3 independent of N, K and ε (for ε small). From this

$$J_3\leqq M_3K^{-2}(1+K_8).$$

Collecting the results on the J_i, we find that

$$\limsup_{\varepsilon\downarrow 0}\limsup_{N\to\infty}\Delta(N,\varepsilon)\leqq 9[M_1K^2e^{-K/p}+M_2K^{-2}+M_3K^{-2}(1+K_8)]. \tag{5.11}$$

Coming back to (5.10), we find that, by Chebyshev's inequality, the probability to be estimated is smaller than

$$\delta^{-2}\cdot\Delta(N,\varepsilon).$$

Since (5.11) is true for every K, assertion (5.10) follows. This completes at the same time the proof of (5.7).

As indicated above, we now know that X satisfies the stochastic differential equation (5.8) and

$$J^N(u^N)=Eg(X^N(1))\to Eg(X(1))=J(u), \tag{5.12}$$

where u^N was the optimal strategy for the Nth stage discrete problem and $u(t,x)$ the limit feedback control determined by assumption (A4). (5.12) follows from the growth

condition imposed on g, which ensures uniform integrability of the $g(X^N(1))$. Hence, if it can be shown that any admissible control U of the original continuous time problem can be approximated by discrete strategies $\pi^N = (U_0^N, \cdots, U_{\Delta^{-N}-1}^N)$ (note that from now on the π^N may be any admissible discrete strategies, not just the optimal ones as before) in the sense that $J^N(\pi^N) \to J(U)$, then it will follow from (5.12) that the control u determined by assumption (A4) is indeed the optimal one for the continuous time problem.

So, to attack this last step, take an arbitrary admissible control U defined on some fixed probability space $(\Omega, \mathcal{F}, P)$ with corresponding solution X of the stochastic differential equation

$$X(t) = x + \int_0^t f(s, X(s), U(s))\, ds + + B(t), \tag{5.13}$$

where X and U are nonanticipative with respect to the Brownian motion B. Define the smoothed controls

$$U_\varepsilon(t) = \frac{1}{\varepsilon} \int_{t-\varepsilon}^t U(s)\, ds$$

with corresponding solutions

$$X_\varepsilon(t) = x + \int_0^t f(s, X_\varepsilon(s), U_\varepsilon(s))\, ds + B(t).$$

It is easy to see that the U_ε are indeed admissible. Since $U_\varepsilon(t) \to U(t)$ for almost every t, from the estimate

$$|X(t) - X_\varepsilon(t)| \leqq L_1 \left(\int_0^t |X(s) - X_\varepsilon(s)|\, ds + \int_0^t |U(s) - U_\varepsilon(s)|\, ds \right)$$

it follows with the Gronwall–Bellman inequality and by bounded convergence that $X_\varepsilon(t) \to X(t)$ almost everywhere for all t. In particular, $J(U_\varepsilon) = Eg(X_\varepsilon(1)) \to Eg(X(1)) = J(U)$. Hence, for showing that any admissible control U can be approximated by discrete strategies in the sense mentioned above, we may suppose that U has continuous sample paths.

For such U, look again at equation (5.13) and set

$$U_n^N = U(n\Delta^N), \quad \varepsilon_n^N = B((n+1)\Delta^N) - B(n\Delta^N), \qquad n = 0, 1, \cdots, \Delta^{-N} - 1.$$

Then $\pi^N = (U_0^N, \cdots, U_{\Delta^{-N}-1}^N)$ is an admissible strategy for the Nth stage discrete problem. Let $X^N = (X_0^N, \cdots, X_{\Delta^{-N}}^N)$ be the solution of the difference equation

$$X_{n+1}^N = X_n^N + f(n\Delta^N, X_n^N, U_n^N)\Delta^N + \varepsilon_n^N, \qquad n = 0, 1, \cdots, \Delta^{-N} - 1,$$
$$X_0^N = x.$$

Let the processes $X^N(t)$, $0 \leqq t \leqq 1$, be defined as in the preceding sections. Then

$$X^N(t) = x + \int_0^t f([s]_N, X^N([s]_N), U([s]_N))\, ds + B^N(t)$$

with

$$B^N(t) = B([t]_N) + (t - [t]_N)\Delta^{-N}[B([t]_N + \Delta^N) - B([t]_N)]$$

(cf. the beginning of § 5). From the assumptions made about the function f it then follows that

$$
\begin{aligned}
&|X(t)-X^N([t]_N)| \\
&\quad \leqq \int_0^{[t]_N} |f(s, X(s), U(s))-f([s]_N, X^N([s]_N), U([s]_N))|\, ds \\
&\qquad + \int_{[t]_N}^{t} |f(s, X(s), U(s))|\, ds + |B(t)-B([t]_N)| \\
&\quad \leqq L_1 \left(\int_0^t |X(s)-X^N([s]_N)|\, ds + \int_0^t |s-[s]_N|\, ds + \int_0^t |U(s)-U([s]_N)|\, ds \right) \\
&\qquad + C_f^{1/2} \int_{[t]_N}^{t} (1+|X(s)|)\, ds + |B(t)-B([t]_N)|.
\end{aligned}
$$

By the Gronwall–Bellman inequality (cf. [5]) this implies that

$$
|X(t)-X^N([t]_N)| \leqq \delta_N(t) + L_1 \int_0^t e^{L_1(t-s)} \delta_N(s)\, ds,
$$

where

$$
\begin{aligned}
\delta_N(t) &= L_1 \left(\int_0^t |U(s)-U([s]_N)|\, ds + \int_0^t (s-[s]_N)\, ds \right) \\
&\quad + C_f^{1/2} \int_{[t]_N}^{t} (1+|X(s)|)\, ds + |B(t)-B([t]_N)|.
\end{aligned}
$$

Since, with probability one, $\delta_N(t) \to 0$ boundedly for all t, it follows that $X^N([t]_N) \to X(t)$ almost everywhere for all t. In particular, $X^N(1) \to X(1)$ almost everywhere, whence $J(U^N) \to J(U)$, which gives the desired approximation.

Remarks. a) If in assumption (A4) the u^N are only assumed to be ε-optimal—i.e., $J^N(u^N) \leqq V^N + \varepsilon$, where V^N denotes the value of the Nth stage problem—then the same proof shows that u is ε-optimal for the continuous time problem. Moreover, if the u^N are ε^N-optimal with $\varepsilon^N \downarrow 0$, then u is optimal.

b) It can be seen from the last step of the proof that continuity of the cost functional g may be replaced by the weaker requirement that the set G of discontinuity points has Lebesgue measure zero. Then, since $X(1)$ takes values in G with probability 0, the convergence $Eg(X^N(1)) \to Eg(X(1))$ will continue to hold. This allows one, for example, to treat maximum inclusion probability problems in the presented framework.

Acknowledgment. The author would like to thank the referee for his valuable comments and suggestions.

REFERENCES

[1] V. E. BENES, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.

[2] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[3] N. CHRISTOPEIT, *Existence of optimal stochastic controls under partial observation*, Z. Wahrsch. Verw. Geb., 51 (1980), pp. 201–213.

[4] N. CHRISTOPEIT AND K. HELMES, *On Benes's bang-bang control problem*, to appear.
[5] R. F. CURTAIN AND A. J. PRITCHARD, *Functional Analysis in Modern Applied Mathematics*, Academic Press, New York, 1977.
[6] M. H. A. DAVIS AND P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, this Journal, 11 (1973), pp. 226–261.
[7] E. B. DYNKIN AND A. A. YUSHKEWICH, *Controlled Markov Processes and Their Applications*, Springer-Verlag, New York, 1979.
[8] I. I. GIKHMAN AND A. V. SKOROKHOD, *Stochastische Differentialgleichungen*, Akademie Verlag, Berlin, 1971.
[9] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Prob. Appl., 5 (1960), pp. 285–301.
[10] W. GRÖBNER AND N. HOFREITER, *Integraltafel*, Springer-Verlag, New York, 1973.
[11] E. HEWITT AND K. STROMBERG, *Real and Abstract Analysis*, Springer-Verlag, New York, 1965.
[12] H. J. KUSHNER, *Existence results for optimal stochastic controls*, J. Optim. Theory Appl., 15 (1975), pp. 347–359.
[13] ———, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.
[14] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes* I, Springer-Verlag, New York, 1977.
[15] D. W. STROOCK AND S. R. S. VARADHAN, *Multidimensional Diffusion Processes*, Springer-Verlag, New York, 1979.
[16] K. YAMADA, *Approximation of Markovian control systems by discrete control policies*, Proc. International Symposium SDE, Kyoto, 1976, pp. 463–491.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 1, January 1983

0363-0129/83/2101-0003 $01.25/0

DIRICHLET BOUNDARY CONTROL PROBLEM FOR PARABOLIC EQUATIONS WITH QUADRATIC COST: ANALYTICITY AND RICCATI'S FEEDBACK SYNTHESIS*

I. LASIECKA† AND R. TRIGGIANI†

Abstract. For a parabolic equation in y defined on a bounded open domain Ω with boundary Γ and with control function u acting in the Dirichlet boundary condition, we study the optimal quadratic cost problem, which penalizes over an assigned time interval $[0, T]$ the $L_2(0, T; L_2(\cdot))$-norm of the solution y and of the control u, as well as the $L_2(\Omega)$-norm of the final state $y(T)$. Feedback synthesis (pointwise in time) of Riccati type: $u^0(t) = CP(t)y^0(t)$ of the optimal solution u^0, y^0 is established through a semigroup approach. Moreover, in contrast with the indirect approach of much of the literature, which relies on a Riccati equation to establish existence and numerical computability of the operator $P(t)$, the present approach is instead direct: i.e., the operator $P(t)$ is first defined by an explicit formula in terms of the system data, and only subsequently shown to satisfy, in an appropriate sense, a Riccati-type operator equation.

Solution to the Riccati feedback synthesis (§ 3) required some regularity results of the optimal solution u^0, y^0. Accordingly, the regularity question is taken up preliminarily (in § 2) and is carried out, in its own right and in full generality, much beyond the need of the Riccati synthesis.

Key words. parabolic boundary control, Riccati equation

1. Introduction.

1.1. Statement of the problem and reference to the literature. Let Ω be a bounded open domain in R^n with boundary Γ, assumed to be an $(n-1)$-dimensional variety with Ω locally on one side of Γ. Here, Γ may have finitely many conical points with Ω convex [K2, p. 227]. Let $A(\xi, \partial)$ be a uniformly strongly elliptic operator of order two in Ω of the form

$$A(\xi, \partial) = \sum_{|\alpha| \leqq 2} a_\alpha(\xi)\partial^\alpha,$$

with smooth real coefficients a_α, where the symbol ∂ denotes differentiation. We consider the optimal boundary control problem (P_Σ):

Minimize the performance index[1]

$$J(u, y(u)) \equiv |u|_\Sigma^2 + |y|_Q^2 + \alpha |y(T)|_\Omega^2 \tag{1.0}$$

over all $u \in L_2(\Sigma)$, subject to

$$\frac{\partial y}{\partial t}(t, \xi) = -A(\xi, \partial)y(t, \xi) \quad \text{in } (0, T] \times \Omega \equiv Q, \tag{1.1}$$

$$y(0, \xi) = y_0(\xi), \qquad \xi \in \Omega, \tag{1.2}$$

$$y(t, \sigma) = u(t, \sigma) \quad \text{in } (0, T] \times \Gamma \equiv \Sigma. \tag{1.3}$$

Here and throughout, $u(t, \cdot)$ is the boundary control acting in the Dirichlet B.C. and α denotes either 1 or else 0.

* Received by the editors February 9, 1981, and in revised form January 26, 1982. This research was supported in part by the National Science Foundation under grant MCS 81-02837.

† Department of Mathematics, University of Florida, Gainesville, Florida 32611.

[1] All norms are L_2-norms over the specified domains, unless a completely self-explanatory subindex is used. Moreover, the results of this paper trivially extend to the cost functional that includes, in the usual way, selfadjoint nonnegative operators on y and $y(T)$, and strictly positive on u.

Remark 1.1. It is known [L4, p. 202] that the response y to an $L_2(\Sigma)$-control may not have a well-defined final point $y(T)$ in the sense that it may well happen that $y(T) \notin L_2(\Omega)$ for some $u \in L_2(\Sigma)$. In this case, the corresponding value of J is $J(u, y(u)) = \infty$, for $\alpha = 1$. This is a pathology that will have to be treated.

A main goal of the present article can be informally described as follows: Establish the feasibility of a *pointwise* (in t) *feedback synthesis* of the optimal control u^0 in terms of the corresponding optimal solution y^0, as expressed by[2]

$$(*) \qquad u^0(t) = CP(t)y^0(t), \qquad 0 \leqq t < T.$$

Here, C is a time independent operator known in terms of the original parabolic equation (1.1)–(1.3), and $P(t)$ is a suitable operator. Further extending known theory, we shall prove the validity of (*) and, moreover, that $P(t)$ satisfies in an appropriate sense a Riccati type (nonlinear, in fact quadratic) operator differential equation in $0 \leqq t < T$ with assigned condition at $t = T$.

Feedback synthesis of optimal control problems with quadratic cost has been, of course, the object of extensive investigation over the past twenty years, since the fundamental work of R. E. Kalman originally centered on linear ordinary differential equations. The intimate relationship between the pointwise feedback realization of the optimal control and Riccati equations has been extended in various meanings and directions to a large variety of dynamical systems. This being the case, the burden is on us to justify another contribution to the subject, particularly when the final result is somehow expected and taken for granted.

The extensive literature on quadratic control problems and Riccati equations shrinks, however, to only a few references, when it comes to *boundary* control problems for partial differential equations; this is even more so in the most challenging case, where the control function acts within the Dirichlet boundary conditions.[3] For instance, a satisfactory fully L_2-theory of the optimal quadratic cost problem for Dirichlet boundary input hyperbolic equations is altogether lacking.[4] As to an L_2-theory for parabolic equations, the only article that treats the Dirichlet boundary input case—the one of our present paper—is Balakrishnan [B2]. The fundamental book by Lions treats only the parabolic boundary control problem with control function acting through the Neumann boundary condition, where higher regularity properties of the solutions are available (see [L4, pp. 159–163]). As to Curtain and Pritchard [C1], [C2], their evolution model[5] is not capable of treating $L_2(0, T; L_2(\Gamma))$-Dirichlet boundary control for parabolic equations with solutions $y \in L_2(0, T; L_2(\Omega))$ even for a one-

[2] The pointwise feedback synthesis is of established crucial importance in engineering practice, where it is termed "real time" or "on line" implementation.

[3] The Dirichlet boundary control problem is the most challenging case, due to the lowest regularity properties of the solutions of a second order P.D.E., as compared to the other two cases of Neumann or elastic (mixed) boundary conditions.

[4] See: Lions [L4, Ch. IV, § 8, p. 325] under the "very strong assumption" $u \in H_0^2(\Sigma)$; Curtain and Pritchard [C2, pp. 602–604], explicitly stating that their model cannot treat $L_2(\Sigma)$-control, whereby control action in $L_2(0, T; H^{1/2}(\Gamma))$ is considered; similarly Vinter and Johnson [VI] with control in $L_2(0, T; H^{1/2}(\Gamma))$.

[5] In Curtain and Pritchard [C2], the overall viewpoint is considerably different from ours. Their analysis starts off, in its first stage, from a dynamical model consisting of an inhomogeneous time dependent evolution equation, on which they endeavor to impose the weakest possible assumptions that they can. Only at their second stage do they attempt to cast various specific dynamical equations, including some boundary control problems, within the original framework of their evolution equation, by trying to match the original standing assumptions. Also centered on an evolution operator model is the paper [G1] by Gibson: This, however, says nothing about boundary control, as weakest possible assumptions are attempted only on the free dynamics (evolution operator of the free system), while the control acts on the system through a bounded time dependent operator. We will have more to say about [G1] later on.

dimensional $\Omega = (0, 1)$: See their example [C2, Ex. 6.2, p. 600], where y is taken in the undesirable space $L_2(0, T; H^{-\frac{1}{2}-\varepsilon}(\Omega))$ and hence the performance index J is penalized in y with respect to the "artificial" norm $L_2(0, T; H^{-\frac{1}{2}-\varepsilon}(\Omega))$, with final state penalized in the corresponding norm $H^{-\frac{1}{2}-\varepsilon}(\Omega)$. Thus, the only work specifically pertinent to our parabolic problem here, is Balakrishnan [B2], where indeed an L_2-theory for Dirichlet boundary control with no final state ($\alpha = 0$ in (1.0)) is presented.

Our approach to modeling the dynamics of boundary input parabolic equations owes much to Balakrishnan, as we also work not directly with the parabolic equation, but rather with a semigroup rooted input-solution formula (a sort of variation of parameter formula) first introduced and studied by Balakrishnan and his former student Washburn [B1], [B2], [W1], [W2]. Regarding, however, the Riccati feedback synthesis, there are notably important differences between Balakrishnan's treatment [B1], [B2] and ours here in § 3. The main ones are the following:

(i) As to the problem studied, our cost functional penalizes also the final state ($\alpha = 1$ in (1.0)), which was instead omitted in [B2] ($\alpha = 0$ in Eq. (1.0)). This addition is responsible for further technical difficulties and complications, in view of the general Remark 1.1.

(ii) As to the method of solution proposed for finding the Riccati operator $P(t)$, Balakrishnan's approach falls in with the general trend that may be labeled as *indirect*. In this trend, which is motivated by finite dimensional theory, one attempts to express the optimal control u^0 in terms of the optimal trajectory y^0 by means of the pointwise relation (*), for some postulated operator $P(t)$ and then deduce a Riccati-type equation for $P(t)$. This way, one runs into the difficulty of having to settle the technical issue of existence, i.e., that the obtained Riccati equation does admit a solution. On this point Balakrishnan's strategy consists in: (i) first proving existence of the Riccati equation in the case of distributed control [B1, § 5.2 and § 6.8]; (ii) then in cleverly recovering the Dirichlet boundary case as a limit process of a suitable sequence of distributed control problems [B2], to which the previous theory from [B1] applies. It is undesirable, however, that the existence proof of the distributed case, and hence of the boundary case, is based on a stochastic technique (for the corresponding filtering problem [B2, § 6.8]). In any case, in Balakrishnan's approach, the numerical solution of the Riccati operator equation is the only way available to determine $P(t)$.

By contrast, our approach is *direct* and *explicit*, in that we first define (construct) an operator $P(t)$ in terms of the original parabolic equation (1.1)–(1.3) and then deduce that $P(t)$ must, in fact, satisfy a Riccati-type equation. Here, therefore, $P(t)$ is not postulated but actually defined in terms of the system's data, and the existence question of the Riccati equation is automatically taken care of.[6,7] See (3.10).) We are also grateful to an anonymous referee for pointing out to us, that our present derivation of the closed loop evolution operator and the solution of the Riccati equation in § 3 bears a closer connection with Gibson's corresponding derivation for the regulator

[6] The first to use a direct approach is, apparently, Lions [L4, pp. 135–6; p. 266] in connection with parabolic equations with distributed control: here the operator $P(t)$ is first defined by

$$P(t)x = p^0(t, x) \qquad (r(t) \equiv 0 \text{ with zero final state}),$$

where $p^0(t, x)$ is the solution (not explicitly available, though!) of a coupled system of P.D.E.s that arise from application of the Hamilton–Jacobi theory. Next, Lions verifies that $P(t)$ satisfies, in fact, a Riccati equation. Gibson's approach in [G1] is likewise direct and explicit.

[7] Our approach, therefore, provides by means of the defining formula (3.10) for $P(t)$ an alternative route to the convergence analysis of numerical approximation schemes of $P(t)$. This way, the need—present in existing literature—to compute $P(t)$ as a solution of a (quadratic) operator Riccati equation is altogether eliminated.

problem in [G1] than we were originally aware of. In fact, even though [G1] says nothing about boundary control, if however one formally replaces the bounded operator B in [G1] with the operator AD (whose domain in $L_2(\Omega)$ consists, technically, only of the zero vector!), one will obtain some remarkable formal similarities between the two treatments involving $P(t)$. (We leave it to the footnotes to point out formulas in [G1] which formally correspond to ours.) This is an inherent bonus of our present semigroup approach to the study of the significantly different *boundary* control problems. At the conceptual level, the operators we introduce and study come directly from the original partial differential equation with nonhomogeneous boundary control. At the technical level, they introduce further pathologies, due to intrinsically unbounded operators which describe the action of the boundary control as well as the not-always well defined final state: all this is reflected in the milder statements of our final results.

To achieve the Riccati (pointwise) feedback synthesis of the optimal control u^0 (in § 3), we shall need, to be sure, certain *regularity* properties of the optimal solution u^0 and y^0. Therefore, we shall first study (in § 2) the question of regularity of the optimal solution to problem (P_Σ). This will be done in its own right and full generality, regardless of what is strictly needed for the Riccati synthesis. Only a much weaker version of the regularity results of § 2 will be actually invoked in § 3 (e.g., rather than full analyticity of u^0 and y^0 in $(0, T)$, the Riccati synthesis will need only continuity of y^0 in $[0, T]$ (endpoints included) and, say, C^1-smoothness in $(0, T)$, together with continuity of u^0 in $[0, T)$ (when $\alpha = 1$ in (1.0)): compare with full statement of Theorem 1.1). Accordingly, some of the technical analysis of § 2, including the extension to complex variables, should *not* be viewed as a necessary prerequisite to establish Riccati's synthesis.

The question of regularity of the optimal boundary control u^0, which solves the problem (P_Σ), has already been the subject of certain recent studies. In this connection, we cite particularly [L1], [L3, § 6] and [S1]. Both works [L1] and [L3, § 6], in which the author was primarily interested in describing the regularity of u^0 over the *entire* interval $[0, T]$, refer to the problem (P_Σ) with $\alpha = 0$. Her results are: (i) that $u^0 \in H^{2-2\varepsilon, 1-\varepsilon}(\Gamma \times [0, T])$ for smooth Γ $(\varepsilon > 0)$, and (ii) that $u^0 \in H^{1-2\varepsilon, \frac{1}{2}-\varepsilon}(\Gamma \times [0, T])$ in the case of conical domains Ω. In addition, for smooth Γ, [L3] proves also that $u^0 \in C^\infty(\Gamma \times \{0 < t < T\})$. No attempt was made, however, to treat the case $\alpha = 1$ in (1.0), which will be needed in § 3.

Credit goes to T. Seidman for having first pointed out that, in the case of smooth Γ and for the problem (P_Σ) with $\alpha = 1$, the optimal control u^0 is, in fact, *analytic in* $(0, T)$ *as an abstract function with values in* $H^s(\Gamma)$, *for any* $s \geqq 0$; see his main theorem in [S1]. As a result of their contacts with Seidman, the authors have consequently re-examined the problem of regularity for (P_Σ) with general nonsmooth boundary Γ (conical points are allowed), but this time with $\alpha = 1$. It was then realized that a new, quicker alternative proof of the analyticity of u^0 in $(0, T)$ could be given, different in spirit and in many technicalities from Seidman's original proof. This is the object of § 2, which is essentially self-contained. Our approach starts with the line of argument previously used by I. Lasiecka to study (P_Σ) with $\alpha = 0$: in particular, an explicit representation[8] of u^0, obtained via a novel application of optimization techniques to the integral version (2.2b) of the dynamics, rather than to the differential version (1.1)–(1.3).

[8] Seidman's approach in [S1] does not provide an explicit representation for u^0. Rather, u^0 is identified, and hence studied, through two coupled equations.

1.2. Statement of main results. The main result on regularity of the optimal solution u^0 and y^0 to problem (P_Σ) is as follows (see § 2):

THEOREM 1.1. *The unique optimal control u^0 for problem (P_Σ) is an $L_2(\Gamma)$-function analytic in $(0, T)$ and continuous at $t = 0$. If $\alpha = 0$ in (1.0), then u^0 is also continuous at $t = T$. The corresponding optimal solution y^0 is an $L_2(\Omega)$-function analytic in $(0, T)$ and continuous at the endpoints $t = 0$ and $t = T$.*

Once analyticity is achieved in $L_2(\Gamma) = H^0(\Gamma)$, it can readily be extended to higher order Sobolev spaces based on Γ.

COROLLARY 1.2. *Let the boundary Γ be such that, for all $s \geqq 0$,*

$$\text{the Dirichlet map } D^9 \text{ is continuous from } H^s(\Gamma) \to H^{s+1/2}(\Omega),$$

and likewise for D^*. Then u^0 is analytic in $(0, T)$ as a function with values in $H^s(\Gamma)$ for any $s \geqq 0$ and likewise y^0 in $H^s(\Omega)$.

Remark 1.2. This assumption on Γ holds for all smooth Ω [L5, I, p. 188] and also for all Ω with conical points, provided that certain relations between the dimension of Ω and the solid angles of the conical points hold, as specified in [K2, Thms. 4.1, 4.2]. As to the fulfillment of the assumption on D^*, however, the literature that we are aware of assumes smooth Γ.

As to the feedback synthesis problem, our analysis in § 3 culminates with the statement that *the optimal control u^0 can be synthesized* (implemented) *as a pointwise (in t) feedback realization of the corresponding optimal solution y^0*, as dictated by (*), *where $P(t)$ is a Riccati operator.* The full description of the situation is contained in the results of § 3 which, in the more demanding case of final state penalization ($\alpha = 1$ in (1.0)), claim in particular that: (i) for $0 \leqq t < T$ the operator $P(t)$ satisfies a Riccati operator equation only in a sense slightly weaker than the weak topology of $L_2(\Omega)$ (i.e., $x, y \in \mathcal{D}(A^\varepsilon)$ rather than $x, y \in L_2(\Omega)$ in subsequent eq. (3.22)); (ii) the terminal value $P(T)$ is defined in the sense of the strong topology; (iii) the Riccati feedback synthesis holds only for $0 \leqq t < T$ (even though u^0 and y^0 are analytic in $0 < t < T$): see subsequent equation (3.9), which also shows that the operator C of (*) is, in fact, $-D^*A^*$ (i.e., the normal derivative operator [B1, p. 220]).

By contrast, the situation without final state penalization (i.e., $\alpha = 0$ in (1.0) is regular including the final point T.

2. Proof of Theorem 1.1 and Corollary 1.2.

2.1. Analyticity of the optimal control u^0. Our proof will be articulated in the following steps.

A) Introduction of a semigroup-rooted integral model to study the non-homogeneous mixed problem (1.1)–(1.3), essentially due to [B1, § 4.1], [B2], [W1], [W2] in the form (2.2b) (left), as modified by the authors by means of fractional powers in (2.2b) (right).

B) Derivation of an explicit expression for the optimal control u^0 (see (2.10)), by means of the Lagrange multiplier theory (Luisternik's theorem) as applied to the *integral* version (2.2b), rather than the *differential* form (1.1)–(1.3), of the dynamics. This is, we believe, a much more advantageous route. It was first used in [L2, § 7] for the case $\alpha = 0$ in (1.0) and led to a quick derivation of the same expression for u^0 as was previously obtained in [B2] through a lengthier limit process based on finite dimension theory.

C) Extension from real t to complex z (in a suitable sector $\mathcal{F}$ of $\mathbb{C}$ based on $[0, T]$) of the quantities entering in the definition of u^0 in (2.10). In particular,

[9] D is defined in § 2, Step A(i).

assertion of elementary but useful properties of the operators L and L^* on the space $\mathscr{A}(\mathscr{F}; L_2(\cdot))$ (see Lemma 2.1 below).

D) (This step is needed only when $\alpha = 1$ in (1.0).) Modification of the expression defining u^0 to an equivalent form, more suitable for the inversion of the operator $(I + L^*L)$ in $\mathscr{A}(\mathscr{F}; L_2(\Gamma))$.

E) Inversion[10] of $(I + L^*L)$ over $\mathscr{A}(\mathscr{F}; L_2(\Gamma))$ through the crucial fact that L is a compact operator from $\mathscr{A}(\mathscr{F}; L_2(\Gamma))$ into $\mathscr{A}(\mathscr{F}; L_2(\Omega))$.

Step A. *The semigroup model for* (1.1)–(1.3) *and preliminaries.*

(i) Let D be the *Dirichlet map*, defined by $y = Dv$, where

$$-A(\xi, \partial)y = 0 \quad \text{in } \Omega; \qquad y|_\Gamma = v.$$

It is known, through the combination of various results ([F2], [L3, App. B], [L5, p. 107], [L6, Thm. 5.1], [K2, p. 227]) that for smooth and for conical convex domains

(2.1) *D is a continuous operator from* $L_2(\Gamma) \to \mathscr{D}(A^{\frac{1}{4}-\varepsilon})$,[11] $\quad 0 < \varepsilon \leqq \frac{1}{4}$

In [N1, Prob. 1.1, p. 252], this property is believed to be true for all Ω belonging to the very large class $\mathscr{M}$.

(ii) Let $(-A)$: $L_2(\Omega) \supset \mathscr{D}(-A) \to L_2(\Omega)$ be the elliptic operator $-A(\xi, \partial)$ with *zero* Dirichlet B.C.

(iii) Let $S(t)$ be the C_0-semigroup on $L_2(\Omega)$, $t \geqq 0$, generated by $(-A)$ [F1, Ex. p. 101], which is analytic (holomorphic) on a triangular sector containing the positive real line.

Recently, Balakrishnan and Washburn [B1], [B2], [W1], [W2] have established that the nonhomogeneous mixed problem (1.1)–(1.3) admits the following semigroup model on $L_2(\Omega)$:

$$y(t) = S(t)y_0 + (Lu)(t), \qquad y_0 \in L_2(\Omega), \tag{2.2a}$$

where Lu is given by the left-hand side integral of

$$\begin{aligned}(Lu)(t) &= \int_0^t AS(t-\tau)Du(\tau)\, d\tau \\ &= \int_0^t A^{\frac{3}{4}+\varepsilon} S(t-\tau)A^{\frac{1}{4}-\varepsilon}Du(\tau)\, d\tau, \qquad 0 < \varepsilon \leqq \tfrac{1}{4},\end{aligned} \tag{2.2b}$$

to be interpreted in the sense that a control $u \in L_2(\Sigma)$ produces, through (2.2), a response $y \in L_2(Q)$.[12] (Compare with Remark 1.1; see also [L3] for a quite general treatment of the regularity problem based on (2.2).) As in our previous work on parabolic boundary input problems, we prefer, however, to use the crucial fact (2.1) and rewrite Lu as the right-hand side integral in (2.2b), by means of fractional powers. This is a more convenient way which makes some of the properties of L more apparent. From (2.2b), on can deduce directly the property (recalled above) that L is a continuous ([B2], [W1], [W2]) linear operator form $L_2(\Sigma)$ into $L_2(Q)$. The dual operator L^* of

[10] In [L3], instead, inversion of $(I + L^*L)$ was proved in the space $H^{2-\varepsilon, 1-\varepsilon}(\Gamma \times [0, T])$ and used to conclude that u^0 belongs to such a space.

[11] At the price of considering a suitable translation of A, rather than A itself, we may assume that the fractional powers of A are well defined; this does not change the local regularity in time of the solutions, the object of this section.

[12] However, $u \in L_p(0, T; L_2(\Gamma))$ for $p > 4$ produces a meaningful pointwise response $y(T) \in L_2(\Omega)$ [W1], [W2].

L, continuous from $L_2(Q)$ into $L_2(\Sigma)$, is [B1, p. 219]:

$$(L^*v)(t) = \int_t^T D^*(AS(\tau - t))^* v(\tau)\, d\tau, \tag{2.3}$$

as one can easily deduce from the definition[13] $(Lu, v)_Q = (u, L^*v)_\Sigma$. We next introduce the operator L_T:

$$L_T u = \int_0^T A^{\frac{3}{4}+\varepsilon} S(T-t) A^{\frac{1}{4}-\varepsilon} Du(t)\, dt, \qquad 0 < \varepsilon \leqq \tfrac{1}{4}, \tag{2.4}$$

unbounded from $L_2(\Sigma) \supset \mathscr{D}(L_T)$ into $L_2(\Omega)$, as noted in Remark 1.1 with

$$\mathscr{D}(L_T) = \{u \in L_2(\Sigma) : L_T u \in L_2(\Omega)\}. \tag{2.5}$$

Notice that $A^{-\frac{1}{4}-2\varepsilon} L_T$ is a bounded operator from $L_2(\Sigma)$ into $L_2(\Omega)$, since the integral kernel then becomes $O(1/(T-t)^{\frac{1}{2}-\varepsilon})$. Therefore, L_T *is* closed, being the product of A, which is closed and invertible in $L_2(\Omega)$, with $A^{-1}L_T$, which is bounded on $L_2(\Sigma)$ [K1, Prob. 5.7, p. 164].

The dual L_T^* is given by [W1, p. 69]:

$$L_T^* y = D^* A^* S^*(T-t) y, \qquad 0 \leqq t < T, \tag{2.6}$$

as can easily be seen from the definition $(L_T u, y)_\Omega = (u, L_T^* y)_\Sigma$ for $u \in \mathscr{D}(L_T)$ and $y \in \mathscr{D}(L_T^*)$. The operator L_T^* *is unbounded from* $L_2(\Omega) \supset \mathscr{D}(L_T^*)$ *into* $L_2(\Sigma)$. Notice that $\mathscr{D}(L_T)$ is dense in $L_2(\Sigma)$ (e.g., $C([0, T]; L_2(\Gamma)) \subset \mathscr{D}(L_T)$). Since L_T is closed, we then introduce, with reference to Remark 1.1, the Hilbert space U given by $U \equiv \mathscr{D}(L_T)$ (eq. (2.5)), equipped with the graph norm $|u|_U^2 = |u|_\Sigma^2 + |L_T u|_\Omega^2$.

Instead of the original minimization problem (P_Σ), we are then led to consider the following minimization problem (P_U):

Minimize the performance index

$$J(u, y(u)) \equiv \tfrac{1}{2}\{|u|_\Sigma^2 + |y|_Q^2 + |S(T)y_0 + L_T u|_\Omega^2\}$$

over all $u \in U$, subject to (2.2a), where L is given by (2.2b).

Problem (P_U) is a classical quadratic problem with continuous, strictly convex $J(u, y)$, having therefore a unique solution in U. Such a solution is then the unique minimizing solution also of problem (P_Σ); i.e., what we have denoted by u^0. As for $u \in L_2(\Sigma)$ but $u \notin U$, the corresponding index value $J(u, y(u)) = \infty$.

Step B. *Explicit characterization of the optimal* u^0 *via Lagrange multiplier theory as applied to* (P_U). For $u \in U$, $y \in L_2(Q)$, $p \in L_2(Q)$, we define the Lagrangian

$$\mathscr{L}(u, y, p) \equiv \tfrac{1}{2}\{|u|_\Sigma^2 + |y|_Q^2 + |S(T)y_0 + L_T u|_\Omega^2\} + (p, y - S(\cdot)y_0 - Lu)_Q. \tag{2.7}$$

Since $y - S(\cdot)y_0 - Lu$ obviously maps $U \times L_2(Q)$ onto $L_2(Q)$ (for any $g \in L_2(Q)$ take $u = 0$ and $y = S(\cdot)y_0 + g$), Liusternik's general Lagrange multiplier theorem, as in [L7, Thm. 1, p. 243] applies and gives: *there exist* $u^0 \in U$, $y^0 \in L_2(Q)$, $p^0 \in L_2(Q)$ *such that* $\mathscr{L}_u = \mathscr{L}_y = \mathscr{L}_p = 0$ *at* (u^0, y^0, p^0). From (2.7), we compute $\mathscr{L}_y = \mathscr{L}_u = 0$ to obtain, respectively,

$$(y^0, \delta y)_Q + (p^0, \delta y)_Q = 0 \quad \forall \delta y \in L_2(Q) \text{ or } p^0 = -y^0,$$

$$(u^0 - L^* p^0, \delta u)_\Sigma = -(S(T)y_0 + L_T u^0, L_T \delta u)_\Omega \quad \forall \delta u \in U.$$

[13] Inner product notation on $L_2(\cdot)$ will specify only the domain $(\cdot)$.

As the left-hand side of the above equation is well defined and continuous for all $\delta u \in L_2(\Sigma)$, we deduce that

(2.8)[14] $\quad S(T)y_0 + L_T u^0 \in \mathscr{D}(L_T^*)$, and hence that: $L_T u^0 \in \mathscr{D}(L_T^*)$,

since $S(T)y_0 \in \mathscr{D}(L_T^*)$. (This is apparent from the definition of L_T^* in (2.6), which makes the t-trajectory $L^*S(T)y_0$ well defined also at $t = T$, since $S(T)y_0 \in \mathscr{D}(A^*) = \mathscr{D}(A)$) [L5, p. 196].) Since (as recalled) U is dense in $L_2(\Sigma)$, it follows that

(2.9a)[14] $\quad u^0 - L^*p^0 = -L_T^* y^0(T) = -L_T^*[S(T)y_0 + L_T u^0]$.

After replacing $-p^0$ with y^0 and using the dynamics for y^0, we obtain

(2.9b) $\quad [I + L^*L + L_T^* L_T]u^0 = -L^*S(\cdot)y_0 - L_T^* S(T)y_0$.

The operator in square brackets on the left is obviously invertible with bounded inverse on all of $L_2(\Sigma)$[15] and therefore

(2.10)[16] $\quad u^0 = -[I + L^*L + L_T^* L_T]^{-1}[L^*S(\cdot)y_0 + L_T^* S(T)y_0]$

is the explicit expression of the optimal control that we seek. However, to deduce the analyticity of u^0, further work is needed for which we find (2.9) more suitable. We rewrite it as

(2.11) $\quad (I + L^*L)u^0 = b + g$,

where the vectors b and g, defined as

(2.12)
$$b \equiv -L^*S(\cdot)y_0 \equiv b_t \equiv \{-L^*S(t)y_0,\ 0 \leqq t \leqq T\},$$
$$g \equiv -L_T^*[S(T)y_0 + L_T u^0] \equiv g_t \equiv \{-D^*A^*S^*(T-t)[S(T)y_0 + L_T u^0],\ 0 \leqq t \leqq T\}$$

are well defined vectors of $L_2(\Sigma) \equiv L_2(0, T; L_2(\Gamma))$ (see (2.8) and (2.6)) and where the subindex t is added to emphasize the t-dependence of the vectors as $L_2(\Gamma)$-trajectories. The fact that g is defined in terms of u^0 will have no implication, so explicit dependence is omitted.

Step C. *Extension from real t to complex z; properties of L and L* on* $\mathscr{A}(\mathscr{F}; L_2(\cdot))$. In order to obtain the desired analyticity result of u^0, it is essential to extend the definition of the quantities entering into (2.11) from $t \in (0, T)$ to $z \in \mathscr{F}$, with $\mathscr{F}$ an appropriate complex sector [S1]. Let $\mathscr{F}$ be the open symmetric sector of $\mathbb{C}$ based on $(0, T)$ and delimited by the four line segments $\rho e^{\pm i\phi}, \rho e^{\pm i(\pi-\phi)} + T$, $0 \leqq \rho \leqq \rho_{\max}$ for some $\phi: 0 < \phi < \pi/2$ so that $\mathscr{F}$ is entirely contained in the infinite sector of analyticity of the semigroups $S(z)$ and $S^*(z)$. Let $\bar{\mathscr{F}}$ be its closure.

Remark 2.1. The choice of this particular sector, based on $(0, T)$, is not really essential, and other sectors will do as well. However, $\mathscr{F}$ is particularly convenient, since the transformation $z \to T - z$, needed below, maps $\mathscr{F}$ onto itself.

We introduce the space $\mathscr{A}(\mathscr{F}; L_2(\Omega))$ of all $L_2(\Omega)$-functions $f(z)$ that are: (i) *analytic (holomorphic) on* $\mathscr{F}$ *and* (ii) *continuous on* $\bar{\mathscr{F}}$. Equipped with the norm

$$|f|_{\mathscr{A},\Omega} \equiv \max_{z \in \bar{\mathscr{F}}} |f(z)|_\Omega,$$

[14] As was pointed out by an anonymous referee, the final conclusions in (2.8) and (2.9a) can be arrived at in a more elementary way (i.e., without use of Lagrange multiplier theory) through a simple direct computation of the nonnegative difference: $J(u^0 + v, y(u^0 + v)) - J(u^0, y^0)$, for $v \in \mathscr{D}(L_T)$. Setting the Fréchet derivative of J to zero will also do the job.

[15] This follows since, for $u \in \mathscr{D}(L_T)$, $([I + L^*L + L_T^* L_T]u, u)_\Sigma = |u|_\Sigma^2 + |Lu|_Q^2 + |L_T u|_\Omega^2 \geqq |u|_\Sigma^2$. Moreover, the operator $I + L^*L + L_T^* L_T$, being selfadjoint in $L_2(\Sigma)$, has an empty residual spectrum.

[16] Compare with [G1, eq. (3.14)].

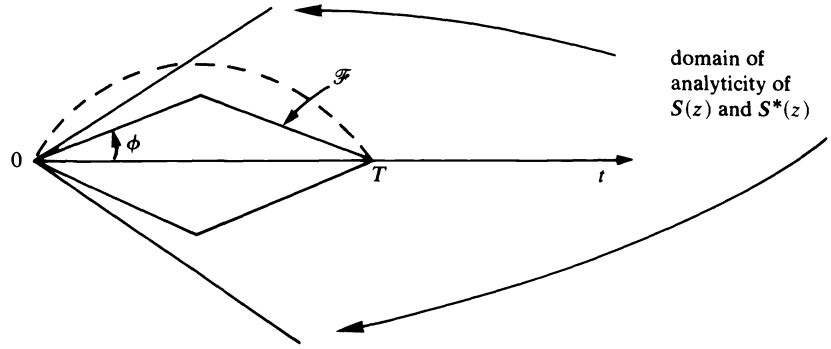


FIG. 1

the space $\mathcal{A}$ is a Banach space, completeness being a consequence of Weierstrass' uniform convergence theorem.[17] A similar definition applies for $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$ and its relative norm.

The following simple but important lemma will be used repeatedly below.

LEMMA 2.1. *The following properties hold for the operators L and L^*, once extended in a natural way to $z \in \mathcal{F}$.*

(i) *L continuously maps $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$ into $\mathcal{A}(\mathcal{F}; L_2(\Omega))$:*

$$|Lu|_{\mathcal{A},\Omega} \leqq T^{\frac{1}{4}-\varepsilon} K |u|_{\mathcal{A},\Gamma};$$

(ii) *L^* continuously maps $\mathcal{A}(\mathcal{F}; L_2(\Omega))$ into $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$:*

$$|L^*v|_{\mathcal{A},\Gamma} \leqq T^{\frac{1}{4}-\varepsilon} K |v|_{\mathcal{A},\Omega}.$$

Here, the constant K is defined by

$$K = M\|A^{\frac{1}{4}-\varepsilon} D\| = M\|D^* A^{*\frac{1}{4}-\varepsilon}\|, \tag{2.13}$$

where $\|\cdot\|$ are the unambiguous operator norms from $L_2(\Gamma)$ into $L_2(\Omega)$ in one case, and from $L_2(\Omega)$ into $L_2(\Gamma)$ in the other case (see eq. (2.1)). Moreover, M is a constant so that

$$|A^{\frac{3}{4}+\varepsilon} S(z)|_\Omega \leqq \frac{M}{|z|^{\frac{3}{4}+\varepsilon}} \quad \text{and} \quad |A^{*\frac{3}{4}+\varepsilon} S^*(z)|_\Omega \leqq \frac{M}{|z|^{\frac{3}{4}+\varepsilon}} \tag{2.14}$$

([F1; p. 101] *plus interpolation*) *for $z \in \mathcal{F}$.*

Proof of Lemma 2.1. (i) Let $u(z) \in \mathcal{A}(\mathcal{F}; L_2(\Gamma))$. With $z = r\,e^{i\theta} \in \mathcal{F}$, take a line segment $\zeta = \rho\, e^{i\theta}$, $0 \leqq \rho \leqq r$, from 0 to z and extend Lu in (2.2b) by

$$(Lu)(z) = \int_0^z AS(z-\zeta) Du(\zeta)\, d\zeta = \int_0^z A^{\frac{3}{4}+\varepsilon} S(z-\zeta) A^{\frac{1}{4}-\varepsilon} Du(\zeta)\, d\zeta, \tag{2.15}$$

Dividing the integration of the left integral into two parts and integrating by parts, we find that the second term yields

$$(Lu)(z) = AS(z-\varepsilon\, e^{i\theta}) \int_0^{\varepsilon e^{i\theta}} S(\varepsilon\, e^{i\theta} - \zeta) Du(\zeta)\, d\zeta + Du(z) - S(z-\varepsilon\, e^{i\theta}) Du(\varepsilon\, e^{i\theta})$$

$$- \int_{\varepsilon e^{i\theta}}^z S(z-\zeta) D \frac{du}{d\zeta}\, d\zeta$$

[17] The *real* version of this theorem does *not* hold and, for this reason, extension from t into z is essential.

and exhibits $(Lu)(z)$ as the sum of four $L_2(\Omega)$-analytic functions in $\mathscr{F}$. This is so because $u(\cdot)$ and $S(\cdot)$ are analytic in $\mathscr{F}$, $AS(z-\varepsilon e^{i\theta}) = S(z-3(\varepsilon/2)e^{i\theta})AS((\varepsilon/2)e^{i\theta})$ and, moreover, because the integral of the first term is a well defined $L_2(\Omega)$-vector, u being an $L_2(\Gamma)$-continuous function in $\mathscr{F}$. To show the continuity of L, we use instead the more convenient right integral in (2.15) so that

$$|(Lu)(z)|_\Omega \leqq \int_0^r \frac{K}{(r-\rho)^{\frac{3}{4}+\varepsilon}}\, d\rho\, |u(\cdot)|_{\mathscr{A},\Gamma},$$

and the desired conclusion for L follows.

Remark 2.2. In the Appendix, we shall also make use of the following properties:

a) The operator L takes an $L_2(\Gamma)$-function, which is L_∞ in $z \in \mathscr{F}$, into an $L_2(\Omega)$-function, which is continuous in $z \in \bar{\mathscr{F}}$. (The proof is identical to that given above, using sup rather than max.)

b) The operator L continuously maps the space

$$\{h(z)\colon h \in L_2(\mathscr{F}; L_2(\Gamma)),\ h(z) \text{ analytic in } \mathscr{F}\}$$

into the space

$$\{\psi(z)\colon \psi \in L_2(\mathscr{F}; L_2(\Omega)),\ \psi(z) \text{ analytic in } \mathscr{F}\}$$

with norm

$$|h|^2 = \int_{\mathscr{F}} |h(z)|_\Gamma^2\, d\mathscr{F},$$

and similarly for ψ. The same proof applies as that following (2.15), except that now the vector $\int_0^{\varepsilon e^{i\theta}} S(\varepsilon e^{i\theta} - \zeta)Du(\zeta)\, d\zeta$ need not be well defined in $L_2(\Omega)$. It is well defined, however, in, say, the dual space $[\mathscr{D}(A^{\frac{1}{4}+\varepsilon})]'$, to which the semigroup can be extended to be still analytic.[18]

Property (b) has a similar counterpart for L^*.

(ii) The proof of L^* is similar, using (2.3) instead of (2.2b) and noticing that $D^*A^{*\frac{1}{4}-\varepsilon}$ is bounded in $L_2(\Omega) \to L_2(\Gamma)$, being the dual of a bounded operator (see (2.2)). □

We now return to (2.11)–(2.12). We first extend b_t and g_t to $L_2(\Gamma)$-trajectories b_z and g_z in a natural way by setting, in accordance with (2.12),

$$b_z \equiv \{-L^*S(z)y_0,\ z \in \bar{\mathscr{F}}\}$$

so that, by Lemma 2.1(ii),

$$b_z \in \mathscr{A}(\mathscr{F}; L_2(\Gamma)), \tag{2.16}$$

and setting

$$g_z \equiv \{-D^*A^*S^*(T-z)[S(T)y_0 + L_T u^0],\ z \in \mathscr{F}\},$$

well defined by Remark 2.1. Notice that, at $z = T$, g_z is not well defined, so extension to $\bar{\mathscr{F}}$ is not allowed. Therefore, while $g_z \notin \mathscr{A}(\mathscr{F}; L_2(\Gamma))$, what holds instead is

$$g_z \text{ is an } L_2(\Gamma)\text{-function, analytic in } \bar{\mathscr{F}} - \{T\}, \tag{2.17}$$

since $S^*(z)L_2(\Omega) \subset \mathscr{D}(A^*)$ for $z \neq 0$, and $\bar{\mathscr{F}}$ was chosen properly contained in the sector of analyticity of $S(\cdot)$, hence of $S^*(\cdot)$.

[18] By standard isomorphism techniques; see e.g. [L5], [L9].

Remark 2.3. If $\alpha = 0$ in (1.0), the term g is missing from (2.11). Interpreted for $z \in \bar{\mathcal{F}}$, (2.11) is in this case

$$(I + L^*L)u^0 = b_z.$$

Later arguments will show that the needed inversion is justified in $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$ and that $u^0 = (I + L^*L)^{-1} b_z$. Restricting to real t, we conclude that: *The optimal control* $u^0(t)$ *for the problem* (P_Σ), *with* $\alpha = 0$, *is an* $L_2(\Gamma)$*-function, analytic for* $0 < t < T$, *and continuous in* $[0, T]$. This conclusion complements, but does not replace, the results of [L1]–[L3] over $[0, T]$, recalled in the Introduction.

Step D. *A more suitable expression for* u^0. Since, as remarked, $g_z \notin \mathcal{A}(\mathcal{F}; L_2(\Gamma))$, we are not allowed to write $(I + L^*L)^{-1}(b_z + g_z)$ in $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$. To overcome this difficulty, we rewrite (2.11) with complex variable z as

$$\begin{aligned}(I + L^*L)u^0 = (I + L^*L)(b_z + g_z) - (I + L^*L)L^*L(b_z + g_z) \\ + (L^*L)^2(b_z + g_z),\end{aligned} \tag{2.18}$$

obtained by adding and subtracting the same quantities. It will be shown in the Appendix that:

$$(L^*L)^2 g_z \text{ is an } L_2(\Gamma)\text{-function continuous in } \bar{\mathcal{F}}, \tag{2.19}$$

so that, by (2.16), (2.17) and Lemma 2.1, it follows that

$$(L^*L)^2(b_z + g_z) \in \mathcal{A}(\mathcal{F}; L_2(\Gamma)). \tag{2.20}$$

If we can show that $(I + L^*L)$ is invertible with bounded *inverse in* $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$, then (2.18) provides

$$u^0 = b_z + g_z - L^*L(b_z + g_z) + (I + L^*L)^{-1}(L^*L)^2(b_z + g_z), \tag{2.21}$$

which exhibits u^0 as an $L_2(\Gamma)$-function analytic over $\mathcal{F}$, as desired. In fact, analyticity of the first two terms is contained in (2.16)–(2.17), analyticity of the third term stems from Remark 2.2(b) via (A.1) of the Appendix, and analyticity of the fourth term is contained in (2.20).

Therefore, all that remains to show is:

Step E. $(I + L^*L)$ *is invertible with bounded inverse in* $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$.[19] The value $\lambda = -1$ cannot be an eigenvalue of the operator in L^*L in $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$ since, as we know, it is not an eigenvalue of L^*L in $L_2(\mathcal{F}; L_2(\Gamma))$. To complete step E, it is therefore enough to show that L *is compact as an operator from* $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$ *into* $\mathcal{A}(\mathcal{F}; L_2(\Omega))$. The compactness of L as an operator from $C([0, T]; L_2(\Gamma))$ into $C([0, T]; L_2(\Omega))$ was already observed (with no explicit proof given) in [B2, below (4.1)]. As a matter of fact, the following standard proof is technically the same on C or on $\mathcal{A}$.

The operators L_δ defined from $\mathcal{A}(\mathcal{F}; L_2(\Gamma))$ into $\mathcal{A}(\mathcal{F}; L_2(\Omega))$ by

$$(L_\delta u)(z) = \int_0^{z - \delta e^{i\theta}} A^{\frac{3}{4}+\varepsilon} S(z - \zeta) A^{\frac{1}{4}-\varepsilon} Du(\zeta)\, d\zeta, \qquad z = re^{i\theta}, \quad \zeta = \rho e^{i\theta},$$

converge to L in the uniform operator norm as $\delta \downarrow 0$

$$\begin{aligned}|(L_\delta - L)u|_{\mathcal{A}(\mathcal{F}; L_2(\Gamma))} &\leqq \sup_{z \in \bar{\mathcal{F}}} \left\{ \int_{r-\delta}^{r} \frac{K}{(r-\rho)^{\frac{3}{4}+\varepsilon}} |u(\zeta)|_{L_2(\Gamma)} |d\zeta| \right\} \\ &\leqq \delta^{\frac{1}{4}-\varepsilon} K |u|_{\mathcal{A}(\mathcal{F}; L_2(\Gamma))}.\end{aligned}$$

[19] This is the only step where our proof conceptually benefits from [S1], although it is technically much simplified over [S1]. In fact, our analysis has revealed the perturbation of the identity factorized as L^*L; thus while [S1] has to prove compactness of the full perturbation, we need only prove compactness of its factor L, a much easier task.

To show that L_δ is compact for fixed δ, we apply the generalized Ascoli theorem [R1, p. 179] to the family of functions $\{L_\delta u_k\}$ from $\mathscr{F}$ into $L_2(\Omega)$, with unit norm in $\mathscr{A}(\mathscr{F}; L_2(\Gamma))$ for the u_k. The family is plainly equicontinuous (in k), the kernel being continuous. Moreover, for each fixed $z \in \bar{\mathscr{F}}$, the totality of points

$$(L_\delta u_k)(z) = A^{-1} \int_0^{z-\delta e^{i\theta}} A^2 S(z-\zeta) Du(\zeta)\, d\zeta$$

clearly lies in a precompact set of $L_2(\Omega)$, since the integral points lie in a bounded set of $L_2(\Omega)$ and A^{-1} is compact. The Ascoli theorem then guarantees uniform convergence on $\bar{\mathscr{F}}$, i.e., convergence in $\mathscr{A}$, of a subsequence $\{L_\delta u_{k_n}\}$, and L_δ is compact.

The proof of analyticity of u^0 in $\mathscr{F}$, hence in $(0, T)$, is thus complete.

Note, however, that all terms on the right side of (2.21) are continuous at the origin $z = 0$. Thus, $u^0(t)$ is continuous at $t = 0$. The properties of u^0 in Theorem 1.1 are thus all proved.

2.2. Properties of the optimal trajectory y^0. Once analyticity of the optimal control u^0 as $L_2(\Gamma)$-function in $\mathscr{F}$ is established, it follows (see the proof of Lemma 2.1, following (2.15)) that the optimal trajectory y^0 as $L_2(\Omega)$-function is likewise analytic in $\mathscr{F}$, hence in $(0, T)$. As to continuity of $y^0(t)$ at $t = T$, we readily deduce from (2.2) with $u = u^0$ that $\lim [y^0(t), x] = [y^0(T), x]$ as $t \uparrow T$, (the last term is well defined by (2.8)) at least when, say, $x \in \mathscr{D}(A^*)$. Furthermore, this limit can be extended to all x in $L_2(\Omega)$, hence in the strong topology of $L_2(\Omega)$—since $\mathscr{D}(A^*)$ is dense in it and $y^0(t)$ is uniformly bounded[20] in the $L_2(\Omega)$-norm on $[0, T]$. □

Proof of Corollary 1.2. The proof is based on the following steps:

(i) The vectors b_t and g_t are analytic in $(0, T)$ with values in any $H^s(\Gamma)$, $s \geqq 0$.

(ii) The operator L maps $\{h(t): h(t) \in H^2(\Gamma)$, and analytic in $(0, T)\} \cap L_2(0, T; L_2(\Gamma))$

into

$$\{\phi(t): \phi(t) \in H^{s+\frac{1}{2}}(\Omega), \text{ and analytic in } (0, T)\} \cap L_2(0, T; L_2(\Omega)),$$

for $s \geqq 0$. A similar statement holds for L^*, mutatis mutandis, i.e., Γ with Ω.

(iii) By (i) and (ii), an iteration procedure applied to (2.11), now rewritten as $u^0 = -L^* L u^0 + b + g$, provides the desired conclusion.

To justify (i), we note that $\mathscr{D}(A^n) = \mathscr{D}(A^{*n})$, $n = 0, 1, \cdots$ and that

$$\int_t^T A^* S^*(\tau - t) S(\tau) y_0\, d\tau \in \mathscr{D}^\infty(A^*) \equiv \bigcap_{n=1}^{\infty} \mathscr{D}(A^{*n}) \subset \bigcap_{n=1}^{\infty} H^{2n}(\Omega),$$

and therefore $b_t = -\int_t^T D^* A^* S^*(\tau - t) S(\tau) y_0\, d\tau$ has values in any $H^s(\Gamma)$.

The analyticity of b_t in $(0, T)$ is settled, as in Lemma 2.1, by integration by parts. As to g_t, the desired conclusions are apparent from its very definition in (2.12).

(ii) For a function u which is in $L_2(0, T; L_2(\Gamma))$ and which is also an $H^2(\Gamma)$-function analytic in $(0, T)$, we can write:

$$(Lu)(t) = AS(t-\varepsilon) \int_0^\varepsilon S(\varepsilon - \tau) Du(\tau)\, d\tau + \sum_{l=0}^{k} A^{-l} Du^{(l)}(t)$$

$$+ \sum_{l=0}^{k} A^{-l} S(t-\varepsilon) Du^{(l)}(\varepsilon) - A^{-k+1} \int_\varepsilon^t S(t-\tau) Du^{(k+1)}(\tau)\, d\tau,$$

[20] Of course, $y^0(T)$ is a priori known to be a well defined vector in $L_2(\Omega)$.

where $k(s)=k$ is an integer such that $k<s/2<k+1$. This expression is obtained by proceeding as above, following (2.15), except that now integration by parts over (ε, T) is carried out $(k+1)$ times (see, e.g. [L3, Eq. (5.1)]). Then the first and third terms in (2.23) are analytic and in $\mathscr{D}^{\infty}(A)\subset H^{\infty}(\Omega)$ for $t>\varepsilon$. (The first integral in (2.23) is well defined by Remark 2.2(b).) The second term in (2.23) is analytic in $(0, t)$, since so is $u^{(l)}(t)$, and in $H^{s+\frac{1}{2}}(\Omega)$ $(l=0)$. Moreover, since $u^{(k+1)}$ is analytic on (ε, t), the term

$$A\int_{\varepsilon}^{t} S(t-\tau)Du^{(k+1)}(\tau)\,d\tau \in H^{\frac{1}{2}}(\Omega)$$

and is analytic. Therefore, the fourth term in (2.23) is in $H^{\frac{1}{2}+2k}(\Omega)\subset H^{\frac{1}{2}+s}(\Omega)$, as desired, as A^{-k} is bounded from $H^{\alpha}(\Omega)\to H^{\alpha+2k}(\Omega)$, $\alpha\geqq 0$ (e.g., [L3, following (6.8iii)]).

A similar analysis applies to L^*, after we notice that D^* is continuous in $H^s(\Omega)\to H^{s+\frac{1}{2}}(\Gamma)$. This is so either as a consequence of D being an isomorphism[21] from $H^{-s-\frac{1}{2}}(\Gamma)$ onto $D_A^{-s}(\Omega)\subset(H^2(\Omega))'$ [L5, Thm. 5.7, p. 179; (6.24), (6.20)], or else (more simply, in the present Dirichlet second-order case) as a consequence of Green's formula $(h, Dv)_{\Omega}=((\partial/\partial\eta)w, v)_{\Gamma}$ where, for all $h\in H^s(\Omega)$, there is $w\in H^{s+2}(\Omega)$, $s\geqq 0$, such that $-A(\xi,\partial)w=h$ in Ω and $w|_{\Gamma}=0$. Then,[21] $(\partial/\partial\eta)w\in H^{s+\frac{1}{2}}(\Omega)$ and $(h, Dv)_{\Omega}$ is well defined for all $h\in H^s(\Omega)$ and all $v\in(H^{s+\frac{1}{2}}(\Omega))'$, by extending inner products to duality pairings; i.e., D is continuous: $(H^{\frac{1}{2}+s}(\Gamma))'=H^{-\frac{1}{2}-s}(\Gamma)\to(H^2(\Omega))'$.

As a third alternative route, one may use Kellogg's review [A1, Thm. 3.8.1, p. 71]. □

3. Feedback synthesis of optimal control via a Riccati operator. In the present section, the problem of synthesizing the optimal control $u^0(t)$ as a "real time" feedback of the optimal solution $y^0(t)$ is considered. Informally, what one seeks is a realization of $u^0(t)$ as expressed by

$$(*)\qquad u^0(t)=CP(t)y^0(t)$$

pointwise in $0\leqq t<T$. Here C is an operator known in terms of the original parabolic system (1.1)–(1.3) while the operator $P(t)$ is expected (from finite [C3] and some infinite dimensional feedback theory, e.g., [B1], [C1]–[C3], [G1], [L4], [L8]) to be a Riccati operator, i.e., to satisfy in an appropriate sense a Riccati operator equation (quadratic) for $0\leqq t<T$. The real time ("on line") realization $(*)$ should be contrasted with (2.10), which gives, instead, u^0 as an $L_2(\Gamma)$-trajectory over $[0, T]$ (analytic over $(0, T)$ and continuous at $t=0$, cf. Theorem 1.1) in terms of the initial datum y_0 and of the operators describing $y^0(t)$ as an $L_2(\Omega)$-trajectory. In order to achieve the pointwise feedback realization $(*)$ of u^0, what is clearly needed is a pointwise description of the dynamics of $y^0(t)$ in terms of the initial datum at an arbitrary initial time: such description will be accomplished by an appropriate evolution operator, as described below. Henceforth, unless otherwise stated explicitly, our analysis *will refer specifically to the case where the performance index penalizes also the final state* (i.e., $\alpha=1$ in eq. (1.0)). This inclusion, as we shall see, adds further technical difficulties to the problem of synthesis through a Riccati operator. Corresponding relevant results without penalization of the final state ($\alpha=0$ in (1.0)) will be relegated to the footnotes.

Step. 1. *Derivation of* $\Phi(t, s)$ *and* $P(t)$ *for* $\alpha=1$ *in* (1.0)., Let s be an arbitrary time $0\leqq s<T$. Henceforth, we take s as the new initial time of our control problem with corresponding initial datum $y_s\in L_2(\Omega)$; i.e., we consider the optimal control problem of the introduction over the time interval $[s, T]$, rather than over $[0, T]$. We

[21] These are the only places where smoothness of the boundary Γ is used.

shall denote[22] the corresponding optimal solution by $y^0(t, s; y_s)$ and $u^0(t, s; y_s)$. The procedure of § 2 leading to (2.9a) and (2.10), once applied to the new problem, gives for $u^0(\cdot, s; y_s)$ as an element of $L_2(s, T; L_2(\Omega))$ the following corresponding expressions:

$$-u^0(\cdot, s; y_s) = L_s^* y^0 + L_{sT}^* y^0(T) \tag{3.1}$$

and

$$-u^0(\cdot, s; y_s) = [I_s + L_s^* L_s + L_{sT}^* L_{sT}]^{-1}[L_s^* S(\cdot - s)y_s + L_{sT}^* S(T-s)y_s], \tag{3.2}$$

where L_s and L_{sT} are the operators obtained, respectively, from the operators L and L_T by changing the initial time from zero to s. Explicitly, L_s is given by

$$(L_s u)(t) = \int_s^t A^{\frac{3}{4}+\rho} S(t-\tau) A^{\frac{1}{4}-\rho} Du(\tau)\, d\tau, \qquad s \leqq t \leqq T, \tag{3.3a}$$

bounded from $L_2(s, T; L_2(\Gamma))$ into $L_2(s, T; L_2(\Omega))$ (in fact, even $L_2(s, T; H^{\frac{1}{2}}(\Omega))$. Its corresponding adjoint is now

$$(L_s^* v)(t) = \int_t^T D^* A^* S^*(\tau - t) v(\tau)\, d\tau, \qquad s \leqq t \leqq T, \tag{3.3b}$$

bounded from $L_2(s, T; L_2(\Omega))$ into $L_2(s, T; L_2(\Gamma))$. The operator L_{sT} is given by

$$L_{sT} u = \int_s^T A^{\frac{3}{4}+\rho} S(T-\tau) A^{\frac{1}{4}-\rho} Du(\tau)\, d\tau, \tag{3.4a}$$

unbounded from $L_2(s, T; L_2(\Gamma)) \supset \mathscr{D}(L_{sT})$ into $L_2(\Omega)$, while its dual is now

$$L_{sT}^* y = D^* A^* S^*(T-t) y, \qquad s \leqq t < T, \tag{3.4b}$$

unbounded from $L_2(\Omega) \supset \mathscr{D}(L_{sT}^*)$ into $L_2(s, T; L_2(\Gamma))$.

Substituting $u^0(\cdot, s; y_s)$ from (3.2) into the dynamics

$$y^0(t, s; y_s) = S(t-s)y_s + (L_s u^0(\cdot, s; y_s))(t), \qquad 0 \leqq s \leqq t \leqq T, \tag{3.5a}$$

$$(L_s u^0(\cdot, s; y_s))(t) = \int_s^t A^{\frac{3}{4}+\rho} S(t-\tau) A^{\frac{1}{4}-\rho} Du^0(\tau, s; y_s)\, d\tau \tag{3.5b}$$

yields

$$y^0(t, s; y_s) = \Phi(t, s) y_s, \qquad 0 \leqq s \leqq t \leqq T, \tag{3.6}$$

where, for $x \in L_2(\Omega)$, we have defined Φ by

$$\Phi(t, s)x \equiv S(t-s)x + \{L_2[I_s + L_s^* L_s + L_{sT}^* L_{sT}]^{-1}[L_s^* S(\cdot - s)x + L_{sT}^* S(T-s)x]\}(t). \tag{3.7}[23]$$

The operator $\Phi(t, s)$ is well defined on the triangle $\Delta_T \equiv \{(s, t): 0 \leqq s \leqq t \leqq T\}$ and acts from all of $L_2(\Omega)$ into itself; moreover, at least by virtue of (3.7), $\Phi(\cdot, s)$ is a bounded operator from $L_2(\Omega)$ into $L_2(s, T; L_2(\Omega))$. But, in fact, $\Phi(t, s)x$ is continuous in $t \in [s, T]$, and thus $\Phi(\cdot, s)$ is a *bounded operator from* $L_2(\Omega)$ *into* $C([s, T]; L_2(\Omega))$, by virtue of the closed graph theorem, applied to the composition of the following three maps:

$$\begin{array}{ccccccc} y_s & \xrightarrow[\text{by (3.2)}]{\text{bounded}} & u^0(\cdot, s; y_s) & \xrightarrow[\text{by (3.5)}]{\text{bounded}} & A^{-1} y^0(\cdot, s; y_s) & \xrightarrow{A} & y^0(\cdot, s; y_s) \\ L_2(\Omega) & & L_2(s, t; L_2(\Omega)) & & C([s, T]; L_2(\Omega)), & & \end{array}$$

[22] In the new notation, the optimal solution in $[0, T]$, so far denoted by $y^0(t)$ and $u^0(t)$, will be $y^0(t, 0; y_0)$ and $u^0(t, 0; y_0)$, respectively.

[23] Compare with [G1, Eq. (3.18)].

along with A being closed and invertible [K1, Prob. 5.7, p. 164]. Some of the properties of $\Phi(t, s)$ will be recorded below in Proposition 3.2 (and subsequent Proposition 3.6), which will show that $\Phi(t, s)$ is the desired evolution operator.

Substituting (3.6) into (3.1) yields

$$-u^0(\cdot, s; y_s) = L_s^*(\Phi(\cdot, s)y_s) + L_{sT}^*(\Phi(T, s)y_s),$$

i.e., explicitly, by (3.3a) and (3.4b).

$$-u^0(t, s; y_s) = \int_t^T D^*A^*S^*(\tau - t)\Phi(\tau, s)y_s\, d\tau + D^*A^*S^*(T-t)\Phi(T, t)y_s, \tag{3.8}$$

$$0 \leqq t < T.$$

If we now choose the initial time s equal to t with corresponding initial datum $y^0(t)$, we obtain from (3.8) the desired pointwise relation

$$-u^0(t) = D^*A^*P(t)y^0(t), \qquad 0 \leqq t < T, \tag{3.9}$$

where, for $x \in L_2(\Omega)$, the operator $P(t)$ is defined by

$$P(t)x \equiv \int_t^T S^*(\tau - t)\Phi(\tau, t)x\, d\tau + S^*(T-t)\Phi(T, t)x, \qquad 0 \leqq t < T. \tag{3.10}$$[24]

Note that, via (3.7) and (3.3)–(3.4), the operator $P(t)$ in (3.10) is given in a constructive manner explicitly in terms of the original parabolic system.

Remark 3.1. As remarked in the Introduction, our approach in studying the quadratic cost problem is within the spirit of Balakrishnan's in that input-solution operators (e.g., L, L^* etc.) in appropriate function spaces are stressed in place of, and substitute for, the equivalent original parabolic equation; yet his approach and ours also differ in some important aspects. In his treatment of the quadratic cost problem with distributed (rather than boundary) control and no final state (cf. [B1, § 5.2, p. 229]) as an abstract ordinary differential equation in Hilbert space, Balakrishnan obtains a formula [B1, Eq. (5.2.12)] which coincides with the first term of our (3.10) above (the second term in (3.10) refers to the final state). There is, however, an important difference in the way this formula occurs, and may be employed, in his development, as opposed to ours, which we now explain. Balakrishnan's approach first postulates the existence of a feedback operator $P(t)$, then introduces the corresponding evolution operator $\Phi(t, s)$ depending on the postulated $P(t)$, hence obtains that $P(t)$ and $\Phi(t, s)$ are related to each other by his equation (5.2.12). Therefore, $P(t)$ given by his equation (5.2.12) is available only implicitly, and not in a direct, computable form. In the final stage of his analysis, Balakrishnan proves that $P(t)$ has to be, in fact, the unique selfadjoint solution of a Riccati equation. (The existence issue is proved via a stochastic technique in the context of the filtering problem, treated in his subsequent chapter.) In Balakrishnan's approach, therefore, actual determination of $P(t)$ rests exclusively with the numerical solution of the Riccati equation. By contrast, our equation (3.10) provides $P(t)$ in an explicit, direct way, as we have already derived an expression for the evolution operator $\Phi(t, s)$ based on the original parabolic equation (our (3.7) above), whose counterpart is missing in Balakrishnan's treatment. We shall prove below that our $P(t)$, constructively defined by (3.10), does satisfy a Riccati equation. The existence issue of the Riccati equation is thus automatically taken care of in our treatment.

[24] Compare with [G1, Eq. (3.25)].

It is also expected that (3.10) can be used for convergence analysis of numerical approximation schemes of $P(t)$.

Finally, comments on the connection with Gibson's work [G1] for the regulator problem were already made in the Introduction, and are being reinforced in the footnotes.

Step 2. *Properties of* $\Phi(t, s)$ *and* $P(t)$. Our next major goal is to show that the operator $P(t)$ defined by (3.10) is, in fact, a Riccati operator, i.e., that it satisfies a Riccati (quadratic) differential equation. To this end, some preliminary properties of $\Phi(t, s)$ and $P(t)$ are needed. We begin with a lemma.

LEMMA 3.1. *With the notation previously introduced* (*see footnote* 22), *for any* $0<\varepsilon<T-s$, *the following uniform bound in s holds*:

$$|u^0(\cdot, s; y_s)|_{C([s,T-\varepsilon];L_2(\Gamma))} \leqq C_{T-\varepsilon}|y_s|_{L_2(\Omega)} \tag{3.11a}$$[25]

where $C_{T-\varepsilon}$ *is a constant depending on* $T-\varepsilon$, *but not on s. Moreover,*

$$|u^0(\cdot, s; y_s)|_{L_2(s,T;L_2(\Gamma))} \leqq C_T|y_s|_{L_2(\Omega)}. \tag{3.11b}$$

Proof. With V_s the selfadjoint operator, $V_s \equiv I_s + L_s^* L_s + L_{sT}^* L_{sT}$ on $L_2(s, T; L_2(\Omega))$, conclusion (3.11b) is a consequence of

$$|V_s^{\frac{1}{2}}x|_{L_2(s,T;L_2(\Omega))} = (V_s x, x)_{L_2(s,T;L_2(\Omega))} \geqq |x|_{L_2(s,T;L_2(\Omega))}, \quad x \in \mathscr{D}(V_s^{\frac{1}{2}}),$$

obtained as in footnote 14. Thus, $V_s^{-\frac{1}{2}}$, and hence V_s^{-1}, is uniformly bounded in s. Equation (3.11b) then follows from (3.2).

To prove (3.11a) we use the counterpart of (2.21) with initial time s; i.e.

$$(\#) \qquad u^0(\cdot, s; y_s) = (I_s - L_s^* L_s)[b_{st} + g_{st}] + [I_s + L_s^* L_s]^{-1}(L_s^* L_s)^2[b_{st} + g_{st}],$$

where

$$b_{st} \equiv \{-L_s^* S(t-s)y_s,\ s \leqq t \leqq T-\varepsilon\},$$

$$g_{st} \equiv \{-D^* A^* S^*(T-t)y^0(T, s; y_s),\ s \leqq t \leqq T-\varepsilon\},$$

so that $b_{st}, g_{st} \in C([s, T-\varepsilon], L_2(\Gamma))$. Next, for a function $f(\cdot, s)$ defined for $s \leqq \cdot \leqq T-\varepsilon$, the following three inequalities hold, uniformly in s:

$$|L_s f(\cdot, s)|_{C([s,T-\varepsilon];L_2(\Omega))} \leqq C_{T-\varepsilon}|f(\cdot, s)|_{C([s,T-\varepsilon];L_2(\Gamma))}, \tag{i}$$

$$|L_s^* f(\cdot, s)|_{C([s,T-\varepsilon];L_2(\Omega))} \leqq C_{T-\varepsilon}|f(\cdot, s)|_{C([s,T-\varepsilon];L_2(\Omega))}, \tag{ii}$$

$$|(L_s^* L_s)^2 f(\cdot, s)|_{C([s,T-\varepsilon];L_2(\Omega))} \leqq C_{T-\varepsilon}|f(\cdot, s)|_{L_2(s,T-\varepsilon;L_2(\Gamma))}. \tag{iii}$$

Here and below $C_{T-\varepsilon}$ denotes a generic constant depending on $T-\varepsilon$ but not on s, and f is assumed in each case to have the appropriate regularity claimed in the respective right-hand side. The first two inequalities are obvious, while the third is obtained by examination of the proof in the Appendix (simplified, as now the extension to complex variables is not needed).

Now *the operator* $I_s + L_s^* L_s$ *has a bounded inverse in* $C([s, T_1]; L_2(\Gamma))$, *for any* $s \leqq T_1 \leqq T$, *which is uniformly bounded in s.*

In fact, given $y = y(\cdot, s)$ in this space, we seek a unique $x = x(\cdot, s)$ in the same space, which solves the equation

$$x + L_s^* L_s x = y.$$

[25] If $\alpha = 0$ in (1.0) (i.e., there is no final state in the performance index), then ε can be taken to be equal to zero in (3.11).

That a unique solution $x \in L_2(s, T_1; L_2(\Gamma))$ exists is a consequence of footnote 14, and moreover

$$(\#\#) \qquad |x(\cdot, s)|_{L_2(s,T_1;\, L_2(\Gamma))} \leqq C_{T_1} |y(\cdot, s)|_{L_2(s,T_1;L_2(\Gamma))}$$

holds. But then $(L_s^* L_s)^2 x$ is, in fact, in $C([s, T_1]; L_2(\Gamma))$ (from Lemma A.1). From

$$L_s^* L_s x = -(L_s^* L_s)^2 x + L_s^* L_s y$$

we next deduce that $L_s^* L_s x \in C([s, T_1]; L_2(\Gamma))$ and, returning to the above original equation, that $x \in C([s, T_1]; L_2(\Gamma))$. Moreover, inequalities (i)–(iii) and $(\#\#)$ give

$$\text{(iv)} \qquad |[I_s + L_s^* L_s]^{-1} y(\cdot, s)|_{C([s,T_1];L_2(\Gamma))} \leqq C_{T_1} |y(\cdot, s)|_{C([s,T_1];L_2(\Gamma))}.$$

Applying inequalities (i)–(iv) to the above equation $(\#)$, we obtain

$$|u^0(\cdot, s; y_s)|_{C([s,T-\varepsilon];L_2(\Gamma))} \leqq C_{T-\varepsilon} |b_{st} + g_{st}|_{C([s,T-\varepsilon];L_2(\Gamma))}.$$

But, as desired,

$$|b_{st}|_{C([s,T-\varepsilon];L_2(\Gamma))} \leqq C_{T-\varepsilon} |y_s|_{L_2(\Omega)}$$

from the definition of b_{st}. As to g_{st}, we have, with $y^0(T, s; y_s) = S(T-s)y_s + L_{sT} u^0(\cdot, s; y_s)$,

$$(***) \qquad \begin{aligned} |g_{st}|_{C([s,T-\varepsilon];L_2(\Gamma))} &\leqq |D^* A^* S^*(T-t) S(T-s) y_s|_{C([s,T-\varepsilon];L_2(\Gamma))} \\ &\quad + C_{T-\varepsilon} |u^0(\cdot, s; y_s)|_{L_2(s,T;\, y_s)}, \end{aligned}$$

since the integral kernel $D^* A^* S^*(T-t) A S(T-\tau) D$ of L_{sT}, rewritten as

$$(D^* A^{*\frac{1}{4}-\rho})(A^{*+2\rho} S^*(T-\tau))(A^{*--\rho} A^{+\rho})(S(T-\tau) A^{\frac{1}{4}-\rho} D)$$

is strongly continuous, as $t \leqq T - \varepsilon$. Application of (3.11b) to $(***)$ then yields (3.11a), as desired. □

PROPOSITION 3.2. *For the operator $\Phi(t, s)$, defined by* (3.7), $0 \leqq s \leqq t \leqq T$, *as a bounded operator from $L_2(\Omega)$ into itself, the following properties hold true*:

(i) $\Phi(t, t) = I$ (*identity on* $L_2(\Omega)$), $0 \leqq t < T$.

(ii) $\Phi(t, s)\Phi(s, \tau) = \Phi(t, \tau)$ (*transition*), $0 \leqq \tau \leqq s < t \leqq T$.

(iii) *For each fixed* s, $0 \leqq s < T$, *the operator* $\Phi(t, s)$ *is strongly continuous in* $t \in [s, T]$, *and is actually analytic in* (s, T).

(iv) *For any* ε, $0 < \varepsilon < T$, *the operator* $\Phi(t, s)$ *is uniformly bounded in* $(s, t) \in \Delta_{T-\varepsilon}$; *i.e.*,

$$(3.12)^{26} \qquad |\Phi(t, s)| \leqq M_{T-\varepsilon}, \text{ for all } 0 \leqq s \leqq t \leqq T - \varepsilon.$$

(v) *For* $t < T$, *the operator* $\Phi(t, s)$ *is strongly continuous in* $s \in [0, t]$ (*left continuity at* $s = t$); *as to* $t = T$, *the operator* $\Phi(T, s)$ *is strongly continuous in* $s \in [0, T)$. *Moreover,*

$$\lim_{t \uparrow T} A^{-\frac{1}{4}-\varepsilon} \Phi(T, t) x = A^{-\frac{1}{4}-\varepsilon} x, \qquad x \in L_2(\Omega),$$

and likewise

$$A^{*-\frac{1}{4}-\varepsilon} \Phi(T, t) x \to A^{*-\frac{1}{4}-\varepsilon} x.$$

(vi) *Let* $0 \leqq s < t < \tau \leqq T$. *The following identity holds*:[27]

$$(3.13) \qquad \frac{\partial \Phi(\tau, t)}{\partial t} \Phi(t, s) x = -\Phi(\tau, t) \frac{\partial \Phi}{\partial t}(t, s) x, \qquad x \in L_2(\Omega).$$

---

[26] If $\alpha = 0$ in (1.0), i.e. (no final state in the performance index), then ε can be taken as equal to zero in (3.12).

[27] It reduces the second argument derivative of Φ, computed along the optimal trajectory $y^0(t, s; x)$, in terms of its first argument derivative.

Proof. Properties (i) and (ii) follow from (3.6), via uniqueness of the optimal control trajectory y^0 initiating at τ and s, in the usual way. Property (iii) is a restatement of the properties of $y^0(t, s; y_s)$ via (3.6).

Property (iv) is a consequence of Lemma 3.1 and (3.5)–(3.6) and the principle of uniform boundedness.

Property (v). With $s<t<T$ given, choose $h>0$ and $\varepsilon>0$ such that $s+h<t\leqq T-\varepsilon$. Then, right continuity follows from

$$\begin{aligned}|\Phi(t, s+h)x-\Phi(t, s)x| &= |\Phi(t, s+h)[x-\Phi(s+h, s)x]| \\ &\leqq M_{T-\varepsilon}|\Phi(s+h, s)x-x| \qquad \text{(by (iv))}\end{aligned}$$

and the right-hand side goes to zero by (iii). The case $t=T$ is now reduced to the previous one:

$$\Phi(T, s+h)x-\Phi(T, s)x=\Phi(T, t)[\Phi(t, s+h)x-\Phi(t, s)x]$$

with $s+h<t<T$.

As to left continuity, we compute for $s\leqq t<T$ and $h>0$

$$\begin{aligned}|\Phi(t, s-h)x-\Phi(t, s)x| &= |\Phi(t, s)[\Phi(s, s-h)x-x]| \\ &\leqq |\Phi(t, s)|\,|\Phi(s, s-h)x-x|.\end{aligned}$$

That the right-hand side goes to zero as $h\downarrow 0$ is now a consequence of (3.5), whereby

$$|\Phi(s, s-h)x-x|\leqq |S(h)x-x|+\int_{s-h}^{s}|A^{\frac{3}{4}+\rho}S(s-\tau)A^{\frac{1}{4}-\rho}Du^0(\tau, s-h; x)|\,d\tau$$

follows. Then, the above integral goes to zero as $h\downarrow 0$ by Lemma 3.1 (since $s\leqq T-\varepsilon$, for suitable $\varepsilon>0$) combined with the kernel $C/(s-\tau)^{\frac{3}{4}+\rho}$ (cf. (2.14)).

To complete property (v), we must prove that, for $x\in L_2(\Omega)$, the limit as $t\uparrow T$ of

$$A^{-\frac{1}{4}-\varepsilon}\Phi(T, t)x=S(T-t)A^{-\frac{1}{4}-\varepsilon}x+\int_t^T A^{\frac{1}{2}-\varepsilon+\rho}S(T-\tau)A^{\frac{1}{4}-\rho}Du^0(\tau, t; x)\,d\tau$$

is $A^{-\frac{1}{4}-\varepsilon}x$. By the Schwarz inequality, this is a consequence of

$$|A^{\frac{1}{2}-\varepsilon+\rho}S(T-\tau)|\leqq C/(T-\tau)^{\frac{1}{2}-\varepsilon+\rho}\in L_2(t, T)$$

(by taking ρ less than the preassigned ε) and of the uniform bound (in t):

$$\int_t^T |u^0(\tau, t; x)|^2\,d\tau\leqq \text{const}_T|x|^2$$

which, in fact, was established in (3.11b). The proof of $A^{-\frac{1}{4}-\varepsilon}$ is complete. The analogous result for $A^{*-\frac{1}{4}-\varepsilon}$ follows then from $|A^{*-\frac{1}{4}-\varepsilon}A^{\frac{1}{4}+\varepsilon}|\leqq$ const by the closed graph theorem.

Property (vi). For $h>0$ so that $t+h<\tau\leqq T$, we compute

$$\frac{1}{h}[\Phi(\tau, t+h)\Phi(t, s)x-\Phi(\tau, t)\Phi(t, s)x]=\Phi(\tau, t+h)\frac{1}{h}[\Phi(t, s)x-\Phi(t+h, t)\Phi(t, s)x].$$

By differentiability of Φ in its first argument at t, since $s<t<T$ (property (ii)) and by right continuity of Φ in its second argument, since $t<\tau\leqq T$ (property (v)), we deduce identity (3.13) for the right derivative.

Similarly for the left derivative: with $h>0$ so that $s \leqq t-h$, we compute,

$$\frac{1}{h}[\Phi(\tau, t-h)\Phi(t, s)x - \Phi(\tau, t)\Phi(t, s)x] = \Phi(\tau, t-h)\frac{1}{h}[\Phi(t, s)x - \Phi(t-h, s)x],$$

and the desired conclusion follows again from (ii) and (v). □

Preliminary properties of the operator $P(t)$ are collected next.

PROPOSITION 3.3. *For the operator $P(t)$ from $L_2(\Omega)$ into itself defined by* (3.10), *the following properties hold*:

(i) *$P(t)$ is a bounded operator on $L_2(\Omega)$ for $0 \leqq t \leqq T$.*

(ii) *For each $t \in [0, T)$ and $0<\varepsilon<1$, the following inclusion holds: range of $P(t) = P(t)L_2(\Omega) \subset \mathcal{D}(A^{*1-\varepsilon})$.*

1) *Thus, by the closed graph theorem, the operator $A^{*1-\varepsilon}P(t)$ is a well defined bounded operator on $L_2(\Omega)$.*

2) *Moreover, the operator $D^*A^*P(t)$ is a well defined bounded operator from $L_2(\Omega)$ into $L_2(\Gamma)$, and, in fact, for any $0<\varepsilon<T$ and $x \in L_2(\Omega)$,*

$$D^*A^*P(t)x \in C([0, T-\varepsilon]; L_2(\Gamma)).$$

(iii) *For each $t \in [0, T)$, the following identity, symmetric in x and y (both in $L_2(\Omega)$) holds.*[28]

$$[P(t)x, y] = \int_t^T [\Phi(\tau, t)x, \Phi(\tau, t)y]\, d\tau + [\Phi(T, t)x, \Phi(T, t)y] + \int_t^T (D^*A^*P(\tau)\Phi(\tau, t)x, D^*A^*P(\tau)\Phi(\tau, t)y)\, d\tau. \tag{3.14}$$[29]

Thus

1) *$P(t) = P^*(t)$, and $P(t)$ is selfadjoint.*

2) *$P(t)$ is positive definite.*

(iv) *The minimal (optimal) value of the performance index J as in* (1.0) *of the optimal control problem on $[s, T]$, $s<T$ that initiates at y_s at time s is*

$$J^0(u^0(\cdot, s; y_s), y^0(\cdot, s; y_s)) = [P(s)y_s, y_s].$$

Hence, for any $x \in L_2(\Omega)$, the map $t \to [P(t)x, x]$ is monotone decreasing.

(v) *$P(t)$ is uniformly bounded on $[0, T]$: $|P(t)| \leqq C_T$, $0 \leqq t \leqq T$, and*[30] *$\lim_{t \uparrow T} P(t)x = x$ for all $x \in L_2(\Omega)$.*

Proof. Property (i). Boundedness of $P(t)$ is immediate from (3.10), once continuity of $\Phi(\tau, t)x$ in $\tau \in [t, T]$ (property (iii) of Proposition 3.2) is used.

Property (ii). This same fact, along with estimate (2.14) of the kernel, implies that for $t<T$, the expression

$$A^{*1-\varepsilon}P(t)x = \int_t^T A^{*1-\varepsilon}S^*(\tau - t)\Phi(\tau, t)x + A^{*1-\varepsilon}S^*(T-t)\Phi(T, t)x \tag{3.15}$$

is well defined for any $x \in L_2(\Omega)$, thus establishing (ii.1). Moreover, the operator

$$D^*A^*P(t) = D^*A^{*\frac{1}{4}-\rho}A^{*\frac{3}{4}+\rho}P(t), \qquad 0<\rho<\tfrac{1}{4},$$

is bounded, by virtue of (ii.1) and $D^*A^{*\frac{1}{4}-\rho}$ being bounded (cf. (2.2) or (2.13)). Property (ii.2) is then a consequence of (3.15) via (iii) in Proposition 3.2.

[28] Throughout § 3, $[\cdot, \cdot]$ denotes the $L_2(\Omega)$-inner product and $(\cdot, \cdot)$ the $L_2(\Gamma)$-inner product.

[29] Compare with [G1, eq. (3.32)].

[30] for $\alpha = 0$ in (1.0), $P(T) = 0$.

Property (iii). From the definition of $P(t)$ in (3.10), we have

$$[P(t)x, y] = \int_t^T [\Phi(\tau, t)x, S(\tau - t)y]\, d\tau + [\Phi(T, t)x, S(T-t)y].$$

We next substitute for $S(\tau - t)y$ the expression obtained from the identity

$$y^0(\tau, t; y) \equiv \Phi(\tau, t)y = S(\tau - t)y - \int_t^\tau AS(\tau - \sigma)DD^*A^*P(\sigma)\Phi(\sigma, t)y\, d\sigma,$$

which results from combining (3.5) (ii) with (3.9). We do likewise for $S(T-t)y$. We thus obtain

$$(3.16)^{31} \quad [P(t)x, y] = \int_t^T [\Phi(\tau, t)x, \Phi(\tau, t)y]\, d\tau + [\Phi(T, t)x, \Phi(T, t)y] + I_3 + I_4,$$

where

$$I_3 \equiv \int_t^T \int_t^\tau [\Phi(\tau, t)x, AS(\tau - \sigma)DD^*A^*P(\sigma)\Phi(\sigma, t)y]\, d\sigma\, d\tau,$$

$$I_4 \equiv \int_t^T [\Phi(T, t)x, AS(T-\sigma)DD^*A^*P(\sigma)\Phi(\sigma, t)y]\, d\sigma.$$

After interchanging the order of integration, we rewrite I_3 as

$$I_3 = \int_t^T \int_\sigma^T [\Phi(\tau, t)x, AS(\tau - \sigma)DD^*A^*P(\sigma)\Phi(\sigma, t)y]\, d\tau\, d\sigma$$

$$= \int_t^T \int_\sigma^T [[\Phi(\tau, t)x, S(\tau - \sigma)\tilde{A}DD^*A^*P(\sigma)\Phi(\sigma, t)y]]\, d\tau\, d\sigma$$

where, in the last step, we have extended[32] the original $A: L_2(\Omega) \supset \mathcal{D}(A) \to L_2(\Omega)$ to $\tilde{A}$ continuous operator from $L_2(\Omega) = \mathcal{D}(\tilde{A})$ into the dual space $[\mathcal{D}(A)]'$, and likewise by continuity the $L_2(\Omega)$-inner product $[,]$ to the duality pairing $[[,]]$ on $\mathcal{D}(A) \times [\mathcal{D}(A)]'$. Therefore

$$I_3 = \int_t^T \left[\left[\int_\sigma^T S^*(\tau - \sigma)\Phi(\tau, \sigma)\Phi(\sigma, t)x\, d\tau, \tilde{A}DD^*A^*P(\sigma)\Phi(\sigma, t)y\right]\right] d\sigma.$$

Similarly for I_4:

$$I_4 = \int_t^T [[S^*(T-\sigma)\Phi(T, \sigma)\Phi(\sigma, t)x, \tilde{A}DD^*A^*P(\sigma)\Phi(\sigma, t)y]]\, d\sigma.$$

Thus, by (3.10)

$$I_3 + I_4 = \int_t^T [[P(\sigma)\Phi(\sigma, t)x, \tilde{A}DD^*A^*P(\sigma)\Phi(\sigma, t)y]]\, d\sigma$$

$$= \int_t^T (D^*A^*P(\sigma)\Phi(\sigma, t)x, D^*A^*P(\sigma)\Phi(\sigma, t)y)\, d\sigma \tag{3.17}$$

as desired, since $D^*A^*P(\sigma)$ is well defined in $L_2(\Gamma)$ for $\sigma < T$ (property (ii.2)), and hence $\tilde{A}$ can be replaced by A in the last step. Equations (3.16)–(3.17) provide the desired conclusion (3.14).

[31] Compare with [G1, eq. (3.30)].

[32] By standard isomorphism techniques; see e.g. [L5], [L9]: A(resp. A^*) defines an isomorphism from $\mathcal{D}(A)$ (resp. $\mathcal{D}(A^*)$) onto $L_2(\Omega)$; thus the dual A^{**} of A^* defines an isomorphism from $L_2(\Omega)$ onto $[\mathcal{D}(A^*)]' = [\mathcal{D}(A)]'$.

Property (iv). Setting $x = y$ in (3.14) yields $P(t)$ as a positive definite operator, and, moreover, proves property (iv) via (3.6) and (3.9).

Property (v). It follows from (iv) that $\lim_{t\uparrow T}[P(t)x, x]$ exists and is nonnegative, for each $x \in L_2(\Omega)$. Since $P(t)$ is selfadjoint by (iii.1), we have in the usual way

$$2[P(t)x, y] = [P(t)(x+y), (x+y)] - [P(t)x, x] - [P(t)y, y],$$

and so $\lim[P(t)x, y]$ exists for all x, y in $L_2(\Omega)$ and is finite as $t\uparrow T$. Next, with $\alpha = 1$ in (1.0), and for $x \in L_2(\Omega)$ and $y \in \mathscr{D}(A^{\frac{1}{4}+\varepsilon})$, we have

$$\lim_{t\uparrow T}[P(t)x, y] = \left[\lim_{t\uparrow T} A^{*-\frac{1}{4}-\varepsilon}P(t)x, A^{\frac{1}{4}+\varepsilon}y\right] = [x, y]$$

as soon as we prove that the limit, as $t\uparrow T$, of

$$A^{*-\frac{1}{4}-\varepsilon}P(t)x = \int_t^T S^*(\tau - t)A^{*-\frac{1}{4}-\varepsilon}\Phi(\tau, t)x\, d\tau + S^*(T-t)A^{*-\frac{1}{4}-\varepsilon}\Phi(T, t)x$$

is $A^{*-\frac{1}{4}-\varepsilon}x$. But this is, in fact, true, as a consequence of Proposition 3.2(v). It then routinely follows that $\lim_{t\uparrow T}[P(t)x, y] = [x, y]$ for all x and y in $L_2(\Omega)$, and hence [T1, p. 353] that $\lim_{t\uparrow T}P(t)x = x$, as desired. The principle of uniform boundedness gives the uniform bound of $|P(t)|$ on $[0, T]$. The proof of Proposition 3.3 is thus complete. □

The expected interplay between Φ and P is revealed by the following proposition.

PROPOSITION 3.4. *Given $0 \leqq s < t < T$, the following inclusion holds*:

$$\text{range}\{[I - DD^*A^*P(t)]\Phi(t, s)\} = [I - DD^*A^*P(t)]\Phi(t, s)L_2(\Omega) \subset \mathscr{D}(A). \tag{3.18}$$

Moreover, for any $x \in L_2(\Omega)$, the following differential equation is satisfied

$$\frac{\partial\Phi(t, s)x}{\partial t} = A[I - DD^*A^*P(t)]\Phi(t, s)x \tag{3.19}$$

$$(\textit{i.e.}, \dot{y}^0(t, s; x) = A[I - DD^*A^*P(t)]y^0(t, s; x)).$$

Proof. We already know that

$$\Phi(t, s)x = S(t-s)x - \int_s^t AS(t-\tau)DD^*A^*P(\tau)\Phi(\tau, s)x\, d\tau \tag{3.20}$$

(cf. (3.5b) and (3.9)). With $x \in L_2(\Omega)$ and $y \in \mathscr{D}(A^*) = \mathscr{D}(A)$,[33] we then obtain by differentiation for $0 \leqq s < t < T$

$$\begin{aligned}\frac{d}{dt}[\Phi(t, s)x, y] &= [S(t-s)x, A^*y] - [DD^*A^*P(t)\Phi(t, s)x, A^*y] \\ &\quad - \left[\int_s^t AS(t-\tau)DD^*A^*P(\tau)\Phi(\tau, s)x\, d\tau, A^*y\right] \\ &\qquad\qquad \text{(well defined by Proposition 3.3(iii.2))} \\ &= [\Phi(t, s)x, A^*y] - [DD^*A^*P(t)\Phi(t, s)x, A^*y] \qquad \text{(by 3.20)} \\ &= [(I - DD^*A^*P(t))\Phi(t, s)x, A^*y].\end{aligned} \tag{3.21}$$

Now, observe that for $x \in L_2(\Omega)$ and $s < t$, one obtains from Corollary 1.2 and (2.1) $(I - DD^*A^*P(t))\Phi(t, s)x \in H^2(\Omega)$ a fortiori, and that its restriction on the boundary

[33] See e.g. [L5, p. 196].

vanishes as

$$(I-DD^*A^*P(t))\Phi(t,s)x|_\Gamma = \Phi(t,s)x|_\Gamma - D^*A^*P(t)\Phi(t,s)x$$
$$= u^0(t,s;x) - u^0(t,s;x) \equiv 0.$$

In other words, the desired inclusion (3.18) holds. Therefore, A^* on the inner product in (3.21) can be moved to the left, thereby yielding

$$\frac{d}{dt}[\Phi(t,s)x, y] = [A(I-DD^*A^*P(t))\Phi(t,s)x, y].$$

Since $\mathscr{D}(A^*)$ is dense in $L_2(\Omega)$, the above identity can be extended to all $y \in L_2(\Omega)$, and thus (3.19) is proved. □

We can finally state and prove the property that the operator $P(t)$ defined by (3.10) is, in fact, a Riccati operator.

THEOREM 3.5. *The operator $P(t)$ defined by* (3.10) *satisfies the following Riccati equation*:

$$[\dot{P}(t)x, y] = -[x, y] - [P(t)x, Ay] - [P(t)Ax, y]$$
$$+ (D^*A^*P(t)x, D^*A^*P(t)y), \qquad 0 \leqq t < T, \quad x, y \in \mathscr{D}(A), \tag{3.22}[34]$$
$$\lim_{t \uparrow T} P(t)x = x.$$

Proof. With $x \in L_2(\Omega)$ and $y \in \mathscr{D}(A)$ and $0 \leqq s < t < T$, we compute from (3.10) via (3.13):

$$[\dot{P}(t)\Phi(t,s)x, y] = -[\Phi(t,s)x, y] + \int_t^T \left[\frac{\partial \Phi(\tau,t)}{\partial t}\Phi(t,s)x, S(\tau-t)y\right] d\tau$$
$$- \int_t^T [\Phi(\tau,t)\Phi(t,s)x, S(\tau-t)Ay\, d\tau]$$
$$+ \left[\frac{\partial \Phi(T,t)}{\partial t}\Phi(t,s)x, S(T-t)y\right]$$
$$- [\Phi(T,t)\Phi(t,s)x, S(T-t)Ay]$$
$$= -[\Phi(t,s)x, y] - [P(t)\Phi(t,s)x, Ay]$$
$$+ \int_t^T \left[\frac{\partial \Phi(\tau,t)}{\partial t}\Phi(t,s)x, S(\tau-t)y\right] d\tau$$
$$+ \left[\frac{\partial \Phi(T,t)}{\partial t}\Phi(t,s)x, S(T-t)y\right] \qquad \text{(by (3.10))}$$
$$= -[\Phi(t,s)x, y] - [P(t)\Phi(t,s)x, Ay]$$
$$- \int_t^T [\Phi(\tau,t)A(I-DD^*A^*P(t))\Phi(t,s)x, S(\tau-t)y]\, d\tau$$
$$- [\Phi(T,t)A(I-DD^*A^*P(t))\Phi(t,s)x, S(T-t)y]$$
$$\text{(by combination of (3.13) and (3.19)).}$$

[34] By Proposition 3.3(ii.1) and (iii.1), one may merely require $x, y \in \mathscr{D}(A^\varepsilon)$, for any $\varepsilon > 0$ since the second and third term on the right-hand side of (3.22) can be rewritten as $[A^{*1-\varepsilon}P(t)x, A^\varepsilon y] + [P(t)A^{1-\varepsilon}A^\varepsilon x, y]$.

Using (3.10) again we finally obtain

$$
\begin{aligned}
[\dot{P}(t)\Phi(t,s)x, y] &= -[\Phi(t,s)x, y] - [P(t)\Phi(t,s)x, Ay] \\
&\quad - [P(t)A(I - DD^*A^*P(t))\Phi(t,s)x, y],
\end{aligned} \tag{3.23}
$$

valid for $0 \leqq s < t < T$, for $x \in L_2(\Omega)$ and $y \in \mathcal{D}(A)$. Now, the right-hand side operator,

$$
P(t)A(I - DD^*A^*P(t)) = P(t)A^{1-\varepsilon}A^{\varepsilon} - P(t)A^{1-\varepsilon}A^{\varepsilon}DD^*A^*P(t)
$$

where $\varepsilon < \frac{1}{4}$, is clearly, by Proposition (ii.1) ($(P(t)A^{1-\varepsilon})^* = A^{*1-\varepsilon}P(t)$) and (ii.2), a closed operator with domain $\mathcal{D}(A^{\varepsilon})$; moreover,

$$
\begin{aligned}
&\lim_{s \uparrow t} P(t)A(I - DD^*A^*P(t))\Phi(t,s)x \\
&\qquad = P(t)A^{1-\varepsilon} \lim_{s \uparrow t} A^{\varepsilon}\Phi(t,s)x + P(t)ADD^*A^*P(t) \lim_{s \uparrow t} \Phi(t,s) \\
&\qquad = P(t)A^{1-\varepsilon}A^{\varepsilon}x + P(t)ADD^*P(t)x
\end{aligned} \tag{3.24}
$$

for any $x \in \mathcal{D}(A^{\varepsilon})$. In fact, for $\varepsilon < \frac{1}{4}$

$$
A^{\varepsilon}\Phi(t,s)x = S(t,s)A^{\varepsilon}x + \int_s^t A^{\frac{3}{4}+\rho+\varepsilon}S(t-\tau)A^{\frac{1}{4}-\rho}Du^0(t,s;x)\,d\tau \tag{3.25}
$$

converges to $A^{\varepsilon}x$ as $s \uparrow t$, as it follows from Lemma 3.1 (3.11a) (since $t < T$).

We conclude that, as $s \uparrow t$, the right-hand side of (3.23) converges to the right-hand side of (3.22) as desired. As to the left-hand side of (3.23), consider now the operator $\dot{P}(t)$ well defined at least on the subspace

$$
\mathcal{M}_t = \{y \in L_2(\Omega) : y = \Phi(t,s)x,\ 0 \leqq s < t,\ x \in L_2(\Omega)\} \subset \mathcal{D}(A^{\varepsilon}) \qquad \text{(by (3.25))}.
$$

Then, $\dot{P}(t)$ is *closable*; $x_s \in \mathcal{M}_t$, $x_s \to 0$, $\dot{P}(t)x_s \to v$ implies by the right-hand side of (3.23), i.e., by (3.24), that $v = 0$. We denote the *closure* of $\dot{P}(t)$ (smallest closed extension) still by $\dot{P}(t)$. Then for $x \in \mathcal{D}(A^{\varepsilon})$ by Proposition 3.2(v)

$$
\lim_{s \uparrow t} \dot{P}(t)\Phi(t,s)x = \dot{P}(t)x,
$$

and the left-hand side of (3.23) converges to the left-hand side of (3.22). The theorem is thus proved. □

We conclude by presenting another property of Φ in the second variable, which complements identity (3.13).

PROPOSITION 3.6. *The following existence and regularity properties of the derivative* $\partial\Phi(t,s)x/\partial s$ *hold in* t *for fixed* s. *Let* $\theta < \frac{1}{4}$ *be given such that* $r \equiv \frac{3}{4} + \varepsilon + \theta < 1$, *where* ε *is defined by* (2.1) *and henceforth fixed. Then*

$$
\frac{\partial\Phi(\cdot, s)x}{\partial s} = \begin{cases} L_1(s, T-h; H^{2\theta}(\Omega)), & x \in \mathcal{D}(A^{\gamma}),\quad \gamma > \theta, \\ C([s+h, T-h]; H^{2\theta}(\Omega)), & x \in L_2(\Omega), \end{cases} \tag{3.26}
$$

where h *is an arbitrary positive number* ($< T - s$, *or* $(T-s)/2$ *respectively*).

Proof. Henceforth s is fixed and L_1 and C denote, for brevity, the spaces at the right side of (3.26). From

$$
\Phi(t,s)x = S(t-s)x + \int_s^t A^{\frac{3}{4}+\varepsilon}S(t-\tau)A^{\frac{1}{4}-\varepsilon}DD^*A^*P(\tau)\Phi(\tau,s)x\,d\tau,
$$

with $t \leqq T-h$, one sees that we must show that $\partial\Phi(t,s)/\partial t$ is (in t for fixed s) the unique solution with the stated regularity (3.26) of the integral equation

$$\begin{aligned}\frac{\partial\Phi(t,s)x}{\partial s} &= -AS(t-s)x - A^{\frac{3}{4}+\varepsilon}S(t-s)A^{\frac{1}{4}-\varepsilon}DD^*A^*P(s)x \\ &\quad + \int_s^t A^{\frac{3}{4}+\varepsilon}S(t-\tau)A^{\frac{1}{4}-\varepsilon}DD^*A^*P(\tau)\frac{\partial\Phi(\tau,s)x}{\partial s}\,d\tau.\end{aligned} \tag{3.27}$$

This will be done via the Banach fixed point theorem on L_1 or C, respectively, following the proof in [T2, Thm. 2.1]. Therefore, details are omitted here. For $g(\cdot) \in L_1$ or C, we introduce the operator F by setting

$$\begin{aligned}(Fg)(t) &\equiv -AS(t-s)x - A^{\frac{3}{4}+\varepsilon}S(t-s)A^{\frac{1}{4}-\varepsilon}DD^*A^*P(s)x \\ &\quad + \int_s^t A^{\frac{3}{4}+\varepsilon}S(t-\tau)A^{\frac{1}{4}-\varepsilon}DD^*A^*P(\tau)g(\tau)\,d\tau.\end{aligned} \tag{3.28}$$

With $\theta < \frac{1}{4}$, the following identification applies:

$$\mathscr{D}(A^\theta) = H^{2\theta}(\Omega),$$

the identification being set theoretically and topologically with $|x|_{H^{2\theta}(\Omega)}$ and $|A^\theta x|_{L_2(\Omega)}$ equivalent norms [F2], [L6, Thm. 5.1], [L3, App. B], [L5, I]. Therefore

$$A^{1-\alpha+\theta}S(\cdot - s)A^\alpha x \in L_1 \quad \text{when } x \in \mathscr{D}(A^\gamma), \quad \gamma > \theta \qquad \text{(from (2.14))},$$

$$S(\cdot - s - h)A^{1+\theta}S(h)x \in C \quad \text{when } x \in L_2(\Omega),$$

and the first term at the right side of (3.28) is in L_1 or C, respectively. Similar considerations apply to the second term, since $r < 1$. The $H^{2\theta}(\Omega)$-norm of the integrand is integrable (cf. (3.29) below). Thus, F is well defined as an operator on L_1 or C. For the integral term, we compute for $s \leqq t \leqq T-h$:

$$|A^rS(t-\tau)A^{\frac{1}{4}-\varepsilon}DD^*A^*P(\tau)| \leqq C_{T-h}|A^rS(t-\tau)A^{\frac{1}{4}-\varepsilon}D|$$

(where we have applied the principle of uniform boundedness to property (ii.2) of Proposition 3.3)

$$\leqq \frac{C_{T-h}K}{(t-\tau)^r} \equiv \frac{K_1}{(t-\tau)^r} \qquad \text{(by (2.13) and (2.14))}. \tag{3.29}$$

F has a fixed point in C. In fact, setting $v = g_1 - g_2$ we compute from (3.28)–(3.29):

$$|(Fv)(t)|_{H^{2\theta}(\Omega)} \leqq \int_s^t \frac{K_1}{(t-\tau)^r}\,d\tau\,|v|_C \leqq K_1 t^q B(1,q)|v|_C,$$

where $|v|_C = \max\{|v(t)|_{H^{2\theta}(\Omega)},\ s+h \leqq t \leqq T-h\}$ and $B(p,q)$ is the beta function

$$B(p,q) = \int_0^1 t^{p-1}(1-t)^{q-1}\,dt.$$

In general,

$$|(F^nv)(t)|_{H^{2\theta}(\Omega)} \leqq K_1^n t^{nq} B(1,q)B(1+q,q)\cdots B(1+(n-1)q,q)|v|_C.$$

Taking sup over $[s+h, T-h]$ yields $|F^nv|_C \leqq C_n|v|_C$ where C_n is, by virtue of the known identities $B(p,q) = \Gamma(p)\Gamma(q)/\Gamma(p+q)$ and $\Gamma(x+1) = x\Gamma(x)$, rewritten as

$$C_n = K_1^n T^{nq}\frac{\Gamma(1)\Gamma^n(q)}{\Gamma(1+nq)} = [K_1T^q\Gamma(q)]^n\frac{\Gamma(1)}{nq\Gamma(nq)},$$

and thus $C_n \downarrow 0$ as $n \to \infty$. The Banach fixed point theorem applies and provides the unique solution of (3.27).

The proof is similar for L_1 (see [T2] for details). □

Appendix. Proof of (2.19). We already know that

(i) $g_t \in L_2(0, T; L_2(\Gamma))$ (see (2.12) and subsequent paragraph),

(ii) g_z is an $L_2(\Gamma)$-function, analytic in $\bar{\mathscr{F}} - \{T\}$ (see (2.17)).

Properties (i) and (ii) then readily imply

$$g_z \in L_2(\mathscr{F}; L_2(\Gamma)). \tag{A.1}$$

The following lemma proves a fortiori, via Remark 2.2(a), the desired assertion (2.19).

LEMMA A.1. *Property* (A.1) *for* g_z *implies*

$$LL^*Lg_z \in L_\infty(\mathscr{F}; L_2(\Omega)).$$

Proof. Throughout this appendix, we shall use the complex variable $\zeta = \rho e^{i\theta}$, $0 \leqq \rho \leqq r$, when integrating along the line segment from 0 to $z = re^{i\theta}$, and the complex variable λ when integrating along the line segment from z to T. Integrating by parts on the integral (2.3), as extended to $z \in \mathscr{F}$, gives

$$\begin{aligned}(L^*Lg_\zeta)(z) &= \int_z^T D^*(AS(\lambda - z))^*(Lg_\zeta)(\lambda)\, d\lambda \\ &= -D^*S^*(T-z)(Lg_\zeta)(T) + D^*(Lg_\zeta)(z) \\ &\quad + \int_z^T D^*S^*(\lambda - z)\frac{d}{d\lambda}(Lg_\zeta)(\lambda)\, d\lambda.\end{aligned} \tag{A.2}$$

Assume for the moment the following:

$$A^{*-\frac{3}{4}-\varepsilon}\frac{d}{dz}(Lg_\zeta)(z) \in L_2(\mathscr{F}; L_2(\Omega)), \tag{A.3}$$

to be established at the end. Rewrite (A.2) more conveniently as

$$\begin{aligned}(L^*Lg_\zeta)(z) &= -D^*A^{*\frac{1}{4}-\varepsilon}A^{*3\varepsilon}S^*(T-z)A^{*-\frac{1}{4}-2\varepsilon}(Lg_\zeta)(T) \\ &\quad + D^*A^{*\frac{1}{4}-\varepsilon}A^{*-\frac{1}{4}+\varepsilon}(Lg_\zeta)(z) \\ &\quad + D^*A^{*\frac{1}{4}-\varepsilon}\int_z^T A^{*\frac{1}{2}+2\varepsilon}S^*(\lambda - z)A^{*-\frac{3}{4}-\varepsilon}\frac{d}{d\lambda}(Lg_\zeta)(\lambda)\, d\lambda \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned} \tag{A.4}$$

In what follows, we shall use repeatedly, with no further mention, the fact that the operators $A^{\frac{1}{4}-\varepsilon}D$ and $D^*A^{*\frac{1}{4}-\varepsilon}$ are bounded (see (2.1)) and, moreover, we shall use equations like (2.13)–(2.14). However, instead of keeping track of the explicit constants, we shall combine them into a generic const. As to the first term, I, we get[35]

---

[35] Here and henceforth, we shall use that $\mathscr{D}(A) = \mathscr{D}(A^*)$ in our case; hence, $\mathscr{D}(A^\theta) = \mathscr{D}(A^{*\theta})$, $0 \leqq \theta \leqq 1$ by interpolation. Therefore,

$$|A^{*\theta}y|_\Omega = |y|_{\mathscr{D}(A^{*\theta})} = |y|_{\mathscr{D}(A^\theta)} = |A^\theta y|_\Omega.$$

See, e.g., [I.3] for details.

from (2.2b)

$$\text{(A.5)}^{36} \qquad \begin{aligned} |I|_\Gamma &\leqq \|D^* A^{*\frac{1}{4}-\varepsilon}\| \frac{M}{|T-z|^{3\varepsilon}} \int_0^T \frac{M}{(T-t)^{\frac{1}{3}-\varepsilon}} \|A^{\frac{1}{4}-\varepsilon} D\|\, |g_t|_\Gamma \, dt \\ &\leqq \frac{\text{const}}{|T-z|^{3\varepsilon}} \|g_z\|_\Gamma \end{aligned}$$

via the Schwarz inequality and assumption (A.1). Similarly for the second term, II, we have

$$\text{(A.6)} \qquad |\text{II}|_\Gamma \leqq \|D^* A^{*\frac{1}{4}-\varepsilon}\| \int_0^r \frac{M\|A^{\frac{1}{4}-\varepsilon} D\|}{(r-\rho)^{\frac{1}{3}+2\varepsilon}} |g_\zeta|_\Gamma \, d\rho \leqq \text{const} \int_0^r \frac{|g_\zeta|_\Gamma^2}{(r-\rho)^{6\varepsilon}} \, d\rho.$$

As to the third term, III, by means of claim (A.3) we similarly obtain

$$\text{(A.7)} \qquad |\text{III}|_\Gamma \leqq \int_{z \to T} \frac{\text{const}}{|\lambda - z|^{6\varepsilon}} |A^{*-\frac{3}{4}-\varepsilon} \frac{d}{d\lambda} (Lg_\zeta)(\lambda)|_\Omega^2 |d\lambda|,$$

where $z \to T$ means "along the line segment from z to T". Therefore, by (2.15), we obtain

$$\begin{aligned} |(LL^* Lg_\zeta)(z)|_\Omega &= \left| \int_0^z A^{\frac{3}{4}+\varepsilon} S(z-\zeta) A^{\frac{1}{4}-\varepsilon} D(L^* Lg_\cdot)(\zeta) \, d\zeta \right| \\ \text{(A.8)} \qquad &\leqq \int_0^r \frac{\text{const}}{(r-\rho)^{\frac{3}{4}+\varepsilon}} |(L^* Lg)(\rho e^{i\theta})|_\Gamma \, d\rho \\ &\leqq \text{I}' + \text{II}' + \text{III}', \end{aligned}$$

where by (A.4)–(A.7), the scalar quantities I′, II′, III′, are:

$$\text{(A.9)} \qquad \text{I}' = \int_0^r \frac{\text{const}}{(r-\rho)^{\frac{3}{4}+\varepsilon}} \frac{\|g_\zeta\|_\Gamma}{|T-\zeta|^{3\varepsilon}} \, d\rho \leqq \int_0^r \frac{\text{const}}{(r-\rho)^{\frac{3}{4}+4\varepsilon}} \|g_\zeta\|_\Gamma \, d\rho \leqq \text{const}\, \|g_z\|_\Gamma,$$

since $|T-\zeta| \geqq (r-\rho)$ (see choice of $\mathscr{F}$);

$$\text{(A.10)} \qquad \text{II}' = \int_0^r \frac{\text{const}}{(r-\rho)^{\frac{3}{4}+\varepsilon}} \int_0^\rho \frac{|g_\cdot|_\Gamma^2}{(\rho-\sigma)^{6\varepsilon}} \, d\sigma \, d\rho \leqq \text{const}\, \|g_z\|_\Gamma^2,$$

as follows by changing the order of integration, where $\cdot$ in (A.10) stands for $\sigma e^{i\theta}$;

$$\text{III}' = \int_0^r \frac{\text{const}}{(r-\rho)^{\frac{3}{4}+\varepsilon}} \int_{\zeta \to T} \frac{1}{|\lambda - \zeta|^{6\varepsilon}} \left| A^{*-\frac{3}{4}-\varepsilon} \frac{d}{d\lambda} (Lg_\cdot)(\lambda) \right|_\Omega^2 |d\lambda| \, d\rho$$

and, after setting $\zeta = T + r' e^{i\beta}$, $\lambda = T - \sigma e^{i\beta}$, $-r' \leqq \sigma \leqq 0$ along the line segment from ζ to T,

$$\begin{aligned} \text{III}' &\leqq \int_0^r \frac{\text{const}}{(r-\rho)^{\frac{3}{4}-\varepsilon}} \int_{-r}^0 \left| \frac{A^{*-\frac{3}{4}-\varepsilon}}{(r'+\sigma)^{6\varepsilon}} \frac{d}{d\lambda} (Lg_\cdot)(\lambda) \right|_\Omega^2 d\sigma \, d\rho \\ \text{(A.11)} \qquad &\leqq \text{const} \int_{\mathscr{F}} \left| A^{*-\frac{3}{4}-\varepsilon} \frac{d}{d\lambda} (Lg_\cdot)(\lambda) \right|_\Omega^2 d\lambda, \end{aligned}$$

which follows by changing the order of integration and using (A.3). The conclusion of Lemma A.1 then follows from (A.8) via (A.9)–(A.11).

[36] $\|g_z\|_\Gamma$ is the $L_2(\mathscr{F}; L_2(\Gamma))$-norm of g_z.

Finally, to establish claim (A.3) observe that from, say, (2.15), we obtain

$$A^{-\frac{3}{4}-\varepsilon}\frac{d}{dz}(Lg_\zeta)(z) = A^{\frac{1}{4}-\varepsilon}Dg_z - \int_0^z AS(z-\zeta)A^{\frac{1}{4}-\varepsilon}Dg_\zeta\, d\zeta.$$

We then use (A.1) and footnote 35 for the first term. As to the integral term, we use direct computation based, e.g., on the Laplace transform technique on generic rays (as in [L3, Appendix A]) combined with a uniform bound. In this way, we obtain claim (A.3). □

REFERENCES

[A1] A. K. AZIZ, ed., *The Mathematical Foundation of the Finite Element Method with Applications to Partial Differential Equations*, Academic Press, New York, 1972.

[B1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[B2] ———, *Boundary control of parabolic equations: L–Q–R theory*, Proc. V Intl. Summer Schl., Central Inst. Math. Mech., Acad. Sci. GDR, Berlin, 1977.

[C1] R. CURTAIN AND A. PRITCHARD, *The infinite dimensional Riccati equation for systems defined by evolution operators*, this Journal, 14 (1976), pp. 951–983. (Results also in [C3].)

[C2] ———, *An abstract theory for unbounded control action for distributed parameter systems*, this Journal, 15 (1977), pp. 566–611. (Results also in [C3].)

[C3] ———, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control 8, Springer-Verlag, New York, 1978.

[F1] A. FRIEDMAN, *Partial Differential Equations*, reprinted by Robert E. Krieger Publ., Huntington, NY, 1976.

[F2] D. FUJIWARA, *Concrete characterizations of the domains of fractional powers of some elliptic differential operators of the second order*, Proc. Japan Acad., 43 (1967), pp. 82–86.

[G1] J. S. GIBSON, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.

[K1] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.

[K2] V. A. KONDRATIEV, *Boundary problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967).

[L1] I. LASIECKA, *Boundary control of parabolic systems*, Appl. Math. Optim. 4 (1978), pp. 301–327.

[L2] ———, *Boundary control of parabolic systems: Finite-element approximation*, Appl. Math. Optim., 6 (1980), pp. 31–62.

[L3] ———, *Unified theory for abstract parabolic boundary problems: A semigroup approach*, Appl. Math. Optim. 6 (1980), pp. 287–333.

[L4] J.-L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[L5] J.-L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, I, II, Lecture Notes in Mathematics 181, 182, Springer, Berlin, 1972.

[L6] ———, *Problème aux limites non homogènes*, IV, Ann. Scuola Norm. Sup. Pisa, 15 (1961), pp. 311–325.

[L7] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley, New York, 1969.

[L8] D. LUKES AND D. L. RUSSELL, *The quadratic criterion for distributed systems*, this Journal, 7 (1969), pp. 101–121.

[L9] I. LASIECKA AND R. TRIGGIANI, *Feedback semigroups and cosine operators for boundary feedback parabolic and hyperbolic equations*, J. Differential Equations, to appear.

[N1] J. NECAS, *Les méthodes directes en théorie des équations élliptiques*, Masson et Cie, Paris, 1967.

[R1] H. L. ROYDEN, *Real Analysis*, Macmillan, New York, 2nd Ed., 1968.

[S1] T. SEIDMAN, *Regularity of optimal boundary control for parabolic equation, I: Semigroup methods and analyticity*, this Journal, 20 (1982), pp. xxx–xxx, to appear.

[T1] A. E. TAYLOR AND D. C. LAY, *Introduction to Functional Analysis*, John Wiley, New York, Second Ed., 1980.

[T2] R. TRIGGIANI, *Well posedness and regularity of boundary feedback parabolic systems*, J. Differential Equations, 36 (1980), pp. 347–362.

[V1] R. B. VINTER AND T. JOHNSON, *Optimal control of nonsymmetric hyperbolic systems in n-variables on the half-space*, this Journal, 15 (1977), pp. 129–143.

[W1] D. C. WASHBURN, *A semigroup approach to time optimal boundary control of diffusion processes*, Ph.D. dissertation, Univ. of California, Los Angeles, 1974.

[W2] ———, *A bound on the boundary input map for parabolic equations with application to time optimal control*, this Journal, 17 (1979), pp. 652–571.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 1, January 1983

0363-0129/83/2101-0004 $01.25/0

CONTROL OF WAVE PROCESSES WITH DISTRIBUTED CONTROLS SUPPORTED ON A SUBREGION*

JOHN LAGNESE†

Abstract. It is proved that solutions of one-dimensional wave equations satisfying general boundary conditions at the ends of a bounded interval I can be exactly controlled to any finite energy state by means of distributed controls which vanish outside of any fixed nonempty subinterval of I. An example is given which shows that no such general analogous result can hold in higher dimensions. In this case, for a spherical region, those states are characterized which can be exactly controlled to zero by means of controls supported in an annulus within the region. It is found that very strong controllability obtains when the controls are distributed near the boundary, but that only rather weak controllability is possible with controls supported in an interior annulus. Applications of these results to boundary control problems in annular regions are also discussed.

Key words. wave equation, distributed controls, boundary controls

1. Introduction. In this paper we will consider the question of exact controllability of solutions of wave equations by means of distributed control forces constrained to vanish outside of a given subset of the region in which the process evolves. Thus, for example, we will consider the problem of controlling the motion of a vibrating string or a vibrating membrane by means of external forces exerted only on a given portion of the elastic medium. It will be shown that a vibrating string can be controlled to any finite energy configuration whatever by means of such forces, but that the same is not true for a vibrating membrane. In fact, given any open set $\mathcal{O}$ whose complement in the membrane admits "trapped rays", in general there will be finite energy combinations of frequencies and amplitudes which cannot be induced in or eliminated from the membrane by means of external forces exerted on $\mathcal{O}$ alone. On the other hand, there are results in the opposite direction. For a circular membrane, we will characterize those states which may be reached by means of control forces distributed in an annulus within and concentric to the membrane. As is to be expected, only rather weak controllability obtains for an annulus in the interior of the membrane due to the presence of reflected waves which can be trapped in the uncontrolled portion of the membrane. However, very strong controllability results are obtained if the controls are distributed near the boundary of the membrane.

Controllability results of the type just discussed can be used to obtain *boundary controllability* results. For example, consider a vibrating membrane in the shape of an annulus with the outer boundary clamped. Suppose no external forces are present and that motion is to be controlled by means of control forces acting on the inner boundary. This problem, originally treated in [9], may be approached by solving the control problem for the clamped circular membrane using distributed controls supported in a small neighborhood of the center. The solution to this control problem, when restricted to the original annulus, will provide a solution to the original boundary control problem. Similarly, by controlling a clamped circular membrane by means of distributed controls supported near the boundary, one obtains boundary controllability results for a slightly smaller circular membrane. In this case our results are in agreement with those of Graham and Russell [6]. This approach to boundary controllability will be carried out in detail in § 4.

* Received by the editors July 24, 1981.

† Department of Mathematics, Georgetown University, Washington, DC 20057, and Center for Applied Mathematics, National Bureau of Standards, Washington, DC 20234.

Our results for one-dimensional problems will be stated and proved in the next section. The higher dimensional analogs are treated in § 3.

2. One-dimensional problems. We consider the problem

$$\rho(x)\frac{\partial^2 u}{\partial t^2}-\frac{\partial}{\partial x}\left(p(x)\frac{\partial u}{\partial x}\right)=g(x,t), \qquad 0\leqq x\leqq l, \quad t>0, \tag{2.1}$$

$$u(x,0)=u_0(x), \qquad \frac{\partial u}{\partial t}(x,0)=v_0(x), \qquad 0\leqq x\leqq l, \tag{2.2}$$

$$\alpha_0 u(0,t)+\beta_0\frac{\partial u}{\partial x}(0,t)=\alpha_1 u(l,t)+\beta_1\frac{\partial u}{\partial x}(l,t)=0, \qquad t>0. \tag{2.3}$$

It is assumed that $|\alpha_0|+|\beta_0|>0$, $|\alpha_1|+|\beta_1|>0$, $\alpha_0\beta_0\leqq 0$, $\alpha_1\beta_1\geqq 0$, and that $\rho(x)$ and $p(x)$ are strictly positive and three times continuously differentiable on $[0, l]$. (2.1) then describes the forced motion of an elastic string with local stiffness $p(x)$ and local density $\rho(x)$.

By a standard transformation of the independent and dependent variables, (2.1) may be brought to the form

$$\frac{\partial^2 u}{\partial t^2}-\frac{\partial^2 u}{\partial x^2}-q(x)u=h(x,t), \qquad 0\leqq x\leqq 1, \quad t>0, \tag{2.4}$$

where h differs from g by a factor depending on x only, and $q(x)$ is continuously differentiable on $[0, 1]$. The boundary conditions remain of the same form under this transformation. Thus we consider the problem (2.4) and (2.2), (2.3) with $l=1$. Given $(a, b)\subset(0, 1)$, the object is to select $h\in L^2((0,1)\times(0,T))$ such that the x-support of h is contained in $[a, b]$ for each t and

$$u(x,T)=u_T(x), \qquad \frac{\partial u}{\partial t}(x,T)=v_T(x), \qquad 0\leqq x\leqq 1, \tag{2.5}$$

for some $T>0$ independent of (u_0, v_0) and (u_T, v_T), where each pair is a fixed but arbitrary "finite energy" state, e.g.,

$$\int_0^1 [(u_0')^2+v_0^2]\,dx<+\infty.$$

This is equivalent to (2.10) below. Because of the time reversibility of (2.3), (2.4), there is no loss of generality if we assume $u_T=v_T=0$.

Following Russell (cf. [11]), (2.2)–(2.4) is transformed to a moment problem for h. To do this one introduces the eigenvalues $\{\lambda_k\}_1^\infty$ and corresponding normalized eigenfunctions $\{\phi_k\}_1^\infty$ of the regular Sturm–Liouville problem

$$\phi''(x)+q(x)\phi(x)+\lambda\phi(x)=0, \qquad 0\leqq x\leqq 1, \tag{2.6}$$

$$\alpha_0\phi(0)+\beta_0\phi'(0)=\alpha_1\phi(1)+\beta_1\phi'(1)=0. \tag{2.7}$$

The λ_k are strictly increasing and in the present case $\lambda_1\geqq 0$. To simplify the presentation we assume $\alpha_0^2+\alpha_1^2>0$, so that $\lambda_1>0$, but this is inessential.

Expand u_0, v_0 in Fourier series in $\{\phi_k\}$:

$$u_0=\sum_{k=1}^{\infty}\mu_k\phi_k, \qquad v_0=\sum_{k=1}^{\infty}\nu_k\phi_k,$$

$$\mu_k=\int_0^1 u_0(x)\phi_k(x)\,dx, \qquad \nu_k=\int_0^1 v_0(x)\phi_k(x)\,dx.$$

Then the problem (2.2)–(2.5) ($u_T = v_T = 0$) for a control h supported in $[a, b]$ for each t is equivalent to the moment problem

$$\int_0^T \int_a^b h(x,t)\phi_k(x) \sin \omega_k t \, dx \, dt = \omega_k \mu_k, \tag{2.8}$$

$$\int_0^T \int_a^b h(x,t)\phi_k(x) \cos \omega_k t \, dx \, dt = -\nu_k, \qquad k = 1, 2, \cdots, \tag{2.9}$$

where $\omega_k = \sqrt{\lambda_k}$.

THEOREM 2.1. *Assume that* u_0, v_0 *satisfy*

$$\sum_1^\infty (\omega_k^2 \mu_k^2 + \nu_k^2) < +\infty. \tag{2.10}$$

If $T > 2$, *for any* $(a, b) \subset (0, 1)$ *the problem* (2.8), (2.9) *has a solution* h *which satisfies*

$$\int_0^T \int_a^b h^2 \, dx \, dt \leqq C \sum_1^\infty (\omega_k^2 \mu_k^2 + \nu_k^2). \tag{2.11}$$

Remark 2.1. The constant C in (2.11) depends on both T and the interval (a, b) and becomes unbounded as $b - a \to 0$. The lower bound of 2 for T is optimal for an arbitrary interval of support (a, b), but is not sharp for specific intervals of support. For example, if $(a, b) = (0, 1)$ the control problem is solvable for every $T > 0$. It would be of interest to determine the relation between the interval (a, b) and the optimal time.

Remark 2.2. Some authors [7], [11] have solved (2.2)–(2.5) with $g(x, t) = g(x)h(t)$ in which $g \in L^2(0, 1)$ is given and h is the control parameter. In these works it is essential that the Fourier coefficients $\{g_k\}_1^\infty$ of $g(x)$ with respect to $\{\phi_k\}$ not converge to zero too rapidly as $k \to \infty$. Thus in [11], for example, it is assumed that

$$\liminf_{k \to \infty} k g_k > 0.$$

One can see that even in the simplest cases (e.g., $q(x) \equiv 0$, $\beta_0 = \beta_1 = 0$) this condition cannot hold if $g \in C^1$ is required to be supported in $(a, b) \subset (0, 1)$.

Proof of Theorem 2.1. Because of the asymptotic properties of the ω_k, it is known (see, e.g., [11]) that if $T > 2$ there is a sequence $\{\sigma_k(t), \tilde{\sigma}_k(t)\}$ in $L^2(0, T)$ biorthogonal to $\{\sin \omega_k t, \cos \omega_k t\}$ such that

$$\sup_k \int_0^T (\sigma_k^2(t) + \tilde{\sigma}_k^2(t)) \, dt = M < +\infty,$$

where M depends on T. Thus

$$\int_0^T \sigma_j(t) \sin \omega_k t \, dt = \int_0^T \tilde{\sigma}_j(t) \cos \omega_k t \, dt = \delta_{jk},$$

$$\int_0^T \sigma_j(t) \cos \omega_k t \, dt = \int_0^T \tilde{\sigma}_j(t) \sin \omega_k t \, dt = 0$$

for $j, k = 1, 2, \cdots$. Define

$$g(x, t) = \sum_j A_j^2 (\omega_j \mu_j \sigma_j(t) - \nu_j \tilde{\sigma}_j(t)) \phi_j(x), \tag{2.12}$$

where

$$A_j^2 = 1 \Big/ \int_a^b \phi_j^2(x) \, dx,$$

and set $h(x,t)=\chi_I(x)g(x,t)$, where χ_I is the characteristic function of the interval $I=(a,b)$. Then h is a formal solution to the moment problem (2.8), (2.9), and will be a genuine solution if the right side of (2.12) is in $L^2((0,1)\times(0,T))$. This will be the case if

$$\inf_j \int_a^b \phi_j^2(x)\,dx > 0, \tag{2.13}$$

in which case

$$\int_0^T \int_a^b |h(x,t)|^2\,dx\,dt \leqq (\text{const.}) \sum_j (\omega_j^2\mu_j^2+\nu_j^2).$$

However, (2.13) is a consequence of

LEMMA 2.1. *For any interval* $(a,b)\subset(0,1)$,

$$\lim_{k\to\infty} \int_a^b \phi_k^2(x)\,dx = b-a.$$

Proof. ϕ_k satisfies

$$\phi_k''+q\phi_k+\omega_k^2\phi_k=0, \qquad 0\leqq x\leqq 1, \tag{2.14}$$

the boundary conditions (2.7), and

$$\int_0^1 \phi_k^2(x)\,dx = 1, \qquad k=1,2,\cdots.$$

Multiplying (2.14) by ϕ_k' and integrating from 0 to x gives

$$\frac{1}{\omega_k^2}|\phi_k'(x)|^2+|\phi_k(x)|^2=\frac{1}{\omega_k^2}|\phi_k'(0)|^2+|\phi_k(0)|^2-\frac{2}{\omega_k^2}\int_0^x q\phi_k\phi_k'\,d\xi. \tag{2.15}$$

Integrate this last expression from a to b to obtain

$$\begin{aligned}&\frac{1}{\omega_k^2}\int_a^b |\phi_k'|^2\,d\xi+\int_a^b \phi_k^2\,d\xi\\ &\quad=(b-a)\left[\frac{1}{\omega_k^2}|\phi_k'(0)|^2+|\phi_k(0)|^2\right]-\frac{2}{\omega_k^2}\int_a^b\int_0^\xi q(\eta)\phi_k(\eta)\phi_k'(\eta)\,d\eta\,d\xi.\end{aligned} \tag{2.16}$$

Now multiply (2.14) by ϕ_k and integrate by parts from a to b:

$$\begin{aligned}&\frac{1}{\omega_k^2}[\phi_k(b)\phi_k'(b)-\phi_k(a)\phi_k'(a)]-\frac{1}{\omega_k^2}\int_a^b |\phi_k'|^2\,d\xi+\int_a^b \phi_k^2\,d\xi\\ &\qquad=-\frac{1}{\omega_k^2}\int_a^b q\phi_k^2\,d\xi.\end{aligned} \tag{2.17}$$

Add (2.16) and (2.17) to obtain

$$\begin{aligned}&\frac{1}{\omega_k^2}[\phi_k(b)\phi_k'(b)-\phi_k(a)\phi_k'(a)]+2\int_a^b \phi_k^2\,d\xi\\ &\quad=(b-a)\left[\frac{1}{\omega_k^2}|\phi_k'(0)|^2+|\phi_k(0)|^2+\frac{q(0)}{\omega_k^2}|\phi_k(0)|^2\right]\\ &\qquad-\frac{2}{\omega_k^2}\int_a^b q\phi_k^2\,d\xi+\frac{1}{\omega_k^2}\int_a^b\int_0^\xi q'(\eta)\phi_k^2(\eta)\,d\eta\,d\xi.\end{aligned} \tag{2.18}$$

The last two terms on the right clearly go to zero as $k \to +\infty$. Also, using (2.18) with $a = 0$ and $b = 1$ gives

$$\begin{aligned} &\frac{1}{\omega_k^2}|\phi_k'(0)|^2 + |\phi_k(0)|^2 + \frac{q(0)}{\omega_k^2}|\phi_k(0)|^2 \\ &\quad = 2 + \frac{1}{\omega_k^2}[\phi_k(1)\phi_k'(1) - \phi_k(0)\phi_k'(0)] + o(1) \to 2 \quad \text{as } k \to \infty, \end{aligned} \tag{2.19}$$

in view of the boundary conditions (2.7). In addition, from (2.15) it follows that for every $\varepsilon > 0$, and $x \in [0, 1]$,

$$\begin{aligned} \frac{1}{\omega_k^2}|\phi_k(x)\phi_k'(x)| &\leqq \frac{\varepsilon}{\omega_k^2}|\phi_k'(x)|^2 + \frac{C_\varepsilon}{\omega_k^2}|\phi_k(x)|^2 \\ &\leqq \varepsilon\left[\frac{1}{\omega_k^2}|\phi_k'(0)|^2 + |\phi_k(0)|^2 + \frac{q(0)}{\omega_k^2}|\phi_k(0)|^2\right] + o(1), \qquad k \to \infty, \end{aligned} \tag{2.20}$$

since $\{\phi_k(x)\}$ is uniformly bounded on $[0, 1]$ [2, p. 335]. Lemma 2.1 now follows from (2.18)–(2.20).

3. Higher dimensional problems. Let Ω be a bounded, open, connected set in R^n with a piecewise smooth boundary and let $\mathcal{O}$ be a nonempty open set contained in Ω. We consider the problem

$$\frac{\partial^2 u}{\partial t^2} - \Delta_n u = g \quad \text{in } \Omega \times (0, T), \tag{3.1}$$

$$u(x, 0) = u_0(x), \qquad \frac{\partial u}{\partial t}(x, 0) = v_0(x) \quad \text{in } \Omega, \tag{3.2}$$

$$\alpha u + \beta \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega \times (0, T), \tag{3.3}$$

where α and β are constants with $|\alpha| + |\beta| > 0$, $\alpha\beta \geqq 0$, and ν is the unit normal pointing out of Ω.

Let (u_0, v_0) and (u_T, v_T) be given in, say, $H^1(\Omega) \times L^2(\Omega)$, and let $T > 0$. We seek a function $g \in L^2(\Omega \times (0, T))$ with $x - \operatorname{supp} g(x, t) \subset \bar{\mathcal{O}}$ for each t such that the solution of (3.1)–(3.3) satisfies

$$u(x, T) = u_T(x), \qquad \frac{\partial u}{\partial t}(x, T) = v_T(x) \quad \text{in } \Omega. \tag{3.4}$$

If $\mathcal{O} = \Omega$, it is not difficult to show that (3.1)–(3.4) has a solution g for every $T > 0$. But if there are closed rays in Ω, reflecting at the boundary as usual, which do not meet $\bar{\mathcal{O}}$ in a nonzero angle, one might suspect that in general there will be no T for which (3.1)–(3.4) has a solution g supported in $\mathcal{O}$ for arbitrary initial and final data. We will verify this below in the case where Ω is the parallelepipedon

$$\Omega = \{x \in R^n | 0 < x_i < a_i, i = 1, \cdots, n\}. \tag{3.5}$$

THEOREM 3.1. *Let Ω be given by* (3.5). *Suppose $\alpha = 0$ in* (3.3) *and $u_T = v_T = 0$. Let $\mathcal{O}$ be an open set in Ω such that $\bar{\mathcal{O}}$ does not meet every face of $\partial\Omega$. Then there are functions u_0, v_0 infinitely differentiable in $\bar{\Omega}$ with $\partial u_0/\partial\nu = \partial v_0/\partial\nu = 0$ on $\partial\Omega$ such that* (3.1)–(3.4) *has no solution $g \in L^2(\Omega \times (0, T))$ with $\cup_{0 \leqq t \leqq T} \operatorname{supp} g(x, t) \subset \bar{\mathcal{O}}$ for any $T > 0$.*

Proof. This is a consequence of a result of H. Fattorini, showing the impossibility of controlling all solutions of the Neumann problem for the homogeneous wave equation in a parallelepipedon from only one face of the boundary. In fact, let u_0, v_0 be C^∞ in $\bar{\Omega}$ with $\partial u_0/\partial \nu = \partial v_0/\partial \nu = 0$ on $\partial\Omega$, and suppose that for some $T>0$ there is a $g \in L^2(\Omega \times (0, T))$ with $g(x, t)$ supported in $\bar{\mathcal{O}}$ for each t such that (3.1)–(3.4) hold. Because of the particular geometry of Ω, we know that $u \in H^{2,2}(\Omega \times (0, T))$ (see [10, p. 6] for the definition of $H^{r,s}(Q)$). Suppose that $\mathcal{O}$ does not meet the face $x_1 = a_1$, say. Then there is a number $\tilde{a}_1$, $0 < \tilde{a}_1 < a_1$, such that $g(x, t) = 0$ almost everywhere for $x_1 \geqq \tilde{a}_1$, $x \in \Omega$. Let $\mathcal{R} = \{x \in \Omega | x_1 > \tilde{a}_1\}$ and w be the restriction of u to $\mathcal{R} \times (0, T)$. Then w satisfies the homogeneous wave equation in $\mathcal{R} \times (0, T)$, has zero data in $\mathcal{R}$ at time T, and satisfies the homogeneous Neumann condition on all faces of $\partial\mathcal{R}$ except Γ_1: $x_1 = \tilde{a}_1$. Let $f \in L^2(\Gamma_1 \times (0, T))$ be the restriction of $\partial u/\partial \nu$ to $\Gamma_1 \times (0, T)$, and $(\tilde{u}_0, \tilde{v}_0)$ be the restriction of (u_0, v_0) to $\mathcal{R}$. f may be viewed as a boundary control function on Γ_1 which steers $(\tilde{u}_0, \tilde{v}_0)$ to $(0, 0)$ in time T. Fattorini [4, Thm. 4.1] has proved that there exists C^∞ (in fact, analytic) data in $\mathcal{R}$ satisfying the homogeneous Neumann conditions on $\partial\mathcal{R}$ for which there is no control $f \in L^2(\Gamma_1 \times (0, T))$ steering this data to zero in time T, for any $T > 0$. Such data have expansions in the eigenfunctions of Δ_n in $\mathcal{R}$ which satisfy the homogeneous Neumann conditions on $\partial\mathcal{R}$, that is, in the functions

$$\phi_k(x) = C_k \cos \frac{\pi k_1 (x_1 - \tilde{a}_1)}{a_1 - \tilde{a}_1} \cos \frac{\pi k_2 x_2}{a_2} \cdots \cos \frac{\pi k_n x_n}{a_n}$$

where $k = (k_1, \cdots, k_n)$ is an arbitrary n-tuple of nonnegative integers and C_k is a constant. If we select $\tilde{a}_1$ in the form $(m/(m+1))a_1$ for some positive integer m, then such data, extended to Ω by periodicity, constitute C^∞ data in $\bar{\Omega}$ satisfying the homogeneous Neumann condition on $\partial\Omega$ which cannot be steered to zero in any time T by means of a control $g \in L^2(\Omega \times (0, T))$ supported in $\bar{\mathcal{O}}$ for each t.

It is possible to obtain a positive result for control in a parallelepipedon along the lines of [4, Thm. 4.2]: Data (u_0, v_0) whose eigenfunction expansions converge sufficiently rapidly may be steered to zero in some time T independent of (u_0, v_0) by means of a control supported in $\mathcal{O}$ for each t, where $\mathcal{O}$ is any nonempty open subset of Ω. However, we are instead going to establish such a result for a spherical region Ω with $\mathcal{O}$ an annulus inside and concentric with Ω, because of some applications to boundary control problems which we wish to discuss in the next section. Thus we consider (3.1)–(3.4) with

$$\Omega = \{x \in R^n \,|\, |x| < 1\}.$$

The desired terminal conditions are

$$u(x, T) = \frac{\partial u}{\partial t}(x, T) = 0, \qquad |x| < 1. \tag{3.6}$$

Following Graham and Russell [6], who studied the boundary control problem for a spherical region, the problem (3.1)–(3.4) is transformed to a sequence of moment problems. This is done by introducing the eigenvalues and normalized eigenfunctions of

$$\Delta_n U + \lambda U = 0, \qquad |x| < 1, \tag{3.7}$$

$$\alpha U + \beta \frac{\partial U}{\partial r} = 0, \qquad |x| = 1. \tag{3.8}$$

We are going to assume $\alpha \neq 0$ in order to eliminate the eigenvalue $\lambda = 0$, but our arguments can easily be modified to handle the opposite case.

It is well known that the eigenvalues of (3.7), (3.8) are a doubly indexed sequence $\{\lambda_{kl} | k = 0, 1, \cdots; l = 1, 2, \cdots\}$ of positive numbers which, for each k, are defined by $\lambda_{kl} = \omega_{kl}^2$, $\omega_{kl} > 0$,

$$\left(\alpha - \beta \frac{p}{2}\right) J_{k+p/2}(\omega_{kl}) + \beta \omega_{kl} J'_{k+p/2}(\omega_{kl}) = 0,$$

$l = 1, 2, \cdots$, where $p = n - 2$ and J_ν is the Bessel function of first kind of order ν. One has $\omega_{k1} < \omega_{k2} < \cdots \to +\infty$. In addition, $\omega_{kl} \to +\infty$ as $k \to \infty$, $l = 1, 2, \cdots$.

The corresponding normalized eigenfunctions are a triply indexed sequence given by

$$U_{kml}(x) = U_{kml}(r, \theta, \phi) = R_{kl}(r) Y_{km}(\theta, \phi),$$

$$k = 0, 1, \cdots, \quad m = 1, \cdots, m(k, p), \quad l = 1, 2, \cdots.$$

In these relations, r, θ, ϕ are hyperspherical coordinates [3, p. 233] with $0 \leqq r \leqq 1$ the radial coordinate, $\theta = \{\theta_j | 0 \leqq \theta_j \leqq \pi, j = 1, \cdots, p\}$ the coordinates of longitude, and $0 \leqq \phi \leqq 2\pi$ the coordinate of latitude. The number $m(k, p)$ is given by

$$m(k, p) = (2k + p) \frac{(k + p - 1)!}{p!\, k!}$$

and $\{Y_{km}(\theta, \phi) | m = 1, 2, \cdots, m(k, p)\}$ is an orthonormal basis in $L^2(\partial\Omega)$ for the surface spherical harmonics of degree k. The functions $R_{kl}(r)$ are defined as follows: If $\beta = 0$,

$$R_{kl}(r) = \sqrt{2}\, r^{-p/2} \frac{J_{k+p/2}(\omega_{kl} r)}{|J'_{k+p/2}(\omega_{kl})|},$$

and if $\beta \neq 0$,

$$R_{kl}(r) = \frac{\sqrt{2}\, \omega_{kl} r^{-p/2}}{(\omega_{kl}^2 - (k + \alpha/\beta)(k + p - \alpha/\beta))^{1/2}} \frac{J_{k+p/2}(\omega_{kl} r)}{|J_{k+p/2}(\omega_{kl})|}.$$

For each k, these functions are orthonormal in the sense that

$$\int_0^1 R_{kl}(r) R_{kj}(r) r^{n-1}\, dr = \delta_{jl}.$$

Furthermore, they have other properties listed in the following lemma. These will be crucial to the interpretation of the controllability results given below and may be of interest in themselves. (3.9) is analogous to Lemma 2.1, but concerns the normalized eigenfunctions of a singular, rather than regular, Sturm–Liouville problem.

LEMMA 3.1. *For any* $a \in [0, 1)$,

$$\lim_{l \to \infty} \int_0^a R_{kl}^2(r) r^{n-1}\, dr = a, \qquad k = 0, 1, \cdots, \tag{3.9}$$

$$\lim_{k \to \infty} \int_0^a R_{kl}^2(r) r^{n-1}\, dr = 0, \qquad l = 1, 2, \cdots, \tag{3.10}$$

$$\liminf_{\substack{k \to \infty \\ l \to \infty}} \int_a^1 R_{kl}^2(r) r^{n-1}\, dr > 0. \tag{3.11}$$

The proof of this lemma is given in the Appendix.

Now let u be a sufficiently smooth solution to (3.1)–(3.3), (3.6) for some $g \in L^2(\Omega \times (0, T))$ supported in a region

$$\mathcal{O} = \{x \in \Omega | r_0 < |x| < r_1\}$$

for each t, where $0 \leqq r_0 < r_1 \leqq 1$ are fixed. Also, let v be a sufficiently smooth solution of (3.1) with $g = 0$, and of (3.3). Then

$$\int_0^T \int_{\mathcal{O}} vg \, dx \, dt = \int_0^T \int_{\Omega} [v(u_{tt} - \Delta_n u) - u(v_{tt} - \Delta_n v)] \, dx \, dt$$

$$= \int_{|x|<1} [u_0(x) v_t(x, 0) - v_0(x) v(x, 0)] \, dx.$$

To obtain a moment problem for g, one substitutes for $v(x, t)$ the separated solutions

$$v(x, t) = \begin{cases} U_{kml}(x) \sin \omega_{kl} t, \\ U_{kml}(x) \cos \omega_{kl} t, \end{cases}$$

and for u_0, v_0 their expansions in U_{kml}:

$$u_0(x) = \sum_{k=0}^{\infty} \sum_{m=1}^{m(k,p)} \sum_{l=1}^{\infty} \mu_{kml} U_{kml}(x), \tag{3.12}$$

$$v_0(x) = \sum_{k=0}^{\infty} \sum_{m=1}^{m(k,p)} \sum_{l=1}^{\infty} \nu_{kml} U_{kml}(x),$$

$$\mu_{kml} = \int_{|x|<1} u_0(x) U_{kml}(x) \, dx, \qquad \nu_{kml} = \int_{|x|<1} v_0(x) U_{kml}(x) \, dx. \tag{3.13}$$

One then obtains the following moment problem for g:

$$\int_0^T \int_{\mathcal{O}} g(x, t) U_{kml}(x) \sin \omega_{kl} t \, dx \, dt = \omega_{kl} \mu_{kml}, \tag{3.14}$$

$$\int_0^T \int_{\mathcal{O}} g(x, t) U_{kml}(x) \cos \omega_{kl} t \, dx \, dt = -\nu_{kml}, \tag{3.15}$$

with k, m, l varying as above.

THEOREM 3.2. *Let u_0, v_0 be given* by (3.12), (3.13) *with*

$$\tau(u_0, v_0) \doteqdot \sum_{k,m,l} a_{kl}^2(r_0, r_1)(\omega_{kl}^2 \mu_{kml}^2 + \nu_{kml}^2) < +\infty, \tag{3.16}$$

where

$$a_{kl}^2(r_0, r_1) = 1 \Big/ \int_{r_0}^{r_1} R_{kl}^2(r) r^{n-1} \, dr. \tag{3.17}$$

For $T > 2$, the moment problem (3.14), (3.15) *has a solution g satisfying*

$$\int_0^T \int_{\mathcal{O}} |g(x, t)|^2 \, dx \, dt \leqq C\tau(u_0, v_0) \tag{3.18}$$

for some constant $C = C(T)$.

The degree of controllability guaranteed by Theorem 3.2 is thus determined by the behavior of $a_{kl}^2(r_0, r_1)$ for large k and l. Controllability of all finite energy states can occur if and only if $\{a_{kl}^2(r_0, r_1)\}$ is bounded. Lemma 3.1 shows this is possible if

and only if $r_1 = 1$. Consequently, the strongest controllability is obtained when the controls are distributed near the boundary of the sphere. More precisely, Theorem 3.2 has the following corollaries.

COROLLARY 3.1. *Let* $u_0 \in H^1(\Omega)$, $v_0 \in L^2(\Omega)$ *and suppose that* $r_1 = 1$. *For* $T > 2$ *the moment problem* (3.14), (3.15) *has a solution* g *satisfying*

$$\int_0^T \int_{\mathcal{O}} |g(x,t)|^2 \, dx \, dt \leqq C(\|u_0\|^2_{H^1(\Omega)} + \|v_0\|^2_{L^2(\Omega)}) \tag{3.19}$$

for some constant $C = C(r_0, T)$.

COROLLARY 3.2. *Suppose that* $u_0 \in H^1(\Omega)$, $v_0 \in L^2(\Omega)$ *have expansions of the form*

$$u_0(x) = u_0(r,\theta,\phi) = \sum_{k=0}^{K} \sum_{m=1}^{m(k,p)} u_{km}(r) Y_{km}(\theta,\phi), \tag{3.20}$$

$$v_0(x) = v_0(r,\theta,\phi) = \sum_{k=0}^{K} \sum_{m=1}^{m(k,p)} v_{km}(r) Y_{km}(\theta,\phi) \tag{3.21}$$

for some finite K. *For* $T > 2$, *the moment problem* (3.14), (3.15) *has a solution* g *satisfying* (3.19) *for some constant* $C = C(r_1 - r_0, K, T)$.

Remark 3.1. The constant C of Corollary 3.2 becomes unbounded as $r_1 - r_0 \to 0$. The same is true as $K \to \infty$ unless $r_1 = 1$.

Proof of Theorem 3.2. Expand g in hyperspherical coordinates:

$$g(x,t) = \sum_{k=0}^{\infty} \sum_{m=1}^{m(k,p)} g_{km}(r,t) Y_{km}(\theta,\phi). \tag{3.22}$$

Substitute this into (3.14), (3.15) and carry out the angular integrations to obtain

$$\int_0^T \int_{r_0}^{r_1} g_{km}(r,t) R_{kl}(r) (\sin \omega_{kl} t) r^{n-1} \, dr \, dt = \omega_{kl} \mu_{kml}, \tag{3.23}$$

$$\int_0^T \int_{r_0}^{r_1} g_{km}(r,t) R_{kl}(r) (\cos \omega_{kl} t) r^{n-1} \, dr \, dt = -\nu_{kml}. \tag{3.24}$$

For each fixed k and m, (3.23) and (3.24) comprise a moment problem for g_{km}. To solve it we use the fact that if $T > 2$, for each fixed k there is a sequence $\{\sigma_{kj}(t), \tilde{\sigma}_{kj}(t)\}$ biorthogonal in $L^2(0, T)$ to $\{\sin \omega_{kl} t, \cos \omega_{kl} t\}$ such that

$$\int_0^T (\sigma^2_{kj}(t) + (\tilde{\sigma}^2_{kj}(t)) \, dt \leqq K < +\infty \tag{3.25}$$

where K depends on T but not on j or k. This follows from

$$\omega_{k,l+1} - \omega_{k,l} > \pi, \qquad l = 1, 2, \cdots,$$

proved by K. D. Graham [5], and a result of A. E. Ingham [8]. Thus, for each k,

$$\int_0^T \sigma_{kj}(t) \sin \omega_{kl} t \, dt = \int_0^T \tilde{\sigma}_{kj}(t) \cos \omega_{kl} t \, dt = \delta_{jl},$$

$$\int_0^T \sigma_{kj}(t) \cos \omega_{kl} t \, dt = \int_0^T \tilde{\sigma}_{kj} \sin \omega_{kl} t \, dt = 0, \qquad j, l = 1, 2, \cdots.$$

Define

$$\tilde{g}_{km}(r,t) = \sum_{j=1}^{\infty} [\omega_{kj} \mu_{kmj} \sigma_{kj}(t) - \nu_{kmj} \tilde{\sigma}_{kj}(t)] a^2_{kj}(r_0, r_1) R_{kj}(r), \tag{3.26}$$

where $a_{kl}^2(r_0, r_1)$ is given by (3.17), and set

$$g_{km}(r, t) = \chi(r)\tilde{g}_{km}(r, t) \tag{3.27}$$

where χ is the characteristic function of the annulus $r_0 < |x| < r_1$. Then g_{km} satisfies (3.23), (3.24), and using the orthogonality of $\{R_{kj}\}_{j=1}^{\infty}$ together with (3.25) gives

$$\int_0^T \int_{r_0}^{r_1} g_{km}^2(r, t) r^{n-1}\, dr\, dt \leqq \int_0^T \int_0^1 \tilde{g}_{km}^2(r, t) r^{n-1}\, dr\, dt$$

$$\leqq K \sum_j a_{kj}^2(r_0, r_1)(\omega_{kj}^2 \mu_{kmj}^2 + \nu_{kmj}^2).$$

Thus g defined by (3.22) is a solution to (3.14), (3.15) satisfying (3.18). □

Proof of Corollary 3.1. This is an immediate consequence of (3.11) and the following result which, when $\alpha = 0$, was proved in [6, Lemma 7.1).

LEMMA 3.2. *If* $u_0 \in H^1(\Omega)$, *then* $\sum_{k,m,l} \omega_{kl}^2 \mu_{kml}^2 \leqq C\|u_0\|_{H^1(\Omega)}^2$.

Proof. We first note that

$$\int_\Omega \nabla V \cdot \nabla W\, dx + \frac{\alpha}{\beta} \int_{\partial\Omega} VW\, d\sigma$$

defines a scalar product on $H^1(\Omega)$ (the second term is omitted if $\beta = 0$) and $\{\omega_{kl}^{-1} U_{kml}\}$ is an orthonormal system in $H^1(\Omega)$ with respect to this scalar product. In fact

$$\int_\Omega \nabla U_{kml} \cdot \nabla U_{hij}\, dx + \frac{\alpha}{\beta} \int_{\partial\Omega} U_{kml} U_{hij}\, d\sigma$$

$$= -\int_\Omega U_{hij} \Delta_n U_{kml}\, dx + \int_{\partial\Omega} U_{hij} \left(\frac{\partial U_{kml}}{\partial \nu} + \frac{\alpha}{\beta} U_{kml} \right) d\sigma$$

$$= \omega_{kl}^2 \int_\Omega U_{hij} U_{kml}\, dx = \omega_{hj} \omega_{kl} \delta_{hk} \delta_{im} \delta_{jl}.$$

The Fourier coefficients $\{\tilde{\mu}_{kml}\}$ of u_0 with respect to this orthonormal system are given by

$$\tilde{\mu}_{kml} = \frac{1}{\omega_{kl}} \int_\Omega \nabla u_0 \cdot \nabla U_{kml}\, dx + \frac{\alpha}{\beta \omega_{kl}} \int_{\partial\Omega} u_0 U_{kml}\, d\sigma$$

$$= -\frac{1}{\omega_{kl}} \int_\Omega u_0 \Delta_n U_{kml}\, dx + \frac{1}{\omega_{kl}} \int_{\partial\Omega} u_0 \left(\frac{\partial U_{kml}}{\partial r} + \frac{\alpha}{\beta} U_{kml} \right) d\sigma$$

$$= \omega_{kl} \int_\Omega u_0 U_{kml}\, dx = \omega_{kl} \mu_{kml}.$$

Hence, from Bessel's inequality,

$$\sum_k \sum_m \sum_l \omega_{kl}^2 \mu_{kml}^2 \leqq \int_\Omega |\nabla u_0|^2\, dx + \frac{\alpha}{\beta} \int_{\partial\Omega} u_0^2\, d\sigma \leqq C\|u_0\|_{H^1(\Omega)}^2. \qquad \square$$

Remark 3.2. One can also obtain the reverse inequality: If the left side is finite, then $u_0 \in H^1(\Omega)$ and

$$\|u_0\|_{H^1(\Omega)}^2 \leqq C_1 \sum_k \sum_m \sum_l \omega_{kl}^2 \mu_{kml}^2$$

(cf. [6, Lemma 8.1]).

Proof of Corollary 3.2. The hypothesis on u_0, v_0 implies $\mu_{kml} = \nu_{kml} = 0$ for all $k > K$, $m = 1, \cdots, m(k, p)$, $l = 1, 2, \cdots$. Thus the solution g constructed above is given by

$$g(x, t) = \chi(r) \sum_{k=0}^{K} \sum_{m=1}^{m(k,p)} \tilde{g}_{km}(r, t) Y_{km}(\theta, \phi),$$

where $\tilde{g}_{km}$ is given by (3.26). The conclusion therefore follows from (3.9) and Lemma 3.2. □

The next result shows that if u_0, v_0 are smoother than (3.16), a smoother solution g to (3.14), (3.15) can be found. This fact will be needed for the boundary control problems considered in the next section.

THEOREM 3.3. *Let u_0, v_0 be given by* (3.12), (3.13) *with*

$$\tilde{\tau}(u_0, v_0) \doteqdot \sum_{k,m,l} a_{kl}^2(r_0, r_1)\omega_{kl}^2(\omega_{kl}^2 \mu_{kml}^2 + \nu_{kml}^2) < +\infty,$$

where $a_{kl}^2(r_0, r_1)$ is given by (3.17). *For $T > 2$, the moment problem* (3.14), (3.15) *has a solution $g \in H^{0,1}(\mathcal{O} \times (0, T))$ satisfying*

$$\int_0^T \int_{\mathcal{O}} \left[g^2 + \left(\frac{\partial g}{\partial t} \right)^2 \right] dx\, dt \leqq C\tilde{\tau}(u_0, v_0) \tag{3.28}$$

for some constant $C = C(T)$.

Proof. The proof is the same as the proof of Theorem 3.2, except that a different biorthogonal sequence to $\{\sin \omega_{kl}t, \cos \omega_{kl}t\}$ is used to ensure that $g_{km}(r, t)$ is smoother with respect to t. Thus for $T > 2$ and for each k let $\{\eta_k(t), \eta_{kj}(t), \tilde{\eta}_{kj}(t)\}$ be biorthogonal in $L^2(0, T)$ to $\{1, \omega_{kl}^{-1} \sin \omega_{kl}t, \omega_{kl}^{-1} \cos \omega_{kl}t\}$ such that

$$\int_0^T (\eta_{kj}^2(t) + \tilde{\eta}_{kj}^2(t))\, dt \leqq K\omega_{kj}^2$$

where K is independent of k and j. Such a sequence exists for the same reasons as before. Set

$$\sigma_{kj}(t) = \int_0^t \tilde{\eta}_{kj}(s)\, ds, \qquad \tilde{\sigma}_{kj}(t) = \int_0^t \eta_{kj}(s)\, ds.$$

Then for each k, $\{\sigma_{kj}, \tilde{\sigma}_{kj}\}$ is biorthogonal in $L^2(0, T)$ to $\{\sin \omega_{kl}t, \cos \omega_{kl}t\}$ and

$$\int_0^T [\sigma_{kj}^2 + (\sigma_{kj}')^2 + \tilde{\sigma}_{kj}^2 + (\tilde{\sigma}_{kj}')^2]\, dt \leqq K\omega_{kj}^2.$$

Thus g_{km} defined by (3.26), (3.27) satisfies (3.23) and (3.24), and

$$\int_0^T \int_0^1 \left[g_{km}^2 + \left(\frac{\partial g_{km}}{\partial t} \right)^2 \right] dr\, dt \leqq K \sum_j a_{kj}^2(r_0, r_1)\omega_{kj}^2(\omega_{kj}^2 \mu_{kjm}^2 + \nu_{kjm}^2).$$

Hence g defined by (3.22) satisfies (3.14), (3.15) as well as (3.28). □

COROLLARY 3.3. *Suppose $r_1 = 1$. Assume that $u_0 \in H^2(\Omega)$, $u_1 \in H^1(\Omega)$ and that u_0 satisfies* (3.8). *For $T > 2$, the problem* (3.14), (3.15) *has a solution $g \in H^{0,1}(\mathcal{O} \times (0, T))$ satisfying*

$$\int_0^T \int_{\mathcal{O}} \left[g^2 + \left(\frac{\partial g}{\partial t} \right)^2 \right] dx\, dt \leqq C(\|u_0\|_{H^2(\Omega)}^2 + \|v_0\|_{H^1(\Omega)}^2) \tag{3.29}$$

for some constant $C = C(r_0, T)$.

COROLLARY 3.4. *Assume that* $u_0 \in H^2(\Omega)$, $u_1 \in H^1(\Omega)$ *have expansions of the form* (3.20), (3.21) *for some finite* K, *and that* u_0 *satisfies* (3.8). *For* $T > 2$, *the problem* (3.14), (3.15) *has a solution* $g \in H^{0,1}(\mathcal{O} \times (0, T))$ *satisfying* (3.29) *for some constant* $C = C(r_1 - r_0, K, T)$.

Proof. These corollaries are proved in the same manner as Corollaries 3.1 and 3.2, respectively. We only have to verify that

$$\sum_k \sum_m \sum_l \omega_{kl}^4 \mu_{kml}^2 \leqq C \|u_0\|_{H^2(\Omega)}. \tag{3.30}$$

In fact, if $\tilde{\mu}_{kml}$ are the Fourier coefficients of $\Delta_n u_0$ with respect to $\{U_{kml}\}$, then

$$\begin{aligned} \tilde{\mu}_{kml} + \omega_{kl}^2 \mu_{kml} &= \int_\Omega (U_{kml} \Delta_n u_0 + \omega_{kl}^2 u_0 U_{kml})\, dx \\ &= \int_{\partial\Omega} \left(U_{kml} \frac{\partial u_0}{\partial r} - u_0 \frac{\partial U_{kml}}{\partial r} \right) d\sigma = 0, \end{aligned}$$

since u_0 and U_{kml} each satisfy (3.8). By Parseval's equality,

$$\sum_k \sum_m \sum_l \omega_{kl}^4 \mu_{kml}^2 = \|\Delta_n u_0\|_{L^2(\Omega)}^2 \leqq \|u_0\|_{H^2(\Omega)}^2. \qquad \square$$

Remark 3.3. A reverse inequality is also true. In fact, since zero is not an eigenvalue of (3.7), (3.8), one has the a priori estimate

$$\|u_0\|_{H^2(\Omega)}^2 \leqq C \|\Delta_n u_0\|_{L^2(\Omega)}^2 \leqq C \sum_k \sum_m \sum_l \omega_{kl}^4 \mu_{kml}^2.$$

4. Application to boundary control. We first consider the problem of controlling solutions of the wave equation in an annulus by means of controls applied on the surface of the inner sphere. Let $0 < r_0 < 1$ be fixed. For $n \geqq 2$ set

$$\Omega = \{x \in R^n | r_0 < |x| < 1\}, \quad \Gamma_0 = \{x \in R^n | \, |x| = r_0\}, \quad \Gamma_1 = \{x \in R^n | \, |x| = 1\}.$$

THEOREM 4.1. *For* $x \in \Omega$ *let* u_0, v_0 *be given by* (3.12), (3.13). *Suppose* $T > 2 - r_0$ *and that*

$$\tilde{\tau}(u_0, v_0) = \sum_k \sum_m \sum_l a_{kl}^2 \omega_{kl}^2 (\omega_{kl}^2 \mu_{kml}^2 + \nu_{kml}^2) < +\infty,$$

where

$$a_{kl}^2 = 1 \Big/ \int_0^\delta R_{kl}^2(r) r^{n-1}\, dr, \qquad \delta = \min (r_0, \tfrac{1}{2}(T + r_0 - 2)).$$

Assume further that u_0 *satisfies* (4.3) *on* Γ_1. *Then there is a control* $f \in H^{1/2,1/2}(\Gamma_0 \times (0, T))$ *such that the solution to the problem*

$$\frac{\partial^2 u}{\partial t^2} - \Delta_n u = 0 \quad \text{in } \Omega \times (0, T), \tag{4.1}$$

$$u(x, 0) = u_0(x), \qquad \frac{\partial u}{\partial t}(x, 0) = v_0(x) \quad \text{in } \Omega, \tag{4.2}$$

$$\alpha_1 u + \beta_1 \frac{\partial u}{\partial r} = 0 \quad \text{on } \Gamma_1 \times (0, T), \tag{4.3}$$

$$\alpha_0 u - \beta_0 \frac{\partial u}{\partial r} = f \quad \text{on } \Gamma_0 \times (0, T), \tag{4.4}$$

satisfies

$$u(x, T) = \frac{\partial u}{\partial t}(x, T) = 0 \quad \text{in } \Omega. \tag{4.5}$$

Furthermore,

$$\|f\|^2_{H^{1/2,1/2}(\Gamma_0\times(0,T))} \leqq C\tilde{\tau}(u_0, v_0) \tag{4.6}$$

for some constant $C = C(T)$.

In (4.3), (4.4), the alphas and betas are constants satisfying $|\alpha_0| + |\beta_0| > 0$, $|\alpha_1| + |\beta_1| > 0$, $\alpha_0\beta_0 \geqq 0$, $\alpha_1\beta_1 \geqq 0$. We note that since the uncontrolled surface Γ_1 can trap waves, one can expect only rather weak controllability to obtain in Ω if the controls are constrained to act on Γ_0 alone.

Analogous to Corollary 3.4 above, one has

COROLLARY 4.1. *Assume that* $u_0 \in H^2(\Omega)$, $v_0 \in H^1(\Omega)$ *have expansions of the form* (3.20), (3.21) *for some finite* K, *and that* u_0 *satisfies* (4.3) *on* Γ_1. *Then the conclusions of Theorem* 4.1 *hold with* (4.6) *replaced by*

$$\|f\|_{H^{1/2}(\Gamma_0\times(0,T))} \leqq C(\|u_0\|_{H^2(\Omega)} + \|v_0\|_{H^1(\Omega)})$$

for some constant $C = C(r_0, K, T)$.

Remark 4.1. Corollary 4.1 extends a result in [9] where the problem (4.1)–(4.5) was considered with $n = 3$, $\beta_1 = 0$, and with initial data of the form (3.20), (3.21) but slightly smoother, using methods completely different than those we shall use here.

Proof of Theorem 4.1. This is a consequence of Theorem 3.3. Let B denote the ball $|x| < 1$, let $T > 2 - r_0$ and set $\delta = \min(r_0, (T + r_0 - 2)/2)$. Then $\tilde{T} = T + r_0 - \delta > 2$. For the data u_0, v_0, which are defined everywhere in B by (3.12), (3.13), there is a function $g \in H^{0,1}(B \times (0, \tilde{T}))$ such that the x-support of g is contained in $|x| \leqq \delta$ for each t, and such that the solution of

$$\frac{\partial^2 w}{\partial t^2} - \Delta_n w = g \quad \text{in } B \times (0, \tilde{T}),$$

$$w(x, 0) = u_0(x), \qquad \frac{\partial w}{\partial t}(x, 0) = v_0(x) \quad \text{in } B,$$

$$\alpha_1 w + \beta_1 \frac{\partial w}{\partial r} = 0 \quad \text{on } \Gamma_1 \times (0, \tilde{T}),$$

satisfies

$$w(x, \tilde{T}) = \frac{\partial w}{\partial t}(x, \tilde{T}) = 0 \quad \text{in } B. \tag{4.7}$$

Moreover, g satisfies an estimate of the form (3.28). According to [10, Thm. 2.1, p. 95], $w \in H^{2,2}(B \times (0, \tilde{T}))$ and

$$\|w\|^2_{H^{2,2}(B\times(0,\tilde{T}))} \leqq C\left(\|u_0\|^2_{H^2(\Omega)} + \|v_0\|^2_{H^1(\Omega)} + \int_0^{\tilde{T}} \int_B \left(|g|^2 + \left|\frac{\partial g}{\partial t}\right|^2\right) dx\, dt\right)$$

$$\leqq C\tilde{\tau}(u_0, v_0).$$

Let u be the restriction of w to $\Omega \times (0, T)$ and $f = (\alpha_0 w - \beta_0 \partial w/\partial r)|_{\Gamma_0\times(0,T)}$. Then u, f satisfy (4.1)–(4.4), and from the trace theorem [10, Thm. 2.1, p. 9] $f \in H^{1/2,1/2}(\Gamma_0 \times (0, T))$ and

$$\|f\|^2_{H^{1/2,1/2}(\Gamma_0\times(0,T))} \leqq C\|w\|^2_{H^{2,2}(B\times(0,\tilde{T}))} \leqq C\tilde{\tau}(u_0, v_0).$$

It remains to show that u satisfies (4.5). To do so we note that w satisfies the homogeneous wave equation in the region $\delta<|x|<1, 0<t<\tilde{T}$. It then follows from (4.7) that $w=0$ at all points (x, t) such that the forward light cone with vertex (x, t) intersects $t=\tilde{T}$ totally in the region $\delta\leqq|x|\leqq 1$. In particular, $w(x, t)=0$ on $|x|=r_0, \tilde{T}-r_0+\delta\leqq t\leqq\tilde{T}$, because of our choice of δ. Thus, in the annular cylinder $r_0\leqq|x|\leqq 1, T-r_0+\delta\leqq t\leqq T$ (when $\delta<r_0$) w satisfies the homogeneous wave equation, has zero Dirichlet data on the surface $|x|=r_0$ and satisfies $\alpha_1 w+\beta_1\partial w/\partial r=0$ on $|x|=1$. It follows that $w\equiv 0$ in this cylinder and so, in particular,

$$w(x, \tilde{T}-r_0+\delta)=\frac{\partial w}{\partial t}(x, \tilde{T}-r_0+\delta)=0, \qquad r_0<|x|<1.$$

(4.5) follows immediately since $\tilde{T}-r_0+\delta=T$. □

Corollary 4.1 is an immediate consequence of the preceding theorem, Lemmas 3.1 and 3.2, and (3.30). □

One can treat in a similar way the problem of controlling solutions of the homogeneous wave equation in a spherical region $|x|<1$ by means of a control f acting on the boundary through

$$\gamma u+\sigma\frac{\partial u}{\partial\nu}=f \quad \text{on } |x|=1, \quad t>0.$$

To do this, consider the distributed control problem (3.1)–(3.4) in a slightly larger sphere Ω: $|x|<1+\delta$ (with the original data appropriately extended) and solve this problem using controls supported in $1<|x|<1+\delta$. In this case very strong controllability obtains (cf. Corol. 3.1) since the admissible controls are supported near the boundary. The solution w to (3.1)–(3.4), restricted to the original sphere, provides a solution to the boundary control problem provided f is defined as the trace on $|x|=1$ of $\gamma w+\sigma\partial w/\partial\nu$. In order to ensure that $f\in L^2$ on the boundary, it is necessary to restrict the initial data to $H^2(\Omega)\times H^1(\Omega)$, unless $\sigma=0$, and one obtains boundary controllability in this space rather than in the finite energy space $H^1(\Omega)\times L^2(\Omega)$, whenever $T>2$. Thus the result obtained in this manner is similar to but slightly weaker than that obtained by Graham and Russell [6, Thm. 1.1].

Appendix. *Proof of Lemma* 3.1. The proof will be given under the assumption that $\beta\neq 0$. The opposite case is handled in a similar fashion.

R_{kl} satisfies the differential equation (cf. [6])

$$R_{kl}''+\frac{p+1}{r}R_{kl}'+\left[\omega_{kl}^2-\frac{k(k+p)}{r^2}\right]R_{kl}=0 \tag{A.1}$$

in $0<r\leqq 1$, and the boundary conditions

$$\alpha R_{kl}(1)+\beta R_{kl}'(1)=0, \qquad |R_{kl}(0+)|<+\infty.$$

(A.1) is transformed to selfadjoint form by setting

$$\phi_{kl}(r)=r^{(p+1)/2}R_{kl}(r).$$

Then

$$\phi_{kl}''+[\omega_{kl}^2+q(r)]\phi_{kl}=0, \qquad 0<r\leqq 1, \tag{A.2}$$

$$(\alpha-\beta(p+1)/2)\phi_{kl}(1)+\beta\phi_{kl}'(1)=0, \qquad \phi_{kl}(0+)=0, \tag{A.3}$$

where

$$q(r) = -\frac{k(k+p)+(p^2-1)/4}{r^2}. \tag{A.4}$$

and $\int_0^1 \phi_{kl}^2(r)\,dr = 1$. ϕ_{kl} is given explicitly by

$$\phi_{kl}(r) = \frac{\sqrt{2r}\,\omega_{kl}}{[\omega_{kl}^2-(k+\alpha/\beta)(k+p-\alpha/\beta)]^{1/2}} \frac{J_{k+p/2}(\omega_{kl}r)}{|J_{k+p/2}(\omega_{kl})|}. \tag{A.5}$$

The beginning of the proof is similar to the proof of Lemma 2.1. We multiply (A.2) by $\phi_{kl}'(r)$ and integrate from ρ to 1, where $0<\rho\leqq 1$, and then integrate again with respect to ρ from a to 1, assuming $a>0$. The result is

$$\omega_{kl}^2\int_a^1 \phi_{kl}^2(r)\,dr + \int_a^1 |\phi_{kl}'(r)|^2\,dr + \int_a^1 [(r-a)q'(r)+q(r)]\phi_{kl}^2(r)\,dr$$
$$= (1-a)[|\phi_{kl}'(1)|^2 + \omega_{kl}^2\phi_{kl}^2(1) + q(1)\phi_{kl}^2(1)].$$

Next, multiply (A.2) by $\phi_{kl}(r)$ and integrate the first product by parts from a to 1 to obtain

$$\omega_{kl}^2\int_a^1 \phi_{kl}^2(r)\,dr - \int_a^1 |\phi_{kl}'(r)|^2\,dr + \int_a^1 q(r)\phi_{kl}^2(r)\,dr$$
$$= \phi_{kl}(a)\phi_{kl}'(a) - \phi_{kl}(1)\phi_{kl}'(1).$$

Add this equation to the preceding one:

$$2\omega_{kl}^2\int_a^1 \phi_{kl}^2(r)\,dr = (1-a)[|\phi_{kl}'(1)|^2 + \omega_{kl}^2\phi_{kl}^2(1) + q(1)\phi_{kl}^2(1)]$$
$$-\phi_{kl}(1)\phi_{kl}'(1) + \phi_{kl}(a)\phi_{kl}'(a) - \int_a^1 [(r-a)q'(r)+2q(r)]\phi_{kl}^2(r)\,dr.$$

Using the boundary condition (A.3) at $r=1$, and the definition (A.4) of $q(r)$, the last equation can be written

$$2\omega_{kl}^2\int_a^1 \phi_{kl}^2(r)\,dr = (1-a)\left[\omega_{kl}^2 - \left(k+\frac{\alpha}{\beta}\right)\left(k+p-\frac{\alpha}{\beta}\right)\right]\phi_{kl}^2(1)$$
$$+ a\left(\frac{\alpha}{\beta}-\frac{p+1}{2}\right)\phi_{kl}^2(1) + \phi_{kl}(a)\phi_{kl}'(a) + 2ac_k\int_a^1 \frac{\phi_{kl}^2(r)}{r^3}\,dr$$

where $c_k = k(k+p)+(p^2-1)/4$. Noting the value of $\phi_{kl}(1)$ from (A.5) and dividing by $2\omega_{kl}^2$ leads to

$$\int_a^1 \phi_{kl}^2(r)\,dr = (1-a) + \frac{a}{2\omega_{kl}^2}\left(\frac{\alpha}{\beta}-\frac{p+1}{2}\right)\phi_{kl}^2(1)$$
$$+ \frac{1}{2\omega_{kl}^2}\phi_{kl}(a)\phi_{kl}'(a) + \frac{ac_k}{\omega_{kl}^2}\int_a^1 \frac{\phi_{kl}^2(r)}{r^3}\,dr. \tag{A.6}$$

All conclusions of Lemma 3.1 will be deduced from this expression.

For fixed k, the second and fourth terms on the right in (A.6) clearly tend to zero as $l\to\infty$, and so, to prove (3.9), we have to show that

$$\lim_{l\to\infty} \frac{\phi_{kl}(a)\phi_{kl}'(a)}{\omega_{kl}^2} = 0. \tag{A.7}$$

This will follow from the asymptotic expressions [12, p. 199]

$$J_\nu(z)\sim\left(\frac{2}{\pi z}\right)^{1/2}\cos\left(z-\frac{\nu\pi}{2}-\frac{\pi}{4}\right),\qquad z\to\infty,$$

$$J'_\nu(z)\sim\left(\frac{2}{\pi z}\right)^{1/2}\sin\left(z-\frac{\nu\pi}{2}-\frac{\pi}{4}\right),\qquad z\to\infty,$$

valid for fixed ν. Hence, setting $\nu=k+p/2$, we have from (A.5)

$$\text{(A.8)}\qquad \limsup_{l\to\infty}\frac{|\phi_{kl}(a)\phi'_{kl}(a)|}{\omega_{kl}^2}\leqq\lim_{l\to\infty}\frac{\text{const.}}{\omega_{kl}|\cos(\omega_{kl}-\nu\pi/2-\pi/4)|}.$$

The limit of the cosine term in the denominator must be unity as $l\to\infty$ for the following reason. One has

$$0=(\alpha-\beta p/2)J_\nu(\omega_{kl})+\beta\omega_{kl}J'_\nu(\omega_{kl})$$

$$\sim\left(\frac{2}{\pi\omega_{kl}}\right)^{1/2}\left[(\alpha-\beta p/2)\cos\left(\omega_{kl}-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)+\beta\omega_{kl}\sin\left(\omega_{kl}-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)\right],\qquad l\to\infty.$$

The expression on the right can be asymptotic to zero as $l\to\infty$ only if

$$\lim_{l\to\infty}\sin\left(\omega_{kl}-\frac{\nu\pi}{2}-\frac{\pi}{4}\right)=0.$$

(A.7) therefore follows from (A.8) and (3.9) is proved.

We next prove (3.11). It clearly suffices to do so for values of a close to 1. Since $c_k>0$ for $k\geqq 1$, we have from (A.6)

$$\int_a^1\phi_{kl}^2(r)\,dr\geqq(1-a)+\frac{a}{2\omega_{kl}^2}\left(\frac{\alpha}{\beta}-\frac{p+1}{2}\right)\phi_{kl}^2(1)+\frac{1}{2\omega_{kl}^2}\phi_{kl}(a)\phi'_{kl}(a).$$

Replace a by ρ and integrate with respect to ρ from a to 1. After interchange of the order of integration in the double integral on the left we obtain

$$\int_a^1(r-a)\phi_{kl}^2(r)\,dr+\frac{1}{4\omega_{kl}^2}\phi_{kl}^2(a)\geqq\frac{(1-a)^2}{2}+\frac{\phi_{kl}^2(1)}{4\omega_{kl}^2}\left[1+(1-a^2)\left(\frac{\alpha}{\beta}-\frac{p+1}{2}\right)\right].$$

The second term on the right is nonnegative if a is close enough to 1, and may therefore be omitted. Again replace a by ρ and integrate with respect to ρ from a to 1. The result is

$$\frac{1}{2}\int_a^1(r-a)^2\phi_{kl}^2(r)\,dr+\frac{1}{4\omega_{kl}^2}\int_a^1\phi_{kl}^2(r)\,dr\geqq\frac{(1-a)^3}{6},$$

hence

$$\left(\frac{1}{2}(1-a)^2+\frac{1}{4\omega_{kl}^2}\right)\int_a^1\phi_{kl}^2(r)\,dr\geqq\frac{(1-a)^3}{6}$$

and (3.11) is proved.

To prove (3.10) we write

$$\int_0^a\phi_{kl}^2(r)\,dr=1-\int_a^1\phi_{kl}^2(r)\,dr$$

and insert (A.6) on the right. Thus

$$\int_0^a \phi_{kl}^2(r)\,dr = a\Big(1-\frac{c_k}{\omega_{kl}^2}\int_a^1 \frac{\phi_{kl}^2(r)}{r^3}\,dr\Big)-\frac{a}{2\omega_{kl}^2}\Big(\frac{\alpha}{\beta}-\frac{p+1}{2}\Big)\phi_{kl}^2(1)$$
$$-\frac{1}{2\omega_{kl}^2}\phi_{kl}(a)\phi_{kl}'(a).$$

The first term on the right does not exceed

$$a\Big(1-\frac{c_k}{\omega_{kl}^2}\int_a^1 \phi_{kl}^2(r)\,dr\Big) = a\Big[1-\frac{c_k}{\omega_{kl}^2}\Big(1-\int_0^a \phi_{kl}^2(r)\,dr\Big)\Big].$$

Hence

$$\Big(1-a\frac{c_k}{\omega_{kl}^2}\Big)\int_0^a \phi_{kl}^2(r)\,dr \leqq \Big|1-\frac{c_k}{\omega_{kl}^2}\Big|-\frac{a}{2\omega_{kl}^2}\Big(\frac{\alpha}{\beta}-\frac{p+1}{2}\Big)\phi_{kl}^2(1)$$
$$-\frac{1}{2\omega_{kl}^2}\phi_{kl}(a)\phi_{kl}'(a).$$

Replace a by ρ and integrate with respect to ρ from a to 1, assuming $a>0$:

$$\int_a^1 \Big(1-\rho\frac{c_k}{\omega_{kl}^2}\Big)\int_0^\rho \phi_{kl}^2(r)\,dr\,d\rho$$
$$\leqq \Big|1-\frac{c_k}{\omega_{kl}^2}\Big|+\frac{1}{4\omega_{kl}^2}\phi_{kl}^2(a)+\frac{\phi_{kl}^2(1)}{4\omega_{kl}^2}\Big[(a^2-1)\Big(\frac{\alpha}{\beta}-\frac{p+1}{2}\Big)-1\Big].$$

The last term is nonpositive if a is sufficiently close to 1, clearly the only case that need be considered. We drop that term, replace a by σ and integrate with respect to σ from a to 1 to obtain

$$\int_a^1\int_\sigma^1 \Big(1-\rho\frac{c_k}{\omega_{kl}^2}\Big)\int_0^\rho \phi_{kl}^2(r)\,dr\,d\rho\,d\sigma \leqq \Big|1-\frac{c_k}{\omega_{kl}^2}\Big|+\frac{1}{4\omega_{kl}^2}. \tag{A.9}$$

The right-hand side goes to zero as $k\to\infty$ since, as we shall show in a moment,

$$\lim_{k\to\infty}\frac{c_k}{\omega_{kl}^2}=1, \qquad l=1,2,\cdots. \tag{A.10}$$

To show that this implies (3.10) we interchange the ρ and σ integrations on the left side of (A.9). The resulting integral is

$$\int_a^1 \Big(1-\rho\frac{c_k}{\omega_{kl}^2}\Big)(\rho-a)\int_0^\rho \phi_{kl}^2(r)\,dr\,d\rho$$
$$=\int_a^1 (1-\rho)(\rho-a)\int_0^\rho \phi_{kl}^2(r)\,dr\,d\rho+\Big(1-\frac{c_k}{\omega_{kl}^2}\Big)\int_a^1 \rho(\rho-a)\int_0^\rho \phi_{kl}^2(r)\,dr\,d\rho$$
$$\geqq \frac{(1-a)^3}{6}\int_0^a \phi_{kl}^2(r)\,dr+\Big(1-\frac{c_k}{\omega_{kl}^2}\Big)\int_a^1 \rho(\rho-a)\int_0^\rho \phi_{kl}^2(r)\,dr\,d\rho.$$

The last term is zero in the limit as $k\to\infty$ because of (A.10), and (3.10) follows immediately.

To prove (A.10) we first note that for each k, the ω_{kl} are interlaced with the positive zeros j_{kl} of $J_{k+p/2}(j_{kl})=0$, $l=1,2,\cdots$. This fact is known as Dixon's theorem

[12, p. 480]. Secondly, we note the asymptotic expansion [1, p. 371]

$$j_{kl} \sim \left(k+\frac{p}{2}\right) Z\left(\left(k+\frac{p}{2}\right)^{-2/3} a_l\right), \qquad k \to \infty,$$

where Z is a certain continuous function which satisfies $Z(0)=1$, and a_l is the lth negative zero of the Airy function $\mathrm{Ai}(r)$. Consequently,

$$\lim_{k\to\infty} \frac{k}{j_{kl}} = 1, \qquad l = 1, 2, \cdots,$$

and therefore, because of the interlacing property of the ω_{kl} with the j_{kl},

$$\lim_{k\to\infty} \frac{k}{\omega_{kl}} = 1, \qquad l = 1, 2, \cdots. \tag{A.11}$$

(A.10) follows immediately from (A.11). □

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, Applied Mathematics Series, vol. 55, National Bureau of Standards, Washington, DC, 1964.

[2] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, Vol. I, Interscience, New York, 1965.

[3] A. ERDÉLYI, W. MAGNUS, F. OBERHETTINGER AND F. G. TRICOMI, *Higher Transcendental Functions*, Vol. II, McGraw-Hill, New York, 1953.

[4] H. FATTORINI, *Estimates for sequences biorthogonal to certain complex exponentials and boundary control of the wave equation*, in Lecture Notes in Control and Information Science 2, Springer-Verlag, Berlin, 1977.

[5] K. D. GRAHAM, *Separation of eigenvalues of the wave equation for the unit ball in R^n*, Studies in Appl. Math., 52 (1973), pp. 329–343.

[6] K. D. GRAHAM AND D. L. RUSSELL, *Boundary control of the wave equation in a spherical region*, SIAM J. Control, 13 (1975), pp. 174–196.

[7] C. J. HERGET, *On the controllability of distributed parameter systems*, Internat. J. Control, 11 (1970), pp. 827–833.

[8] A. E. INGHAM, *Some trigonometric inequalities with applications to the theory of series*, Math. Z., 41 (1936), pp. 367–379.

[9] J. LAGNESE, *Boundary patch control of the wave equation in some non-star complemented regions*, J. Math. Anal. Appl., 77 (1980), pp. 364–380.

[10] J.-L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. II, Springer-Verlag, Berlin, 1972.

[11] D. L. RUSSELL, *Nonharmonic Fourier series in the control theory of distributed parameter systems*, J. Math. Anal. Appl., 18 (1967), pp. 542–560.

[12] G. N. WATSON, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Univ. Press, London, 1952.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 1, January 1983

0363-0129/83/2101-0005 $01.25/0

ON A CERTAIN CONTROLLABILITY GAP*

IAN R. PETERSEN† AND B. ROSS BARMISH†

Abstract. This paper is concerned with the problem of steering the state $x(t)$ of the system $\dot{x}(t) = Ax(t)+Bu(t)$ to a prescribed closed and convex target set X in R^n. In addition, admissible control values $u(t)$ are constrained to lie in a prescribed compact set Ω in R^m. In light of the constraint, the term *constrained controllability problem* is often used to describe the situation above. More precisely, the system is said to be *globally Ω-controllable* to X if every initial state x_0 can be steered to X in finite time. In a recent paper [B. R. Barmish and W. E. Schmitendorf, IEEE Trans. Automat. Control, AC-25 (1980), pp. 540–547.] two *separate* conditions were given for global Ω-controllability to X. The first of these conditions was a necessary condition and the second of these conditions was a sufficient condition. This led to the possibility of a so-called *controllability gap*; that is, the possible existence of systems which satisfy one condition but not the other.

In this paper, we show that the controllability gap vanishes for a large class of systems. Namely if X is compact, the interior of Ω is nonempty and rank $[B : AB : A^2B : \cdots A^{n-1}B] = n$, then the sufficient condition of Barmish and Schmitendorf is also a necessary condition for global Ω-controllability to X. We also give examples to show that the controllability gap persists if these additional assumptions are not made.

Key words. controllability, target sets, rest points, optimal control

1. Introduction.

In a recent paper [1], some new controllability criteria were developed for systems with constraints on the input. The so-called *constrained controllability problems* considered in [1] involve dynamical systems described by state equations of the form

$$\dot{x}(t) = A(t)x(t)+f(t, u(t)), \qquad t \in [0, \infty)$$

with the input $u(t)$ restricted to a given set Ω. Within this framework, the main result given in [1] is a *necessary* condition and a *sufficient* condition which indicates whether an arbitrary initial state x_0 can be steered to a pre-specified target X, in finite time. If such a control exists then the system is said to be *globally Ω-controllable to X.*

The fact that a separate necessary condition and sufficient condition was obtained leads one to investigate the possibility of a "controllability gap". That is, it is of interest to know if there exist systems which satisfy the sufficient condition but *not* the necessary condition. The objective of this paper is to show that for a large class of linear time invariant systems, the sufficient condition given in [1] is also necessary. Hence, there is no controllability gap for this class. We shall also give examples of systems which are outside of this class and in fact fall within the gap described above.

Henceforth, we restrict our attention to linear time invariant state equations described by

$$\text{(S)} \qquad \dot{x}(t) = Ax(t)+Bu(t), \qquad t \in [0, \infty),$$

where $x(t) \in R^n$ is the *state* and $u(t) \in R^m$ is the *control.* A pre-specified *control restraint set* $\Omega \subseteq R^m$ is taken as given and an admissible control is a Lebesgue measurable vector function $u(\cdot)$ such that $u(t) \in \Omega$ for all $t \geqq 0$. The set of all such control functions is denoted $M(\Omega)$. Finally, a target set $X \subseteq R^n$ is also pre-specified. In the sequel we

* Received by the editors May 29, 1981, and in revised form March 8, 1982. This work was supported in part by Rochester Gas and Electric Corporation and in part by the U.S. Department of Energy under contract ET-78-S-01-3390; it was presented at the 1981 Conference on Systems and Information Sciences, Johns Hopkins University.

† Department of Electrical Engineering, University of Rochester, Rochester, New York 14627.

shall frequently use the following standard assumptions:

A1. X is closed and convex.

A2. Ω is compact.

For the special case when $X = \{0\}$, the problem of steering to the origin is referred to as the *Ω-null controllability* problem. A classical result on the *Ω-null controllability* problem is given in [2] for the case when the interior of Ω (denoted int Ω) contains the origin. In [3], the assumption that $0 \in \operatorname{int} \Omega$ is removed and Ω is assumed to be the closed unit hypercube in the nonnegative orthant. Finally, further criteria for *Ω-null controllability* are given in [4] for the case when Ω is a more general restraint set.

This paper generalizes the results in [4] by providing a *constrained controllability* criterion for the case when the target is any compact set rather than the origin. For this situation, some results are given in [5] and [6]. These results, however, apply only if some further special assumptions are made on the system and target; that is, it must be assumed that X is symmetric about the origin and satisfies a positive invariance condition. For the linear time invariant system (S), this amounts to the requirement that

$$e^{-A'\tau}X \subseteq e^{-A'\tau'}X \quad \text{for all } \tau' \geqq \tau.$$

More recently, it was shown in [1] that one can in fact obtain criteria for Ω-controllability to X under much weaker hypotheses. It is within this context that the previously mentioned controllability gap arises.

The plan of this paper is as follows. In § 2, we provide the basic definitions and notation. Section 3 describes the results of [1] and in particular the so-called *controllability gap* which arises in that paper.

Section 4 presents the main result of this paper. Namely, under a mild strengthening of hypotheses[1] on A, B and X, it is shown that the controllability gap disappears. Hence, the sufficient condition given in [1] is also necessary for global Ω-controllability to X.

Section 5 gives examples to show that if we do not make any further assumptions over and above A1 and A2, then one can easily construct systems which "fall into the gap." That is, there are systems for which the necessary condition is not sufficient and there are systems for which the sufficient condition is not necessary.

2. Definitions and notation. Let $x_0 \in R^n$ and $u(\cdot) \in M(\Omega)$, be given. Then, for initial condition $x(0) \triangleq x_0$, we denote the state of the system (S) at time t by $x(t, x_0, u(\cdot))$. Given the initial condition $x_0 \in R^n$, the system (S) is said to be *Ω-controllable to X from x_0* if there exists a finite time $T \geqq 0$ and an admissible control $u(\cdot) \in M(\Omega)$ such that $x(T, x_0, u(\cdot)) \in X$. The set of initial states x_0 which can be steered to X at time T is denoted by $X_0(T)$; i.e.,

$$X_0(T) \triangleq \{x_0 \in R^n : x(T, x_0, u(\cdot)) \in X \text{ for some } u(\cdot) \in M(\Omega)\}.$$

Next, we say that (S) is *globally Ω-controllable to X* if (S) is Ω-controllable to X *from every $x_0 \in R^n$*. The term *global Ω-null controllability* is coined specially to denote global Ω-controllability to $\{0\}$. Finally, we define the *domain of Ω-null controllability* to be the set of initial states x_0 which are Ω-controllable to $\{0\}$.

[1] Namely, all one requires is controllability of the pair (A, B), that Ω have nonempty interior, and compactness of X.

In [1], two functions are instrumental to describing global Ω-controllability criteria. They are $V: R^n \times [0, \infty) \to R \cup \{+\infty\}$ and $W: [0, \infty) \to R \cup \{+\infty\}$ where

$$V(z_0, t) \triangleq \int_0^t \max_{\omega \in \Omega} \omega' B' e^{-A'\tau} z_0 \, d\tau - \inf_{x \in X} x' e^{-A't} z_0 \tag{2.1}$$

and

$$W(t) \triangleq \min \{V(z_0, t): \|z_0\| = 1\}. \tag{2.2}$$

Note that the terms $e^{-A'\tau}z_0$ and $e^{-A't}z_0$ in (2.1) can be viewed as the response of the adjoint system

$$\dot{z}(t) = -A'z(t), \qquad t \in [0, \infty) \tag{S'}$$

with initial state $z(0) = z_0$.

3. The controllability gap. The take-off point for this paper is the following three theorems which were proven in [1].

THEOREM 3.1. *Consider the system* (S) *satisfying* A1 *and* A2 *and let* $x_0 \in R^n$ *be given. Then* $x_0 \in X_0(T)$ *if and only if*

$$x_0' z_0 + V(z_0, T) \geqq 0 \tag{3.1}$$

for all nonzero vectors $z_0 \in R^n$.

THEOREM 3.2. *Consider the system* (S) *satisfying* A1 *and* A2. *Then the following condition is necessary for global* Ω*-controllability to* X:

$$\sup_{t \geqq 0} V(z_0, t) = +\infty \tag{3.2}$$

for all nonzero vectors $z_0 \in R^n$.

Note that since $V(z_0, t)$ is positively homogeneous in z_0, one need only verify (3.2) for $z_0 \in R^n$ such that $\|z_0\| = 1$.

THEOREM 3.3. *Consider the system* (S) *satisfying* A1 *and* A2. *Then the following condition is sufficient for global* Ω*-controllability to* X:

$$\sup_{t \geqq 0} W(t) = +\infty. \tag{3.3}$$

It is the difference between the conditions given in Theorems 3.2 and 3.3 which gives rise to the so-called *controllability gap*. In [1], it was shown that for the special case $X = \{0\}$, the gap between necessary condition (3.2) and sufficient condition (3.3) disappears.[2] In the sequel, we investigate the controllability gap without this restriction on the target X.

4. The main result. Our main result is given in the theorem below.

THEOREM 4.1. *Consider the system* (S) *satisfying* A1 *and* A2. *Furthermore, assume that*:

A3. X *is compact.*

A4. $\operatorname{int} \Omega \neq \varnothing$.

A5. $\operatorname{rank} [B : AB : A^2B \cdots A^{n-1}B] = n$.

Then the following condition is necessary and sufficient for global Ω*-controllability to* X:

$$\sup_{t \geqq 0} W(t) = +\infty.$$

[2] In fact, for this null controllability situation a single necessary and sufficient condition can be given even for time-varying systems; e.g., see [9].

In order to prove this theorem, we shall exploit the three lemmas to follow.

LEMMA 4.1. *Consider the system* (S) *satisfying* A1 *and* A2. *Furthermore, suppose that*

$$\sup_{t \geqq 0} W(t) \leqq \beta$$

for some finite β. *Then there exists an initial state* $\tilde{x}_0 \in R^n$ *and an increasing sequence of times* $\langle \tilde{t}_k \rangle_{k=1}^{\infty}$ *such that*

(i) $\tilde{t}_k \to \infty$ *as* $k \to \infty$,

(ii) $\tilde{x}_0 \notin X_0(\tilde{t}_k)$ *for all* k.

Proof. If $\sup_{t \geqq 0} W(t) \leqq \beta$, then we first define the sequence $t_k = k$ for $k = 1, 2, 3, \cdots$. Now, it is clear that $t_k \to \infty$ as $k \to \infty$ and $W(t_k) \leqq \beta$ for all k. By definition of $W(\cdot)$ given in (2.2), we can infer that for each k, there exists a vector $z_{0k} \in R^n$ such that $\|z_{0k}\| = 1$ and $V(z_{0k}, t_k) \leqq \beta$. Now, for each k define $x_{0k} \triangleq -2\beta z_{0k}$. Hence,

$$x_{0k}' z_{0k} + V(z_{0k}, t_k) \leqq -\beta \tag{4.1}$$

for all k. Since $\langle x_{0k} \rangle_{k=1}^{\infty}$ is an infinite sequence in the compact set $\{x : \|x\| = 2\beta\}$, there must be a subsequence $\langle \tilde{x}_{0k} \rangle_{k=1}^{\infty}$ which converges to some vector $\tilde{x}_0$ with $\|\tilde{x}_0\| = 2\beta$. Let $\langle \tilde{z}_{0k} \rangle_{k=1}^{\infty}$ denote the corresponding subsequence of $\langle z_{0k} \rangle_{k=1}^{\infty}$ and let $\langle \bar{t}_k \rangle_{k=1}^{\infty}$ be the corresponding subsequence of $\langle t_k \rangle_{k=1}^{\infty}$.

Since $\langle \bar{t}_k \rangle_{k=1}^{\infty}$ is an increasing subsequence of the positive integers, it is clear that $\bar{t}_k \to \infty$ as $k \to \infty$. Also, in light of (4.1), it follows that

$$\tilde{x}_{0k}' \tilde{z}_{0k} + V(\tilde{z}_{0k}, \bar{t}_k) \leqq -\beta \tag{4.2}$$

for all k. Furthermore, since $\tilde{x}_{0k} \to \tilde{x}_0$ as $k \to \infty$, there exists a positive integer N such that $\|\tilde{x}_0 - \tilde{x}_{0k}\| \leqq \beta/2$ for all $k \geqq N$. Hence, for such $k \geqq N$, it is clear using (4.2) that

$$\begin{aligned} \tilde{x}_0' \tilde{z}_{0k} + V(\tilde{z}_{0k}, \bar{t}_k) &= \tilde{x}_{0k}' \tilde{z}_{0k} + V(\tilde{z}_{0k}, \bar{t}_k) + (\tilde{x}_0 - \tilde{x}_{0k})' \tilde{z}_{0k} \\ &\leqq \tilde{x}_{0k}' \tilde{z}_{0k} + V(\tilde{z}_{0k}, \bar{t}_k) + \|\tilde{x}_0 - \tilde{x}_{0k}\| \cdot \|\tilde{z}_{0k}\| \\ &\leqq \tilde{x}_{0k}' \tilde{z}_{0k} + V(\tilde{z}_{0k}, \bar{t}_k) + \frac{\beta}{2} \\ &\leqq -\frac{\beta}{2}. \end{aligned}$$

By invoking Theorem 3.1, we see that $\tilde{x}_0 \notin X_0(\bar{t}_k)$ for all $k \geqq N$. We conclude that $\tilde{x}_0$ is the required state and $\langle \tilde{t}_k \rangle_{k=1}^{\infty} = \{\bar{t}_k : k \geqq N\}$ is the required infinite sequence of times. □

DEFINITION. An int $CH(\Omega)$ *rest-point* for the system (S) is defined to be a pair $(x^*, u^*) \in R^n \times R^m$ such that $Ax^* + Bu^* = 0$ and $u^* \in \text{int } CH(\Omega)$, the interior of the convex hull of Ω.

LEMMA 4.2. *Consider the system* (S) *satisfying Assumptions* A1–A4. *Furthermore, assume that* (S) *is globally* Ω*-controllable to* X. *Then the set of* int $CH(\Omega)$ *rest-points is nonempty.*

Proof. Proceeding by contradiction, suppose that there are no int $CH(\Omega)$ rest-points. Then, in accordance with [7, Lemma 2], there exists a nonzero vector z_0 such that $A'z_0 = 0$ and $\omega' B' z_0 \leqq 0$ for all $\omega \in CH(\Omega)$. Therefore,

$$\max_{\omega \in CH(\Omega)} \omega' B' z_0 \leqq 0. \tag{4.3}$$

Noting that the expression $\omega' B' z_0$ is convex with respect to ω, it follows from [8, Thm. 32.2] that

$$\max_{\omega \in CH(\Omega)} \omega' B' z_0 = \max_{\omega \in \Omega} \omega' B' z_0.$$

This implies that

$$\max_{\omega \in \Omega} \omega' B' z_0 \leqq 0. \tag{4.4}$$

Since $A' z_0 = 0$, it is evident that $e^{-A't} z_0 = z_0$ for all $t \geqq 0$. Noting this fact, (2.1) becomes

$$V(z_0, t) = \int_0^t \max_{\omega \in \Omega} \omega' B' z_0 \, d\tau - \inf_{x \in X} x' z_0.$$

Since X is compact, the second term above is a finite constant. Furthermore, in view of (4.4), the first term must be nonpositive. Hence,

$$\sup_{t \geqq 0} V(z_0, t) < +\infty.$$

This is clearly a contradiction to Theorem 3.1. □

LEMMA 4.3. *Consider the system* (S) *satisfying* A1–A5. *If in addition* $0 \in \text{int } CH(\Omega)$ *and* (S) *is globally* Ω*-controllable to* X, *then* (S) *is globally* Ω*-null controllable.*

Proof. For this system, the classical result given in [2] states that the domain of null controllability is an open set containing the origin. Hence there exists an $\varepsilon > 0$ such that the ball $B_\varepsilon \triangleq \{x : \|x\| < \varepsilon\}$ is contained in the domain of null controllability. We now claim that any initial state $x_0 \in R^n$ can be steered into B_ε. Note that this would be sufficient for global Ω-null controllability.

To prove this claim, we first observe that the compactness of X implies that it can be enclosed in an open ball of finite radius, say R, which is centered at the origin. Given the positive ε above, we now select $\varepsilon_1 > 0$ such that $\varepsilon_1 \leqq \varepsilon$ and $R/\varepsilon_1 > 1$. Now, let $x_0 \in R^n$ be an arbitrary initial state. We must prove that x_0 can be steered to B_ε. By the hypothesis of global Ω-controllability to X, there exists a control $u_1(\cdot) \in M(\Omega)$ which steers the initial state $\tilde{x}_0 \triangleq x_0 R/\varepsilon_1$ to X in some finite time, say T. Hence, if

$$x_1 \triangleq e^{AT} x_0 \frac{R}{\varepsilon_1} + \int_0^T e^{A(T-\tau)} B u_1(\tau) \, d\tau, \tag{4.5}$$

then $x_1 \in X$. Therefore $\|x_1\| < R$. Multiplying (4.5) by the factor ε_1/R makes it apparent that the control $u_0(t) \triangleq (\varepsilon_1/R) u_1(t)$ steers the initial state x_0 to $x_2 \triangleq \varepsilon_1 x_1/R$ at time T. Moreover, note that $x_2 \in B_{\varepsilon_1}$. Recalling that $\varepsilon_1/R < 1$, $0 \in CH(\Omega)$ and $u_1(t) \in \Omega$ for all $t \geqq 0$, it is evident that $u_0(\cdot) \in M(CH(\Omega))$. Hence, we have a control in $M(CH(\Omega))$ which steers the initial state x_0 to B_ε in finite time. Since the attainable set is the same for control restraint sets Ω and $CH(\Omega)$ (see [2]), the preceding analysis implies that there must exist another control $u_2(\cdot) \in M(\Omega)$ which steers x_0 to B_ε in the same time T. This completes the proof. □

Proof of Theorem 4.1 (*Sufficiency*). If $\sup_{t \geqq 0} W(t) = +\infty$, then Theorem 3.3 implies that the system (S) is globally Ω-controllable to X.

Necessity. Proceeding by contradiction, suppose that $\sup_{t \geqq 0} W(t) \leqq \beta < +\infty$. Then by Lemma 4.1, there exists an initial state $\tilde{x}_0 \in R^n$ which *cannot* be steered to X at times $\langle \tilde{t}_k \rangle_{k=1}^{\infty}$ where $\tilde{t}_k \to \infty$ as $k \to \infty$. That is, $\tilde{x}_0 \notin X_0(\tilde{t}_k)$ for all k.

Since the system (S) is globally Ω-controllable to X, Lemma 4.2 implies that there exists an int $CH(\Omega)$ rest-point (x^*, u^*). Letting $y \triangleq x - x^*$ and $v \triangleq u - u^*$, we generate

a *translated system* described by

$$(S_y) \qquad \dot{y}(t) = Ay(t) + Bv(t)$$

with control restraint set $V \triangleq \{v \in R^m : v + u^* \in \Omega\}$ and target $Y \triangleq \{y \in R^n : y + x^* \in X\}$. Since X is compact and convex and Ω is compact, Y must also be compact and convex and V must be compact. Furthermore, since $u^* \in \operatorname{int} CH(\Omega)$ we must have $0 \in \operatorname{int} CH(V)$. Noticing that the system (S_y) has the same A and B matrices as the system (S), it follows that (S_y) satisfies all the assumptions of Lemma 4.3. Therefore, (S_y) is globally V-null controllable which in turn implies that the system (S) is globally Ω-controllable to $\{x^*\}$.

Consequently, there exists a control $\tilde{u}(\cdot) \in M(\Omega)$ which steers the specifically constructed initial state $\tilde{x}_0$ to x^* in some finite time, say $\tilde{T}$. Since (S) is Ω-controllable to X, there also exists a second control $u^*(\cdot) \in M(\Omega)$ which steers the initial state x^* to X in some finite time, say T^*. Recalling that $\tilde{t}_k \to \infty$ as $k \to \infty$, one can select some sufficiently large k, say $k = N$, such that $\tilde{t}_N > \tilde{T} + T^*$.

We now assert that $\tilde{x}_0$ can be steered to X at time $\tilde{t}_N$. Indeed, this transfer of state is accomplished in three stages. First we apply the control $\tilde{u}(\cdot)$ for $t \in [0, \tilde{T}]$. This will lead to $x(\tilde{T}) = x^*$. Since (x^*, u^*) is an $\operatorname{int} CH(\Omega)$ rest-point, the constant control $u(t) \equiv u^* \in CH(\Omega)$ applied for $t \in (\tilde{T}, \tilde{t}_N - T^*)$ leads to $x(\tilde{t}_N - T^*) = x^*$. Recalling that the attainable set is the same for constraint sets Ω and $CH(\Omega)$ (see [2]), there exists a control $\bar{u}(\cdot) \in M(\Omega)$ having the following property: If one applies the control $\bar{u}(\cdot)$ for $t \in (\tilde{T}, \tilde{t}_N - T^*)$ this will lead to $x(\tilde{t}_N - T^*) = x^*$. Finally, one applies the control $u(t) = u^*(t - (\tilde{t}_N - T^*))$ for $t \in [\tilde{t}_N - T^*, \tilde{t}_N]$. The result is that $x(\tilde{t}_N) \in X$ (time invariance of (S) is used implicitly here). This contradicts the fact that

$$\tilde{x}_0 \notin X_0(\tilde{t}_N) \quad \text{for all } k. \qquad \square$$

5. Examples. In this section we give three examples of linear time invariant systems which indicate why one requires A3–A5 in Theorem 4.1. In the first two examples, these assumptions are weakened and we generate systems which fall within the controllability gap. That is, these systems are globally Ω-controllable to X, necessary condition (3.2) holds and sufficient condition (3.3) fails. Hence, (3.3) is not necessary if A3–A5 is violated. In the third example, we consider a system which satisfies all of the assumptions of Theorem 4.1. Despite the fact that necessary condition (3.2) is satisfied, this system is *not* globally Ω-controllable to X. Thus, when A1–A5 are satisfied, Theorem 4.1 provides a stronger necessary condition than (3.2).

Example 1 (A4 and A5 violated). Under these conditions, we show that global Ω-controllability to X is possible and yet $\sup_{t \geqq 0} W(t) < +\infty$. Consider the state equations

$$\dot{x}_1(t) = -x_1(t) + x_2(t), \qquad \dot{x}_2(t) = -x_1(t) - x_2(t), \qquad t \in [0, \infty)$$

with target set $X = \{(x_1, x_2) : |x_1| \leqq 1, x_2 = 0\}$. Note that since no control term appears in the system equations, it does not matter what control restraint set is chosen.

For this system the state transition matrix is

$$\phi(t, 0) = e^{At} = e^{-t} \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

Therefore, the response is given by

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = e^{-t} \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \cdot \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$$

from initial state $x_0 = [x_{01}\ x_{02}]'$. To establish global Ω-controllability to X, notice that for any given initial state x_0, it is possible to choose a sufficiently large time T so that $x_{01} \sin T = x_{02} \cos T$ and $|x_{01}| + |x_{02}| \leqq e^T$. Since $x_1(T) = e^{-T}(x_{01} \cos T + x_{02} \sin T)$, and $x_2(T) = e^{-T}(-x_{01} \sin T + x_{02} \cos T)$, it is clear (from choice of T) that $|x_1(T)| \leqq 1$ and $x_2(T) = 0$. Therefore $x(T) \in X$. This shows that this system is globally Ω-controllable to X.

For this "unforced" system, a straightforward calculation (using (2.1)) yields

$$V(z_0, t) = -\inf_{\substack{|x_1| \leqq 1 \\ x_2 = 0}} [x_1, x_2] e^t \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \cdot \begin{bmatrix} z_{01} \\ z_{02} \end{bmatrix}$$

$$= e^t |z_{01} \cos t + z_{02} \sin t|,$$

where $z_0 = [z_{01}\ z_{02}]'$. Now using (2.2)

$$W(t) = \min_{\|z_0\| = 1} e^t |z_{01} \cos t + z_{02} \sin t|.$$

To see that $\sup W(t) < +\infty$, we observe that at each fixed $t \in [0, \infty)$, the minimum describing $W(t)$ above is achieved by $z_{01} = -\sin t$ and $z_{02} = \cos t$. This implies that $W(t) = 0$ for all $t \in [0, \infty)$. Hence, for this system necessary condition (3.2) holds (implied by Theorem 3.1) but sufficient condition (3.3) fails.

Example 2 (A3 violated). This example illustrates the same phenomenon as in Example 1 can occur even for systems which satisfy the controllability assumption A5. The system is described by the equations

$$\dot{x}_1(t) = x_1(t) + x_2(t),$$

$$\dot{x}_2(t) = -x_1(t) + x_2(t) + u(t), \qquad t \in [0, \infty)$$

with an *unbounded* target set $X = \{(x_1, x_2): x_1 \geqq 0, x_2 = 0\}$ and a control restraint set $\Omega = [-1, 1]$. Note that for this system the controllability matrix $[B : AB] = \left[\begin{smallmatrix} 0 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ has full rank. Hence, A5 is satisfied.

For this system the state transition matrix is

$$\phi(t, 0) = e^{At} = e^t \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}.$$

To see that this system is globally Ω-controllable to X, we consider the solution to the state equation which is

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = e^t \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix} \cdot \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix}$$

with control $u(t) \equiv 0$ and initial condition $x_0 = [x_{01}\ x_{02}]'$. Hence, given any initial state x_0, we can find a time $T \geqq 0$ such that $x_{01} \sin T = x_{02} \cos T$ and $x_{01} \cos T + x_{02} \sin T \geqq 0$. For this T,

$$\begin{bmatrix} x_1(T) \\ x_2(T) \end{bmatrix} = \begin{bmatrix} e^t x_{01} \cos T + e^t x_{02} \sin T \\ 0 \end{bmatrix}$$

which implies that $x(T) \in X$. This proves that this system is globally Ω-controllable to X.

Now to show that $\sup_{t \geqq 0} W(t) < +\infty$, we first observe that any vector $z_0 \in R^n$ satisfying $\|z_0\| = 1$ can be rewritten in the "parameterized" form $z_0 = z_0(\theta) \triangleq \left[\begin{smallmatrix} \cos\theta \\ \sin\theta \end{smallmatrix}\right]$ for some $\theta \in [0, 2\pi]$. Substitution into (2.1) followed by a straightforward calculation

yields

$$V(z_0(\theta), t) = \int_0^t \max_{\omega \in \Omega} \omega\, e^{-\tau} \sin(\theta - \tau)\, d\tau - \inf_{x_1 \geqq 0} e^{-t} x_1 \cos(\theta - t)$$

$$= \int_0^t e^{-\tau} |\sin(\theta - \tau)|\, d\tau - \begin{cases} -\infty & \text{if } \cos(\theta - \tau) < 0 \\ 0 & \text{if } \cos(\theta - \tau) \geqq 0 \end{cases}$$

$$\leqq \int_0^t e^{-\tau}\, d\tau + \begin{cases} +\infty & \text{if } \cos(\theta - \tau) < 0 \\ 0 & \text{if } \cos(\theta - \tau) \geqq 0. \end{cases}$$

Now using (2.2), for each fixed $t \in [0, \infty)$

$$W(t) = \min\{V(z_0(\theta), t) : \theta \in [0, 2\pi]\} \leqq [1 - e^{-t}] \leqq 1.$$

Therefore, $\sup_{t \geqq 0} W(t) < +\infty$ which violates condition (3.3).

Example 3 (condition (3.2) holds, A1–A5 hold, and condition (3.3) fails). The system is described by the equations

$$\dot{x}_1(t) = -x_1(t) + x_2(t), \qquad \dot{x}_2(t) = -x_1(t) - x_2(t) + u(t), \qquad t \in [0, \infty),$$

with control restraint set $\Omega = [-1, 1]$ and "singleton" target set $X = \{\begin{bmatrix} 5 \\ 0 \end{bmatrix}\}$. It is obvious that A1–A4 hold. Note also that the system has controllability matrix $[B : AB] = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$ which has full rank. Hence, Assumption A5 holds as well. For this system

$$e^{-A't} = e^t \begin{bmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{bmatrix}$$

and $B = [0\ 1]'$. Now, given any $z_0 = [z_{01}\ z_{02}]' \in R^n$, substitution into (2.1) yields

$$V(z_0, t) = \int_0^t \max_{\omega \in \Omega} \omega\, e^{\tau} (z_{02} \cos \tau - z_{01} \sin \tau)\, d\tau - 5\, e^t (z_{01} \cos t + z_{02} \sin t).$$

To investigate condition (3.2) and (3.3), we rewrite z_0 as $z_0 \triangleq [\cos\theta \sin\theta]'$. Hence, we are led to examine

$$V(z_0(\theta), t) = \int_0^t \max_{\omega \in \Omega} \omega\, e^{\tau} \sin(\theta - \tau)\, d\tau - 5\, e^t \cos(\theta - \tau)$$

$$= \int_0^t e^{\tau} |\sin(\theta - \tau)|\, d\tau - 5\, e^t \cos(\theta - t).$$

To complete the argument, we observe that

(i) The first term in $V(z_0(\theta), t)$ above is positive.

(ii) For any given $\theta \in [0, 2\pi]$ and any $N \geqq 0$, there exists a $t \geqq 0$ such that $-5\, e^t \cos(\theta - t) \geqq N$. Therefore $\sup_{t \geqq 0} V(z_0(\theta), t) = +\infty$ for all $\theta \in [0, 2\pi]$. Equivalently $\sup_{t \geqq 0} V(z_0, t) = +\infty$ for all $z_0 \in R^2$ such that $\|z_0\| = 1$. Hence, condition (3.2) is satisfied.

To show that condition (3.3) is violated, we shall prove that $W(t)$ is bounded. To accomplish this, we define $\theta_t \triangleq t \bmod 2\pi$ for each $t \in [0, \infty)$. Hence,

$$V(z_0(\theta_t), t) = \int_0^t e^{\tau} |\sin(t - \tau)|\, d\tau - 5\, e^t \leqq \int_0^t e^{\tau}\, d\tau - 5\, e^t < 0.$$

This implies that $W(t) < 0$ since $W(t)$ is obtained by minimizing $V(z_0(\theta), t)$ with respect to $\theta \in [0, 2\pi]$. By Theorem 3.1 and the preceding analysis, it follows that $0 \notin X_0(t)$ for all $t \geqq 0$. Therefore the system is not Ω-controllable to X from the origin which in turn precludes global Ω-controllability to X.

6. Further extension. We recall that the target set X was assumed to be compact (Assumption A3) in Theorem 4.1. In fact, Assumption A3 may be eliminated in many circumstances of interest. In particular, if every eigenvalue λ of the matrix A satisfies $\operatorname{Re}\lambda \leqq 0$ and A is nonsingular, then one can prove Lemmas 4.2 and 4.3 without the assumption that X is compact. Hence the same proof as given in § 4 will suffice to prove Theorem 4.1 for unbounded targets. Such an extension would be useful when considering targets such as the positive orthant. Example 2 shows that some further restrictions must be made on the system (S) if unbounded targets are to be allowed.

REFERENCES

[1] B. R. BARMISH AND W. E. SCHMITENDORF, *New results on controllability of systems of the form* $\dot{x}(t) = A(t)x(t) + f(t, u(t))$, IEEE Trans. Automat. Control, AC-25 (1980), pp. 540–547.

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

[3] S. SAPERSTONE AND J. YORKE, *Controllability of linear oscillatory systems with positive controls*, this Journal, 9 (1971), pp. 253–262.

[4] R. BRAMMER, *Controllability of linear autonomous systems with positive controllers*, this Journal, 10 (1972), pp. 339–353.

[5] E. N. CHUKWU AND S. D. SILLIMAN, *Complete controllability to a closed target set*, J. Optim. Theory Appl., 21 (1977), pp. 369–383.

[6] E. N. CHUKWU AND J. GRONSKI, *Controllability of nonlinear systems with restrained controls to closed convex sets*, Rep. CSUMD 45, Dept. Mathematics, Cleveland State University, Cleveland, OH.

[7] M. HEYMANN AND R. J. STERN, *Controllability of linear systems with positive controls: geometric considerations*, J. Math. Anal. Appl., 52 (1975), pp. 36–41.

[8] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1972.

[9] W. E. SCHMITENDORF AND B. R. BARMISH, *Constrained controllability of linear systems*, this Journal, 18 (1980), pp. 327–345.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 1, January 1983

0363-0129/83/2101-0006 $01.25/0

LINEAR-QUADRATIC OPTIMAL CONTROL OF HEREDITARY DIFFERENTIAL SYSTEMS: INFINITE DIMENSIONAL RICCATI EQUATIONS AND NUMERICAL APPROXIMATIONS*

J. S. GIBSON†

Abstract. Recent theory of infinite dimensional Riccati equations is applied to the linear-quadratic optimal control problem for hereditary differential systems, and it is shown that, for most such problems, the operator solutions of the Riccati equations are of trace class (i.e., nuclear). With special attention to trace-norm convergence, an abstract approximation theory is developed and applied to a particular approximation scheme. Numerical examples are given.

Problems on both finite and infinite time intervals are studied. For both the hereditary system and the approximating systems in the infinite time problem, characteristic equations are derived for the closed-loop eigenvalues, and formulas for the corresponding eigenvectors are given.

TABLE OF CONTENTS

Abstract 95
Notation 95
1. Introduction 95
2. The hereditary system, the adjoint system and stability 97
3. The optimal control problem on a finite interval 99
4. Optimal control on the infinite interval 104
5. The traces of the operators $\Pi(t)$ and Π 110
6. Abstract approximation theory 113
7. An approximation scheme 120
7.1. Approximation of the semigroups $T(\cdot)$ and $T^*(\cdot)$ 120
7.2. Approximation for optimal control on a finite interval 122
7.3. Approximation for optimal control on the infinite interval 125
7.4. Computational aspects of the finite dimensional Riccati algebraic equations 129
8. Examples 134

1. Introduction. For at least two decades now, engineers and mathematicians have studied the linear-quadratic optimal control problem for hereditary differential systems, and many papers have resulted. Among the most important contributions have been controllability, observability and stabilizability results [13], [31], [38], descriptions of the feedback structure of the optimal control laws [1], [10], [11], [12], [26], [37], and approximation theory for computing the optimal control in feedback form [10], [27], [37], [38] and in open-loop form [4], [5]. The principal concerns of this paper are the mathematical properties of the optimal feedback control laws and the resulting closed-loop systems and the convergence of numerical approximations of the feedback control laws. We consider time-invariant control systems only (i.e., the uncontrolled systems are time-invariant), and study control on both finite and infinite intervals.

The foundation of our analysis is the theory of infinite dimensional Riccati equations developed in [19]. As with the linear-quadratic regulator for systems represented by ordinary differential equations, the solution to a Riccati equation defines the feedback structure of the optimal control law. Previous authors have studied infinite dimensional Riccati equations in connection with hereditary control

* Received by the editors July 1, 1980, and in revised form August 17, 1981. 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.

problems [6], [10], [11], [12], [13], but the results of [19] enable us to derive a number of new results for hereditary control. Among the most significant are results concerning the traces of the solutions of the Riccati equations, the eigenvalues and eigenvectors of the optimal closed-loop system for control on the infinite interval and numerical approximation of the solutions of the Riccati equations. Now let us outline the paper and discuss the main results in more detail.

In § 2, we discuss the hereditary differential equation and the equivalent evolution equation on the space $Z = R^n \times L_2(-r, 0; R^n)$ (sometimes called M^2). As usual, we represent the homogeneous solution in Z with a strongly continuous semigroup $T(\cdot)$ with generator $\mathcal{A}$. After defining $\mathcal{A}$, we give its Z-adjoint $\mathcal{A}^*$. While previous papers on control of hereditary systems have not used this adjoint operator explicitly, its explicit use is essential to the analysis in this paper. Finally in § 2, we discuss various equivalent definitions of stability and stabilizability for hereditary differential systems and give the definitions to be used here.

In § 3, we give the pertinent results from [19] for optimal control on finite time intervals and the two Riccati integral equations, and then derive some basic implications for hereditary control problems.

We begin § 4 by summarizing the pertinent results from [19] concerning optimal control on an infinite time interval, primarily existence, uniqueness and stability results for solutions to the Riccati algebraic equation, which is the steady-state version of the Riccati integral equations. Then we derive a characteristic equation for the eigenvalues of the optimal closed-loop system and several other results analogous to well-known results for finite dimensional linear regulator problems, including formulas for the eigenvectors of the optimal closed-loop system.

In § 5, we review some standard properties of trace class operators on Hilbert spaces and show that the solutions of the Riccati equations of §§ 3 and 4 are trace class operators. Also, we discuss the use of the trace of the solution of the Riccati equation for a hereditary optimal control problem as a sort of average-performance measure.

The most important contribution of this paper should be the approximation theory. Section 6 presents an abstract approximation theory for schemes which discretize the history space and leave the time variable continuous. Thus our analysis does not cover the scheme Delfour presented in [9], where he discretized both the history space and the time variable. In our analysis, the key approximation is the approximation of the semigroup $T(\cdot)$ by a sequence of semigroups $T_N(\cdot)$ which converge strongly to $T(\cdot)$. Assuming such an approximation, we define a sequence of optimal control problems for the approximating systems represented by the $T_N(\cdot)$'s, and derive convergence results for the corresponding sequences of Riccati equations and feedback control laws.

A primary objective of this paper is to call attention to the importance of having strong convergence for the solutions of the approximating Riccati equations—as opposed to weak convergence, which was obtained by Delfour [10] and Kunisch [24] for hereditary systems and Lions [30] for parabolic systems. This importance stems from the reasons for using a feedback control law in engineering design. By synthesizing the optimal control law in feedback form, the designer ensures that the system will respond optimally for any initial condition and, in the case of a time-invariant linear regulator, the closed-loop system will be asymptotically stable. For an infinite dimensional control system, a sequence of feedback control laws is based on a sequence of finite dimensional approximations of the actual system, and these control laws should converge in some sense to the optimal feedback control law as the dimension

of the approximation increases. The sense in which these control laws converge determines the answer to the following question: Is it possible to choose the model dimension sufficiently large that, when the control law based on the finite dimensional model is applied to the actual infinite dimensional system, the response of the resulting closed-loop system will be close to optimal for any initial condition and the system will be stable? While this is one of the most important questions to ask about an approximation scheme for control of a distributed system, previous authors have neglected it. As we will show in § 6, the answer is yes if the solutions to the sequence of finite dimensional Riccati equations converge strongly. The reason is, since the control space is finite dimensional, strong convergence of the operator solutions of the Riccati equations yields uniform norm convergence for the feedback control laws. Similar results are given for hyperbolic systems in [18].

As we will see, the key to obtaining strong convergence for the solutions of the approximating Riccati equations is to have the adjoint semigroups $T_N^*(\cdot)$ converge strongly. Because of the finite dimensionality of the performance measures here, strong convergence for the adjoint semigroups actually results in trace-norm convergence for the solutions of the finite dimensional Riccati equations. As discussed in § 5, the trace-norm is the strongest of all the common operator norms.

In § 7, we apply the theory of § 6 to an approximation scheme that in one form or another has been used by a number of authors for hereditary differential control problems [4], [23], [36], [37], [38], [39]. In [4], Banks and Burns have chronicled the evolution of this scheme and cast it in its most recent form as a Ritz method which approximates the history function by a finite number of piecewise constant functions. Banks and Burns then used the Trotter–Kato approximation theorem to show that their sequence of approximating semigroups converge strongly to the semigroup representing the homogeneous solution of the hereditary system, as needed in our § 6. Because Banks and Burns considered only open-loop control on finite time intervals, they did not need to and did not raise the question of whether the adjoint semigroups converge strongly; however, as we will show, strong convergence of the adjoint semigroups follows from Banks and Burns' basic convergence results, once we have defined the adjoint of the generator of the semigroup $T(\cdot)$. Finally, in § 7, we give some results on the eigenvalues and eigenvectors of the approximating closed-loop systems that should be useful for numerical solution of the approximating Riccati equations in control on the infinite interval.

We present three numerical examples in § 8.

2. The hereditary system, the adjoint system and stability. We consider the differential equation

$$\dot{x}(t) = Lx_t + B_0 u(t), \qquad t \geqq t_0, \tag{2.1}$$

where

$$x(t) \in R^n, \; x_t \in L_2(-r, 0; R^n) \quad \text{for some } r \geqq 0$$

and

$$x_t(\theta) = x(t+\theta), \quad B_0 \in \mathscr{L}(R^m, R^n), \quad u \in L_2(t_0, t_1; R^m) \quad \text{for all } t_1 < \infty.$$

The linear operator L has the form

$$L\phi = \sum_{i=0}^{\nu} A_i \phi(-h_i) + \int_{-r}^{0} D(\theta)\phi(\theta)\, d\theta, \tag{2.2}$$

where $A_i \in \mathscr{L}(R^n, R^n)$ for $0 \leqq i \leqq \nu < \infty$, $0 = h_0 < h_1 < \cdots < h_\nu = r$, and $D(\cdot) \in L_2(-r, 0; \mathscr{L}(R^n, R^n))$. It is well known [13], [14], [19] that, for any $x(t_0) \in R^n$ and $x_{t_0} \in L_2(-r, 0; R^n)$, (2.1) has a unique solution $x(\cdot)$ which is absolutely continuous with $\dot{x}(\cdot) \in L_2(t_0, t_1; R^n)$ for $0 \leqq t_1 < \infty$, and which satisfies (2.1) for almost all t.

It is also well known [4], [14], [15] that (2.1) can be posed as an evolution equation on the space $Z = R^n \times L_2(-r, 0; R^n)$, which is a Hilbert space with the inner product

$$\langle (x, \phi), (y, \psi) \rangle_Z = \langle x, y \rangle_{R^n} + \langle \phi, \psi \rangle_{L_2}. \tag{2.3}$$

Throughout this paper, we identify Z with its dual. Written as an evolution equation on Z, (2.1) becomes

$$\dot{z}(t) = \mathscr{A} z(t) + Bu(t), \qquad t \geqq t_0, \tag{2.4}$$

where $z(t) = (x(t), x_t)$ and $Bu(t) = (B_0 u(t), 0)$; the operator $\mathscr{A}$ is defined by $D(\mathscr{A}) = \{(x, \phi)\colon \phi \text{ is absolutely continuous, } \phi' \in L_2(-r, 0; R^n),\ x = \phi(0)\}$ and $\mathscr{A}(\phi(0), \phi) = (L\phi, \phi')$. This $\mathscr{A}$ is the infinitesimal generator of a strongly continuous semigroup $T(t) \in \mathscr{L}(Z, Z)$, $t \geqq 0$, and the integral version of (2.4) is

$$z(t) = T(t-s)z(s) + \int_s^t T(t-\eta)Bu(\eta)\, d\eta, \qquad t_0 \leqq s \leqq t. \tag{2.5}$$[1]

Some important properties of $\mathscr{A}$ and $T(\cdot)$ are (see [4], [12], [20]):

i) $\mathscr{A}$ has compact resolvent, and λ is an eigenvalue of $\mathscr{A}$ if and only if $\det \Delta(\lambda) = 0$, where $\Delta(\lambda) = \lambda I - L_0(\lambda)$ and $L_0(\lambda)$ is given by

$$L_0(\lambda) = \sum_{i=0}^{\nu} e^{-\lambda h_i} A_i + \int_{-r}^{0} e^{\lambda\theta} D(\theta)\, d\theta. \tag{2.6}$$

ii) For $t \geqq r$, $T(t)Z \subset D(\mathscr{A})$ and $T(t)$ is compact.

To obtain some of its most important results, this paper must deal more explicitly than previous papers have with $\mathscr{A}^*$ and $T^*(t)$, the adjoints of $\mathscr{A}$ and $T(t)$ respectively. We emphasize that the $\mathscr{A}^*$ used here is the Z-adjoint of $\mathscr{A}$, instead of the adjoint used in [12], where, through the imbedding $D(\mathscr{A}) = V \subset Z \subset V'$ (the topological dual of V), the authors obtain an $\tilde{\mathscr{A}}^* \in \mathscr{L}(V, V')$.

THEOREM 2.1. *$D(\mathscr{A}^*)$ consists of those $(y, \psi) \in Z$ for which $\psi' \in L_2(-r, 0; R^n)$ and ψ is absolutely continuous on $[-r, 0]$ except at the points $-h_1, \cdots, -h_{\nu-1}$, where*

$$\psi((-h_i)^+) - \psi((-h_i)^-) = A_i^T y, \qquad 1 \leqq i \leqq \nu - 1, \tag{2.7}$$

and

$$\psi(-r) = A_\nu^T y. \tag{2.8}$$

(If $\nu = 1$, ψ is absolutely continuous on $[-r, 0]$.) For $(y, \psi) \in D(\mathscr{A}^*)$,

$$\mathscr{A}^*(y, \psi) = (A_0^T y + \psi(0), D^T y - \psi'). \tag{2.9}$$

Proof. The theorem follows upon substituting into the definition of a Hilbert space adjoint operator, and integrating by parts in the appropriate places. (See also [9], [41].) □

$\mathscr{A}^*$ generates the strongly continuous semigroup $T^*(\cdot)$, which has properties similar to those of $T(\cdot)$. In particular:

THEOREM 2.2. *For $t \geqq r$, $T^*(t)Z \subset D(\mathscr{A}^*)$, and $T^*(t)$ is compact.*

[1] The vector-valued integrals in this paper are Bochner integrals (see [21]).

Proof. Let $t \geqq r$. Since $T(t)Z \subset D(\mathscr{A})$, $\mathscr{A}T(t) \in \mathscr{L}(Z, Z)$. For $x \in D(\mathscr{A})$ and $y \in Z$, $\langle T(t)\mathscr{A}x, y\rangle_Z = \langle \mathscr{A}T(t)x, y\rangle_Z = \langle \mathscr{A}x, T^*(t)y\rangle_Z = \langle x, (\mathscr{A}T(t))^*y\rangle_Z$. Thus $T^*(t)y \in D(\mathscr{A}^*)$ and $\mathscr{A}^*T^*(t)y = (\mathscr{A}T(t))^*y$ $(= T^*(t)\mathscr{A}^*y$ if $y \in D(\mathscr{A}^*))$. Since $T(t)$ is compact, so is $T^*(t)$. □

Now we consider the stability of the open-loop system, i.e., the stability of solutions to (2.1) and (2.4) for u equal to zero. The following theorem says that all notions of asymptotic stability for our homogeneous hereditary system are equivalent.

THEOREM 2.3. *For $(x(0), x_0) = z(0) \in Z$, let $x(\cdot)$ be the solution to (2.1) for $u(\cdot) = 0$, and let $z(\cdot)$ be the corresponding solution to (2.4) and (2.5). Then the following statements are equivalent:*

i) *For each $(x(0), x_0) \in Z$, $x(t) \to 0$ as $t \to \infty$.*

ii) *There exist positive constants M and α such that*

$$\|x(t)\|_{R^n} \leqq M e^{-\alpha t}\|(x(0), x_0)\|_Z, \qquad t \geqq 0, \quad (x(0), x_0) \in Z.$$

iii) *For each $(x(0), x_0) \in Z$, $\int_0^\infty \|x(t)\|_{R^n}^2 dt < \infty$.*

iv) *For each $z(0) \in Z$, $z(t) \to 0$ as $t \to \infty$.*

v) *There exist positive constants M' and α' such that*

$$\|z(t)\|_Z \leqq M' e^{-\alpha' t}\|z(0)\|_Z, \quad t \geqq 0, \qquad z(0) \in Z.$$

vi) *For each $z(0) \in Z$, $\int_0^\infty \|z(t)\|_Z^2 dt < \infty$.*

Proof. The theorem follows from the definitions of x_t and $z(t)$, and from the spectral properties of $\mathscr{A}$ and $T(t)$ (which is compact for $t \geqq r$). For details see Hale [20, Chapt. 7]. Although Hale constrains x_t to be in $C(-r, 0; R^n)$, the spectral arguments that establish the present theorem proceed precisely as Hale's arguments. □

DEFINITION 2.1. The homogeneous system (2.1), or equivalently (2.4) or (2.5), is *asymptotically stable* if statements i)–vi) of Theorem 2.3 hold.

Note that the homogeneous hereditary system is asymptotically stable if and only if the semigroup $T(\cdot)$ is uniformly exponentially stable; i.e., $\|T(t)\| \leqq M' e^{-\alpha' t}$, $t \geqq 0$, with M' and α' the positive constants in statement v) of Theorem 2.3.

DEFINITION 2.2. The system (2.1) (or (2.4) or (2.5)) is *stabilizable* if there is an operator $K \in \mathscr{L}(Z, R^m)$ such that $\mathscr{A} - BK$ generates a uniformly exponentially stable semigroup.

Note that this definition of stabilizability requires the existence of a real $m \times n$ matrix K_0 and a real square-integrable matrix function $K_1(\cdot)$ such that

$$K(x, \phi) = K_0 x + \int_{-r}^{0} K_1(\theta)\phi(\theta)\, d\theta, \qquad (x, \phi) \in Z. \tag{2.10}$$

Stabilizability was defined similarly in [12] and [26] except that $K(x, \phi)$ was allowed to include terms like $K_i\phi(\theta_i)$, where K_i is an $m \times n$ matrix and $-r \leqq \theta_i < 0$, for $1 \leqq i < \infty$. With such terms, K would not be bounded; however, we will see that modifying Definition 2.2 to allow such terms in K would yield a definition of stabilizability that would be equivalent to Definition 2.2. Also, we would arrive at an equivalent definition of stabilizability by requiring K_1 to be C^∞ on $[-r, 0]$ because $C^\infty[-r, 0]$ is dense in $L_2[r, 0]$ and, if $\mathscr{A} - BK$ generates a uniformly exponentially stable semigroup, so does $\mathscr{A} - B\tilde{K}$ for $\|K - \tilde{K}\|$ sufficiently small but nonzero. Definition 2.2 is the same as that used in [31] for stabilizability.

3. The optimal control problem on a finite interval. Let $Q_0 = Q_0^T \in \mathscr{L}(R^n, R^n)$, $G = G^* \in \mathscr{L}(Z, Z)$, and $R = R^T \in \mathscr{L}(R^m, R^m)$, with $Q \geqq 0$, $G \geqq 0$ and $R > 0$. The optimal control problem for this section is: given $-\infty < t_0 \leqq t_f < \infty$ and $(x(t_0), x_{t_0}) \in Z$,

choose the control $u \in L_2(t_0, t_f; R^m)$ to minimize the cost functional

$$J(t_0, (x(t_0), x_{t_0}), u) = \langle G(x(t_f), x_{t_f}), (x(t_f), x_{t_f})\rangle_Z + \int_{t_0}^{t_f} (\langle Q_0 x(t), x(t)\rangle_{R^n} + \langle Ru(t), u(t)\rangle_{R^m})\, dt, \tag{3.1}$$

where $x(\cdot)$ is the solution to (2.1) for the given initial conditions. With $z(\cdot)$ as in § 2, we have

$$J(t_0, (x(t_0), x_{t_0}), u) = J(t_0, z(t_0), u) = \langle Gz(t_f), z(t_f)\rangle_Z + \int_{t_0}^{t_f} (\langle Qz(t), z(t)\rangle_Z + \langle Ru(t), u(t)\rangle_{R^n})\, dt, \tag{3.2}$$

where $Q(x, \phi) = (Q_0 x, 0)$ for $(x, \phi) \in Z$.

As in [19], we denote $L_2(t, t_f; Z)$ by $\mathscr{Z}_t$ and $L_2(t, t_f; R^m)$ by $\mathscr{U}_t$, for $t \leqq t_f$, and define $T_t \in \mathscr{L}(Z, \mathscr{Z}_t)$, $\mathscr{T}_t \in \mathscr{L}(\mathscr{Z}_t, \mathscr{Z}_t)$ and $\mathscr{F}_t \in \mathscr{L}(\mathscr{Z}_t, Z)$ by

$$(T_t z)(s) = T(s-t)z, \qquad z \in Z, \tag{3.3}$$

$$(\mathscr{T}_t g)(s) = \int_t^s T(s-\eta) g(\eta)\, d\eta, \qquad g \in \mathscr{Z}_t, \tag{3.4}$$

$$\mathscr{F}_t g = (\mathscr{T}_t g)(t_f), \qquad g \in \mathscr{Z}_t. \tag{3.5}$$

Then (3.2) becomes

$$J(t_0, z(t_0), u) = \langle G(T(t_f - t_0)z(t_0) + \mathscr{F}_{t_0} Bu), (T(t_f - t_0)z(t_0) + \mathscr{F}_{t_0} Bu)\rangle_Z + \langle Q(T_{t_0} z(t_0) + \mathscr{T}_{t_0} Bu), (T_{t_0} z(t_0) + \mathscr{T}_{t_0} Bu)\rangle_{\mathscr{Z}_{t_0}} + \langle Ru, u\rangle_{\mathscr{U}_{t_0}}, \tag{3.6}$$

or

$$J(t_0, z(t_0), u) = \langle \tilde{R}_{t_0} u, u\rangle_{\mathscr{U}_{t_0}} + 2\langle \tilde{B}^*_{t_0} z(t_0), u\rangle_{\mathscr{U}_{t_0}} + \langle GT(t_f - t_0)z(t_0), T(t_f - t_0)z(t_0)\rangle_Z + \langle QT_{t_0} z(t_0), T_{t_0} z(t_0)\rangle_{\mathscr{Z}_{t_0}}, \tag{3.7}$$

where

$$\tilde{R}_{t_0} = (R + B^* \mathscr{T}^*_{t_0} Q \mathscr{T}_{t_0} B + B^* \mathscr{F}^*_{t_0} G \mathscr{F}_{t_0} B) \in \mathscr{L}(\mathscr{U}_{t_0}, \mathscr{U}_{t_0}) \tag{3.8}$$

and

$$\tilde{B}^*_{t_0} = (B^* \mathscr{T}^*_{t_0} Q T_{t_0} + B^* \mathscr{F}^*_{t_0} G T(t_f - t_0)) \in \mathscr{L}(Z, \mathscr{U}_{t_0}) \tag{3.9}$$

with $\mathscr{T}^*_{t_0}$ and $\mathscr{F}^*_{t_0}$ given by

$$(\mathscr{T}^*_{t_0} g)(t) = \int_t^{t_f} T^*(\eta - t) g(\eta)\, d\eta \tag{3.10}$$

and

$$(\mathscr{F}^*_{t_0} z)(t) = T^*(t_f - t)z. \tag{3.11}$$

Since R is positive definite, so is $\tilde{R}_{t_0}$, and the unique optimal control is given by

$$u(t) = -(\tilde{R}^{-1}_{t_0} \tilde{B}^*_{t_0})(t) z(t_0) \equiv -(\tilde{R}^{-1}_{t_0} \tilde{B}^*_{t_0} z(t_0))(t) \quad \text{a.e. in } [t_0, t_f]. \tag{3.12}$$

Our hypotheses on the operators involved in (3.8) and (3.9) justify (3.12), where $\tilde{R}^{-1}_{t_0} \tilde{B}^*_{t_0} \in \mathscr{B}_\infty(t_0, t_f; Z, R^n)$.[2]

[2] For Banach spaces H and U, $\mathscr{B}_\infty(t_0, t_f; H, U)$ is the Banach space (see [19, Appendix A]) of essentially bounded, strongly measurable functions from (t_0, t_f) to $\mathscr{L}(H, U)$. An operator-valued function $B(\cdot)$: $(t_0, t_f) \to \mathscr{L}(H, U)$ is called strongly measurable if $B(\cdot)\chi$ is strongly measurable (see [21]) for each x in H.

We have given (3.3)–(3.12) to prepare for the approximation theory of § 6. Although (3.12) was central to the development of the feedback control law and the Riccati integral equations in [19], we will skip the derivations here and summarize the results that we need for the present optimal control problem.

From [19], we know that there exists a unique optimal (i.e., minimizing) control u, which can be given in the feedback form

$$u(t) = -R^{-1}B^*\Pi(t)z(t), \qquad t_0 \leqq t \leqq t_f, \tag{3.13}$$

where $\Pi(\cdot)$ satisfies an infinite dimensional Riccati integral equation. According to [19, § 3], the correct $\Pi(\cdot)$ for (3.13) is the unique element of $\mathcal{B}_\infty(t_0, t_f; Z, Z)$ which satisfies the "first Riccati integral equation"

$$\begin{aligned}\Pi(t)z &= T^*(t_f - t)GT(t_f - t)z \\ &\quad + \int_t^{t_f} T^*(\eta - t)[Q - \Pi(\eta)BR^{-1}B^*\Pi(\eta)]T(\eta - t)z\, d\eta, \qquad t_0 \leqq t \leqq t_f, \quad z \in Z,\end{aligned} \tag{3.14}$$

and $\Pi(t)$ is strongly continuous in t. Furthermore, the minimum value of $J(t_0, z(t_0), \cdot)$ is

$$\min_{v \in L_2(t_0, t_f; R^n)} J(t_0, z(t_0), v) = \langle \Pi(t_0)z(t_0), z(t_0)\rangle_Z, \qquad z(t_0) \in Z. \tag{3.15}$$

With the u of (3.13), (2.4) becomes

$$\dot{z}(t) = (\mathcal{A} - BR^{-1}B^*\Pi(t))z(t), \qquad 0 \leqq t \leqq t_f, \tag{3.16}$$

which is not autonomous. For the "fundamental solution" of (3.16), we need the following.

DEFINITION 3.1. Let $-\infty < t_0 < t_f < \infty$, and let Z be a Banach space. $T(\cdot, \cdot): \{(t, s): t_0 \leqq s \leqq t \leqq t_f\} \to \mathcal{L}(Z, Z)$ is an *evolution operator* if

i) $T(t, r)T(r, s) = T(t, s)$, $t_0 \leqq s \leqq r \leqq t \leqq t_f$;

ii) $T(t, t) = I$;

iii) $T(t, s)$ is strongly continuous in s on $[t_0, t]$ and strongly continuous in t on $[s, t_f]$.

For the evolution operators used in this paper, it is not difficult to show that there exists a constant M_1, depending on t_0 and t_f, such that

$$\|T(t, s)\| \leqq M_1, \qquad t_0 \leqq s \leqq t \leqq t_f. \tag{3.17}$$

Clearly, we have this bound when $T(t, r) = T(t - r)$ is a C_0-semigroup. The existence of the bound for our subsequent evolution operator $S(t, s)$ will follow from (3.20) below. We should note that when $T(t, s)$ is the evolution operator for a homogeneous hereditary system with time-varying coefficients, usually we still have (3.17) (see [35]).

For the proof of the following theorem, see Curtain and Pritchard [7].

THEOREM 3.1. *Let $T(\cdot, \cdot)$ be an evolution operator which is uniformly bounded as in* (3.17), *and let C be in $\mathcal{B}_\infty(t_0, t_f; Z, Z)$. Then the operator integral equation*

$$S(t, s)z = T(t, s)z + \int_s^t T(t, \eta)C(\eta)S(\eta, s)z\, d\eta, \qquad z \in Z, \tag{3.18}$$

has a unique solution $S(\cdot, \cdot)$ in the class of strongly continuous (*as in Definition* 3.1) *bounded linear operators on Z. $S(\cdot, \cdot)$ is an evolution operator, and it is called the perturbed evolution operator corresponding to the perturbation of $T(\cdot, \cdot)$ by C. $S(\cdot, \cdot)$ is*

also the unique solution of

$$S(t,s)z = T(t,s)z + \int_s^t S(t,\eta)C(\eta)T(\eta,s)z\,d\eta; \tag{3.19}$$

i.e., $T(\cdot,\cdot)$ is the perturbed evolution operator corresponding to the perturbation of $S(\cdot,\cdot)$ by $-C$. If M_1 is the uniform bound of (3.17), *we have*

$$\|S(t,s)\| \leqq M_1 \exp\left(M_1\|C\|_{\mathcal{B}_\infty}(t-s)\right). \tag{3.20}$$

Of course, a strongly continuous semigroup on Z is an evolution operator on Z. Henceforth, $S(\cdot,\cdot)$ will denote the perturbed evolution operator corresponding to the perturbation of the semigroup $T(\cdot)$ of (2.5) by $-BR^{-1}B^*\Pi$, where Π is the nonnegative, self-adjoint solution of (3.14); i.e.,

$$S(t,s)z = T(t-s)z - \int_s^t T(t-\eta)BR^{-1}B^*\Pi(\eta)S(\eta,s)z\,d\eta, \tag{3.21}$$

$$t_0 \leqq s \leqq t \leqq t_f, \quad z \in Z.$$

The optimal response of our control system is then

$$z(t) = S(t,s)z(s), \qquad t_0 \leqq s \leqq t \leqq t_f, \tag{3.22}$$

where, again, $S(\cdot,\cdot)$ is the solution of (3.21). Note that formal differentiation of (3.18) and (3.19) with respect to t yields (3.16).

We have two more important integral equations for $\Pi(\cdot)$:

$$\Pi(t)z = T^*(t_f-t)GS(t_f,t)z + \int_t^{t_f} T^*(\eta-t)QS(\eta,t)z\,d\eta, \qquad t_0 \leqq t \leqq t_f, \quad z \in Z, \tag{3.23}$$

and the "second Riccati integral equation"

$$\Pi(t)z = S^*(t_f,t)GS(t_f,t)z$$

$$+ \int_t^{t_f} S^*(\eta,t)[Q + \Pi(\eta)BR^{-1}B^*\Pi(\eta)]S(\eta,t)z\,d\eta, \qquad t_0 \leqq t \leqq t_f, \quad z \in Z. \tag{3.24}$$

We can use (3.21) to go from one to another of the integral equations (3.14), (3.23) and (3.24). For details, see [19, § 3].

The following theorem gives the information we need about the present optimal control problem and the integral equations for $\Pi(\cdot)$ (see [19, § 3]).

THEOREM 3.2. *For $s \leqq t_f < \infty$ and $z(s) \in Z$, the unique control $u \in L_2(s, t_f; R^m)$ which minimizes the cost functional $J(s, z(s), \cdot)$ of* (3.1) *and* (3.2) *is the linear feedback control of* (3.13), *where $\Pi(\cdot)$ is the unique element of $\mathcal{B}_\infty(s, t_f; Z, Z)$ which satisfies the first Riccati integral* (3.14). *When solutions of the second Riccati integral* (3.24) *are restricted to be self-adjoint elements of $\mathcal{B}_\infty(s, t_f; Z, Z)$, $\Pi(\cdot)$ and $S(\cdot,\cdot)$ are the unique solution of the system of equations* (3.21) *and* (3.24).

We can write $\Pi(t)$ as a matrix of operators:

$$\Pi(t) = \begin{bmatrix} \Pi^{00}(t) & \Pi^{01}(t) \\ \Pi^{10}(t) & \Pi^{11}(t) \end{bmatrix}, \qquad t \leqq t_f. \tag{3.25}$$

Since $\Pi(t)$ is a nonnegative self-adjoint element of $\mathcal{L}(Z,Z)$, $\Pi^{00}(t)$ is a real nonnegative, symmetric $n \times n$ matrix, $\Pi^{10}(t)$ is a real square-integrable $n \times n$ matrix function $\Pi^{10}(t,\cdot)$ on $[-r, 0]$, $\Pi^{01}(t) = \Pi^{10}(t)^*$ and

$$\Pi^{01}(t)\phi = \int_{-r}^0 \Pi^{10}(t,\theta)^T\phi(\theta)\,d\theta, \qquad \phi \in L_2(-r, 0; R^n) \tag{3.26}$$

and $\Pi^{11}(t)$ is a nonnegative self-adjoint operator on $L_2(-r, 0; R^n)$.

We will see in § 4 that $\Pi^{11}(t)$ is of the trace class and therefore an integral operator, but we shall not pursue a characterization of its kernel or derive a set of coupled ordinary and partial differential equations for $\Pi^{00}(t)$, $\Pi^{10}(t)$ and $\Pi^{11}(t)$, as some previous authors have done (see [1], [10], [11], [26], [38]). Such equations, while they enhance our understanding of the feedback control law, usually can be solved only by numerical approximations which are equivalent to numerical approximations for the solution of the Riccati integral equations, with which this paper will deal extensively. Also, Delfour and Mitter [13] derived the second Riccati integral equation (3.24), and differentiated this equation in a weak, or distributional, sense to obtain an infinite dimensional Riccati differential equation, which we will not pursue here.

From (3.13) and (3.25) we see that the optimal control is given by

$$u(t) = -R^{-1}B_0^T(\Pi^{00}(t)x(t) + \int_{-r}^{0} \Pi^{10}(t,\theta)^T x_t(\theta)\,d\theta), \qquad t \leqq t_f. \tag{3.27}$$

We can deduce some important properties of $\Pi^{00}(t)$ and $\Pi^{10}(t,\cdot)$ from (3.23), Theorems 2.1 and 2.2, and the following lemma.

LEMMA 3.1. *Let $\mathcal{A}$ generate a strongly continuous semigroup $T(\cdot)$ on a Hilbert space Z, $0 \leqq t_1 < \infty$, $g \in L_2(0, t_1; H)$, and $f(s) = f(0) + \int_0^s g(\eta)\,d\eta$ for $0 \leqq s \leqq t_1$. Then $\int_s^t T(s)f(s)\,ds \in D(\mathcal{A})$ for $0 \leqq t \leqq t_1$.*

Proof. By using Bochner integrals instead of Riemann integrals, we generalize a portion of an argument given in [22] for the case where f is continuously differentiable. We have

$$\begin{aligned} \int_0^t T(s)f(s)\,ds &= \int_0^t T(s)\left[f(0) + \int_0^s g(\tau)\,d\tau\right] ds \\ &= \int_0^t T(s)f(0)\,ds + \int_0^t \left[\int_\tau^t T(s)g(\tau)\,ds\right] d\tau. \end{aligned} \tag{3.28}$$

It is a standard result that $\int_0^t T(s)z\,ds \in D(\mathcal{A})$ for $z \in Z$. Also (see [22, pp. 488]),

$$\mathcal{A}\int_\tau^t T(s)g(\tau)\,ds = (T(t) - T(\tau))g(\tau). \tag{3.29}$$

($g(\tau)$ is defined for almost all $\tau \in (0, t_1)$.) Since $\mathcal{A}$ is closed, (3.29) and [21, Thm. 3.7.12, p. 83] imply that $\int_0^t [\int_\tau^t T(s)g(\tau)\,ds]\,d\tau \in D(\mathcal{A})$ and

$$\mathcal{A}\int_0^t \left[\int_\tau^t T(s)g(\tau)\,ds\right] d\tau = \int_0^t (T(t) - T(\tau))g(\tau)\,d\tau. \tag{3.30}$$

Hence the lemma. □

THEOREM 3.3. *Let $\Pi(\cdot)$ be the solution of the Riccati integral equations* (3.14) *and* (3.24). *Then $\Pi(t)Z \subset D(\mathcal{A}^*)$ for $t \leqq t_f - r$; if $G = 0$, $\Pi(t)Z \subset D(\mathcal{A}^*)$ for $t \leqq t_f$.*

Proof. For initial time t, initial state $z(t)$ and $\eta \geqq t$, $S(\eta, t)z(t)$ is the solution to (2.5) for the control of (3.13). (η replaces the t in (2.5) and (3.13).) Thus, $QS(\eta, t) = (Q_0 x(\eta), 0)$, and the derivative of $x(\eta)$ is given by (2.1). Hence Lemma (3.1) shows that the integral in (3.23) is in $D(\mathcal{A}^*)$ for $z \in Z$ and $t \leqq t_f$. By Theorem 2.2, $T^*(t_f - t)GS(t_f, t)z$, the other term on the right side of (3.23), is in $D(\mathcal{A}^*)$ for $z \in Z$ and $t \leqq t_f - r$. □

Now, if $\Pi(t)$ maps all of Z into $D(\mathcal{A}^*)$, then $(\Pi^{00}(t)x, \Pi^{10}(t)x) \in D(\mathcal{A}^*)$ for all $x \in R^n$. Therefore, from Theorems 2.1 and 3.3, we have

THEOREM 3.4. *For $t \leqq t_f - r$, or, if $G = 0$, for $t \leqq t_f$, $\Pi^{10}(t, \cdot)$ is absolutely continuous except at the points $-h_1, \cdots, -h_{\nu-1}$, where*

$$\Pi^{10}(t, (-h_i)^+) - \Pi^{10}(t, (-h_i)^-) = A_i^T \Pi^{00}(t), \qquad 1 \leqq i \leqq \nu - 1. \tag{3.31}$$

Also,

$$\Pi^{10}(t, -r) = A_\nu^T \Pi^{00}(t). \tag{3.32}$$

Note that Theorem 3.4 agrees with results in [10] and [13].

4. Optimal control on the infinite interval. Now we consider the optimal control problem for $t_f = \infty$ and $G = 0$. Since the operators L, B, Q and R are constant, we take the initial time to be zero. For an initial state z, a control u and the corresponding $z(t)$ given by (2.5) with $s = 0$, the cost functional is

$$J(z, u) = \int_0^\infty (\langle Qz(t), z(t)\rangle_Z + \langle Ru(t), u(t)\rangle_{R^m})\, dt. \tag{4.1}$$

DEFINITION 4.1. A function $u: (0, \infty) \to R^m$ is an *admissible control for the initial state z*, or simply an *admissible control for z*, if u is strongly measurable on $(0, \infty)$ and $J(z, u)$ is finite.

For this section we will assume the following.

Hypothesis 4.1. The operators Q_0 and L are such that, if u is an admissible control for z, then $z(t) \to 0$ as $t \to \infty$; i.e., any admissible control drives the state to zero asymptotically.

This is certainly the case if Q_0 is positive definite. Hypothesis 4.1 holds also if (2.1) with output $Q_0 x$ is observable according to definitions in [3] and [13].

DEFINITION 4.2. Let $\mathcal{A}$ be as defined in § 2. An operator $\Pi \in \mathcal{L}(Z, Z)$ is a *solution of the Riccati algebraic equation* if Π maps $D(\mathcal{A})$ into $D(\mathcal{A}^*)$, $\mathcal{A}^*\Pi + \Pi\mathcal{A}$ has a bounded extension to all of Z, and Π satisfies the Riccati algebraic equation

$$\mathcal{A}^*\Pi + \Pi\mathcal{A} - \Pi B R^{-1} B^* \Pi + Q = 0. \tag{4.2}$$

For the results summarized in Theorems 4.1 and 4.2, see [19, § 4].

THEOREM 4.1. *Let $\mathcal{A}$, B, Q and R be as previously defined. There exists a nonnegative, self-adjoint solution of the Riccati algebraic equation if and only if, for each $z \in Z$, there is an admissible control for z. Under the hypothesis that any admissible control drives the state to zero, there exists at most one nonnegative, self-adjoint solution of the Riccati algebraic equation. Suppose that such a solution Π exists. Then for $z \in Z$, the unique control in $L_2(0, \infty; R^m)$ which minimizes $J(z, \cdot)$ can be given in the feedback form*

$$u(t) = -R^{-1} B^* \Pi z(t), \qquad 0 \leqq t < \infty, \tag{4.3}$$

and

$$\min_{v \text{ admissible}} J(z, v) = \langle \Pi z, z\rangle_Z, \qquad z \in Z. \tag{4.4}$$

The optimal trajectory is given by

$$z(t) = S(t - s) z(s), \qquad 0 \leqq s < t < \infty, \tag{4.5}$$

where $S(\cdot)$ is the strongly continuous semigroup generated by $\mathcal{A} - BR^{-1}B^*\Pi$, and $S(\cdot)$ is uniformly exponentially stable.

COROLLARY 4.1. *The system* (2.1) (*see Definition* 2.3) *is stabilizable if and only if, for each $z \in Z$, there is an admissible control.*

THEOREM 4.2. *If* Π *is a nonnegative, self-adjoint solution of the Riccati algebraic equation, and* $S(\cdot)$ *is the strongly continuous semigroup generated by* $\mathcal{A}-BR^{-1}B^*\Pi$, *then* Π *and* $S(\cdot)$ *satisfy*

$$\Pi z = S^*(t)\Pi S(t)z + \int_0^t S^*(\eta)[Q+\Pi BR^{-1}B^*\Pi]S(\eta)z\,d\tau, \qquad 0\leqq t<\infty, \quad z\in Z. \tag{4.6}$$

Conversely, suppose that $\Pi\in\mathcal{L}(Z,Z)$, $\Pi=\Pi^*\geqq 0$, $S(\cdot)$ *is the strongly continuous semigroup generated by* $\mathcal{A}-BR^{-1}B^*\Pi$, *and* Π *and* $S(\cdot)$ *satisfy* (4.6). *Then* Π *satisfies—uniquely—the Riccati algebraic equation.*

COROLLARY 4.2. *Let* $\Pi\in\mathcal{L}(Z,Z)$ *be nonnegative and self-adjoint and suppose that* $S(\cdot)$, *the semigroup generated by* $\mathcal{A}-BR^{-1}B^*\Pi$, *is uniformly exponentially stable. Then* Π *and* $S(\cdot)$ *satisfy* (4.6) *if and only if* Π *and* $S(\cdot)$ *satisfy*

$$\Pi z = \int_0^\infty S^*(\eta)[Q+\Pi BR^{-1}B^*\Pi]S(\eta)z\,d\eta, \qquad z\in Z, \tag{4.7}$$

if and only if Π *satisfies the Riccati algebraic equation.*

Proof. The "only if" of the first "if and only if" follows when $t\to\infty$ in (4.6). Conversely, replace z in (4.7) by $S(t)z$ and apply $S^*(t)$ to both sides of the resulting equation; changing the variable of integration then yields (4.6). The second "if and only if" follows from Theorem 4.2. □

The following theorem, along with our previous theorems, says that the nonnegative, self-adjoint solution of the Riccati algebraic equation is the steady-state solution of the Riccati integral equation of § 3, and that it is stable with respect to nonnegative initial (actually, final) conditions.

THEOREM 4.3. *Let* $\Pi(\cdot)$ *be the solution of the Riccati integral equations for the problem of* § 3 *with* $G\geqq 0$, *and* $t_f=0$, *and suppose that there exists a nonnegative self-adjoint solution* Π *of the Riccati algebraic equation. Then*

$$\lim_{s\to-\infty}\Pi(s)z=\Pi z, \qquad z\in Z. \tag{4.8}$$

If $G=0$,

$$\Pi\geqq\Pi(s)\geqq\Pi(t), \qquad -\infty<s\leqq t\leqq 0. \tag{4.9}$$

If $G\geqq\Pi$ *and* $\|S(\eta)\|\leqq Me^{-\alpha\eta}$ *for positive constants* M *and* α, *then*

$$\Pi\leqq\Pi(s)\leqq\Pi+Me^{2\alpha s}\|G\|, \qquad -\infty<s\leqq 0. \tag{4.10}$$

This theorem follows from [19, Thm. 4.10]. The uniform exponential convergence of $\Pi(s)$ to Π is essential to our most important results for approximation of Π (see Theorem 6.9).

As with the $\Pi(\cdot)$ of § 3, we can represent the Π of this section by the matrix of operators:

$$\Pi=\begin{bmatrix}\Pi^{00} & \Pi^{01}\\ \Pi^{10} & \Pi^{11}\end{bmatrix}, \tag{4.11}$$

where Π^{00} is a nonnegative, symmetric $n\times n$ matrix, Π^{10} is a square-integrable matrix function $\Pi^{10}(\cdot)$ on $[-r,0]$, $\Pi^{01}=\Pi^{10*}$ and

$$\Pi^{01}\phi=\int_{-r}^0\Pi^{10}(\theta)^T\phi(\theta)\,d\theta, \qquad \phi\in L_2(-r,0;R^n), \tag{4.12}$$

and Π^{11} is a nonnegative, self-adjoint operator on $L_2(-r, 0; R^n)$. Again, we will not pursue the characterization of and the relations among the operators Π^{00}, Π^{10} and Π^{11} to the extent to which Delfour, McCalla and Mitter have in [12], but we do have

THEOREM 4.4. *If Π is the nonnegative, self-adjoint solution of the Riccati algebraic equation, then $\Pi Z \subset D(\mathscr{A}^*)$.*

Proof. The proof is quite similar to the proof of Theorem 3.3. Instead of (3.23), we use (4.6) with $t \geqq r$, and recall that, since the generator of $S^*(\cdot)$ results from a bounded perturbation of $\mathscr{A}^*$, the domain of the generator of $S^*(\cdot)$ is $D(\mathscr{A}^*)$. □

THEOREM 4.5. *$\Pi^{10}(\cdot)$ is absolutely continuous on $[-r, 0]$ except at the points $-h_1, \cdots, -h_{\nu-1}$, where*

$$\Pi^{10}((-h_j)^+) - \Pi^{10}((-h_i)^-) = A_i^T \Pi^{00}, \qquad 1 \leqq i \leqq \nu - 1. \tag{4.13}$$

Also,

$$\Pi^{10}(-r) = A_\nu^T \Pi^{00}. \tag{4.14}$$

Proof. The theorem follows from Theorem 2.1 and Theorem 4.4. □

An adjoint state similar to the adjoint state employed in finite dimensional optimal control theory will facilitate the remaining work of this section. If Π is the solution of the Riccati algebraic equation and $z(\cdot)$ is the optimal trajectory for the problem of this section, define the adjoint state $p(\cdot)$ by

$$p(t) = \Pi z(t), \qquad t \geqq 0. \tag{4.15}$$

The optimal control of (4.3) then becomes

$$u(t) = -R^{-1} B^* p(t). \tag{4.16}$$

If $z(0) \in D(\mathscr{A})$, $z(t)$ and $p(t)$ are continuously differentiable for $t \geqq 0$, and

$$\dot{p}(t) = \Pi \dot{z}(t) = \Pi(\mathscr{A} - BR^{-1}B^*\Pi) z(t), \qquad t \geqq 0. \tag{4.17}$$

Then with (4.2) and (4.15), (4.17) yields

$$\dot{p}(t) = -\mathscr{A}^* p(t) - Q z(t), \qquad t \geqq 0. \tag{4.18}$$

We thus have

$$\begin{pmatrix} \dot{z}(t) \\ \dot{p}(t) \end{pmatrix} = \hat{\mathscr{A}} \begin{pmatrix} z(t) \\ p(t) \end{pmatrix}, \qquad t \geqq 0, \tag{4.19}$$

where

$$\hat{\mathscr{A}} = \begin{pmatrix} \mathscr{A} & -C \\ -Q & -\mathscr{A}^* \end{pmatrix}, \qquad D(\hat{\mathscr{A}}) = D(\mathscr{A}) \times D(\mathscr{A}^*) \tag{4.20}$$

and $C = BR^{-1}B^*$.

Note that, as an operator on $Z \times Z$, $\hat{\mathscr{A}}$ is closed and densely defined. However, in spite of (4.19), which holds if $z(0) \in D(\mathscr{A})$ or if $t > r$, neither $\hat{\mathscr{A}}$ nor $-\hat{\mathscr{A}}$ generates a C_0-semigroup on $Z \times Z$. For, since C and Q are bounded, if $\hat{\mathscr{A}}$ generated a C_0-semigroup, so would the operator given by

$$\hat{\tilde{\mathscr{A}}} = \begin{bmatrix} \mathscr{A} & 0 \\ 0 & -\mathscr{A}^* \end{bmatrix}, \qquad D(\hat{\tilde{\mathscr{A}}}) = D(\hat{\mathscr{A}}). \tag{4.21}$$

Hence $-\mathscr{A}^*$ would generate a C_0-semigroup on Z, and therefore $-\mathscr{A}$ would generate a C_0-semigroup on Z. But we know that, for $r > 0$, $-\mathscr{A}$ does not generate a C_0-semigroup on Z.

We will now derive a number of results concerning the eigenvalues of $\hat{\mathscr{A}}$ which will parallel the results for the finite dimensional linear regulator problem (see [2], [25]). To this end, we will denote the complex version of Z by Z_c; i.e., $Z_c = \{z = v + iw: v, w \in Z\}$. Z_c is a complex Hilbert space with

$$\langle z_1, z_2\rangle_{Z_c} = \langle z_1, \bar{z}_2\rangle_Z. \tag{4.22}$$

We extend all linear operators on Z to Z_c in the usual way, and we will not distinguish in notation between the real and complex domains of an operator, since the meaning will be clear in every instance.

THEOREM 4.6. *$\hat{\mathscr{A}}$ has compact resolvent.*[3] *If λ is an eigenvalue of $\hat{\mathscr{A}}$, so is $-\lambda$, and the algebraic*[4] *and geometric multiplicities of λ are, respectively, equal to the algebraic and geometric multiplicities of $-\lambda$.*

Proof. Since $\mathscr{A}$ has compact resolvent, so does $\mathscr{A}^*$ and the $\hat{\mathscr{A}}$ of (4.21). Then, since $\hat{\mathscr{A}}$ results from a bounded perturbation of $\hat{\mathscr{A}}$, $\hat{\mathscr{A}}$ has compact resolvent.

Let λ be an eigenvalue of $\hat{\mathscr{A}}$. For $(z_1, p_1) \in D(\hat{\mathscr{A}})$, it is easy to see that $(\hat{\mathscr{A}} - \lambda)(z_1, p_1) = (z_0, p_0) \Rightarrow (\hat{\mathscr{A}}^* + \lambda)(-p_1, z_1) = (p_0, -z_0)$, which implies a one-to-one correspondence between generalized eigenvectors corresponding to $-\lambda$ as an eigenvalue of $\hat{\mathscr{A}}$ and generalized eigenvectors corresponding to λ as an eigenvector of $\hat{\mathscr{A}}^*$. Thus the spectral characteristics of $-\lambda$ as an eigenvalue of $\hat{\mathscr{A}}^*$ are identical to those of λ as an eigenvalue of $\hat{\mathscr{A}}$. But (see [22, Remark 6.23, p. 184]), since the multiplicities are finite, the spectral characteristics of $-\bar{\lambda}$ as an eigenvalue of $\hat{\mathscr{A}}$ are identical to those of $-\lambda$ as an eigenvalue of $\hat{\mathscr{A}}^*$. Since $\hat{\mathscr{A}}$ is real, the theorem follows. □

Now define a $2n \times 2n$ matrix function of λ by

$$\hat{\Delta}(\lambda) = \lambda I - \begin{bmatrix} L_0(\lambda) & -C_0 \\ -Q_0 & -L_0^T(-\lambda) \end{bmatrix}, \tag{4.23}$$

where $L_0(\lambda)$ is given by (2.6) and $C_0 = B_0 R^{-1} B_0^T$.

THEOREM 4.7. *λ is an eigenvalue of $\hat{\mathscr{A}}$ if and only if*

$$\det \hat{\Delta}(\lambda) = 0. \tag{4.24}$$

Proof. λ is an eigenvalue of $\hat{\mathscr{A}}$ if and only if there exist a nonzero $((x, \phi), (y, \psi)) \in D(\mathscr{A}) \times D(\mathscr{A}^*)$ such that $\hat{\mathscr{A}}((x, \phi), (y, \psi)) = \lambda((x, \phi), (y, \psi))$. This is equivalent to saying that there exists a nontrivial solution to the set of equations

$$L\phi - C_0 y = \lambda x, \tag{4.25a}$$

$$\phi' = \lambda\phi, \tag{4.25b}$$

$$-Q_0 x - A_0^T y - \psi(0) = \lambda y, \tag{4.25c}$$

$$-D^T y + \psi' = \lambda\psi, \tag{4.25d}$$

subject to the conditions $\phi(0) = x$ and (2.7) and (2.8). From here, it is a straightforward exercise to write down $\phi(s)$ and $\psi(s)$ and show that (4.25a)–(4.25d) have a nontrivial solution subject to the required boundary and jump conditions if and only if there exists a nonzero pair (x, y) of complex n-vectors such that

$$L_0(\lambda)x - C_0 y = \lambda x \tag{4.26a}$$

---

[3] Recall (see [22]) that this means that the spectrum of $\hat{\mathscr{A}}$ consists entirely of isolated eigenvalues with finite multiplicities, and, if λ is not an eigenvalue of $\hat{\mathscr{A}}$, $(\lambda - \hat{\mathscr{A}})^{-1}$ is compact.

[4] As usual, the algebraic multiplicity of an eigenvalue is the dimension of the corresponding generalized eigenspace.

and

$$-Q_0x - L_0^T(-\lambda)y = \lambda y. \tag{4.26b}$$

□

We know (see [20], [29]) that the algebraic multiplicity of an eigenvalue λ of $\mathcal{A}$ is equal to the multiplicity of λ as a root of $\det \Delta(\lambda) = 0$, and it appears that the corresponding statement holds for λ as an eigenvalue of $\hat{\mathcal{A}}$ and a root of (4.24). However, we will not pursue a proof here.

For the rest of this section, we will assume:

Hypothesis 4.2. The system (2.1) is stabilizable.

Also, keep in mind Hypothesis 4.1.

THEOREM 4.8. *$\hat{\mathcal{A}}$ has no purely imaginary eigenvalues.*

Proof. Suppose $\hat{\mathcal{A}}(z, p) = i\omega(z, p)$ for some real ω and $(z, p) \in D(\hat{\mathcal{A}})$. Then

$$\mathcal{A}z - Cp = i\omega z \tag{4.27}$$

and

$$-Qz - \mathcal{A}^*p = i\omega p. \tag{4.28}$$

After taking the Z_C inner product of each term in (4.27) with $\bar{p}$ and the Z_C inner product of z with the complex conjugate of each term in (4.28), we add the two resulting equations to obtain

$$-\langle Qz, \bar{z}\rangle_{Z_C} - \langle Cp, \bar{p}\rangle_{Z_C} = 0. \tag{4.29}$$

Since Q and C are nonnegative and self-adjoint, we must have $Qz = Cp = 0$.

Suppose $z \neq 0$. Since $\mathcal{A}z = i\omega z$, $\operatorname{Re} z$ cannot be zero. Then $z(t) = e^{i\omega t}z + e^{-i\omega t}\bar{z}$ is a real, nontrivial solution of (2.4) for $u(t) = 0$, and $Qz(t) = 0$, $t \geqq 0$. Thus, for the initial condition $z + \bar{z}$, $u = 0$ is an admissible control which does not drive the state to zero asymptotically, contradicting our hypothesis on Q and L.

Now suppose $z = 0$ and $p \neq 0$. Then $(\mathcal{A}^* - \Pi BR^{-1}B^*)p = \mathcal{A}^*p = i\omega p$, so that $i\omega$ is an eigenvalue of both $\mathcal{A}^* - \Pi BR^{-1}B^*$ and its adjoint $\mathcal{A} - BR^{-1}B^*\Pi$. But this is impossible because, by Hypothesis 4.2, Theorem 4.1 and Corollary 4.1, $\mathcal{A} - BR^{-1}B^*\Pi$ generates the uniformly exponentially stable semigroup $S(\cdot)$. □

THEOREM 4.9. *A complex number λ with $\operatorname{Re}\lambda < 0$ is an eigenvalue of $\mathcal{A} - BR^{-1}B^*\Pi$ if and only if λ is an eigenvalue of $\hat{\mathcal{A}}$. The algebraic and geometric multiplicities of λ as an eigenvalue of $\mathcal{A} - BR^{-1}B^*\Pi$ are finite and are identical to the respective multiplicities of λ as an eigenvalue of $\hat{\mathcal{A}}$.*

Proof. Suppose λ is an eigenvalue of $\hat{\mathcal{A}}$ and $\operatorname{Re}\lambda < 0$. Then we have a nonzero pair (z, p) such that $(\operatorname{Re} z, \operatorname{Re} p) \in D(\hat{\mathcal{A}})$ and $(\operatorname{Im} z, \operatorname{Im} p) \in D(\hat{\mathcal{A}})$ and $\hat{\mathcal{A}}(z, p) = \lambda(z, p)$. Define $(z(t), p(t)) = \frac{1}{2}(e^{\lambda t}(z, p) + e^{\bar{\lambda} t}(\bar{z}, \bar{p})) = \operatorname{Re}(e^{\lambda t}(z, p))$ and $u(t) = -R^{-1}B^*p(t)$. Then $(z(t), p(t))$ satisfies (4.19), and u is an admissible control for $z(0) = \operatorname{Re} z$. Suppose that $v \in L_2(0, \infty; R^n)$ and $u + v$ is an admissible control for $z(0)$. The solution of (2.5) for the initial condition $z(0)$ and the control $u + v$ is $z(t) + y(t)$, where $y(t)$ is the solution of (2.5) for the initial condition 0 and the control v. Note that v is an admissible control for the initial condition 0. We have

$$\begin{aligned} J(z(0), u+v) - J(z(0), u) = &\int_0^\infty (\langle Qy(t), y(t)\rangle_Z + \langle Rv(t), v(t)\rangle_{R^m})\, dt \\ &+ 2\int_0^\infty (\langle Qz(t), y(t)\rangle_Z + \langle Ru(t), v(t)\rangle_{R^m})\, dt. \end{aligned} \tag{4.30}$$

Also, for $\bar{t} < \infty$,

$$\begin{aligned}\int_0^{\bar{t}} \langle Qz(t), y(t)\rangle_Z \, dt &= \int_0^{\bar{t}} \int_0^t \langle B^*T^*(t-s)Qz(t), v(s)\rangle_{R^m} \, ds \, dt \\ &= \int_0^{\bar{t}} \left\langle B^* \int_t^{\bar{t}} T^*(s-t)Qz(s) \, ds, v(t) \right\rangle_{R^m} dt \quad \text{(Fubini)}.\end{aligned} \tag{4.31}$$

From (4.18), we have

$$p(t) = T^*(\bar{t}-t)p(\bar{t}) + \int_t^{\bar{t}} T^*(s-t)Qz(s) \, ds, \tag{4.32}$$

so that (4.31) becomes

$$\begin{aligned}\int_0^{\bar{t}} \langle Qz(t), y(t)\rangle_Z \, dt &= \int_0^{\bar{t}} \langle B^*p(t), v(t)\rangle_{R^m} \, dt - \left\langle p(\bar{t}), \int_0^{\bar{t}} T(\bar{t}-t)Bv(t) \, dt \right\rangle_Z \\ &= \int_0^{\bar{t}} \langle -Ru(t), v(t)\rangle_{R^m} \, dt - \langle p(\bar{t}), y(\bar{t})\rangle_Z.\end{aligned} \tag{4.33}$$

Since $\langle p(\bar{t}), y(\bar{t})\rangle_Z \to 0$ as $t \to \infty$, (4.30) and (4.33) show that $J(z(0), u+v) - J(z(0), v) > 0$ if $v \neq 0$. Thus $u(t)$ is the optimal control for the initial condition $z(0)$, and we must have $u(t) = -R^{-1}B^*\Pi z(t)$. Hence $\dot{z}(t) = (\mathcal{A} - BR^{-1}B^*\Pi)z(t)$ and $z(t) = S(t)z(0)$, $t \geqq 0$. With this last expression for $z(\cdot)$, (4.18) yields

$$\begin{aligned}\dot{p}(t) &= -(\mathcal{A}^* - \Pi BR^{-1}B^*)p(t) + \Pi Bu(t) - Qz(t) \\ &= -(\mathcal{A}^* - \Pi BR^{-1}B^*)p(t) - (\Pi BR^{-1}B^*\Pi + Q)z(t),\end{aligned} \tag{4.34}$$

of which the integral version is

$$p(t) = S^*(\bar{t}-t)p(\bar{t}) + \int_t^{\bar{t}} S^*(s-t)[Q + \Pi BR^{-1}B\Pi]S(s-t)z(t) \, ds, \; 0 \leqq t \leqq \bar{t} < \infty. \tag{4.35}$$

As $\bar{t} \to \infty$, (4.35) and (4.7) yield

$$p(t) = \Pi z(t), \qquad t \geqq 0. \tag{4.36}$$

In particular, setting $t = 0$ yields

$$\operatorname{Re} p = \Pi \operatorname{Re} z. \tag{4.37}$$

Similarly, we could define $(z(t), p(t)) = \operatorname{Im}(e^{\lambda t}(z, p))$ and obtain

$$\operatorname{Im} p = \Pi \operatorname{Im} z. \tag{4.38}$$

Hence

$$p = \Pi z. \tag{4.39}$$

Since $(z, p) \neq 0$, we see from (4.39) that $z \neq 0$. Also, from (4.39) and the fact that $\hat{\mathcal{A}}(z, p) = \lambda(z, p)$, we have

$$\mathcal{A}z - B^*R^{-1}Bp = (\mathcal{A} - BR^{-1}B^*\Pi)z = \lambda z; \tag{4.40}$$

i.e., λ is an eigenvector of $\mathcal{A} - BR^{-1}B^*\Pi$.

Now suppose λ is an eigenvector of $\mathcal{A} - BR^{-1}B^*\Pi$. Then we have a nonzero $z \in D(\mathcal{A})$ such that $(\mathcal{A} - BR^{-1}B\Pi)z = \lambda z$. Letting $p = \Pi z$, we obtain

$$\mathcal{A}z - BR^{-1}B^*p = \lambda z. \tag{4.41}$$

Since $\Pi(Z) \subset D(\mathscr{A}^*)$, $p \in D(\mathscr{A}^*)$, and (4.2) yields

$$-Qz - \mathscr{A}^* p = -Qz - \mathscr{A}^* \Pi z = \Pi(\mathscr{A} - BR^{-1}B^*\Pi)z = \Pi(\lambda z) = \lambda p. \tag{4.42}$$

From (4.41) and (4.42), we have $\hat{\mathscr{A}}(z, p) = \lambda(z, p)$.

At this point, we have shown that $(z, p) \neq 0$ and $\hat{\mathscr{A}}(z, p) = \lambda(z, p)$ if and only if $z \neq 0$, $p = \Pi z$, and $(\mathscr{A} - BR^{-1}B^*\Pi)z = \lambda z$. From this, it is clear that $\{(z_j, p_j)\}$ is a set of linearly independent eigenvectors of $\hat{\mathscr{A}}$ corresponding to λ if and only if $\{z_j\}$ is a set of linearly independent eigenvectors of $\mathscr{A} - BR^{-1}B^*\Pi$ corresponding to λ. Hence λ has the same geometric multiplicity as an eigenvalue of $\hat{\mathscr{A}}$ as it has as an eigenvalue of $\mathscr{A} - BR^{-1}B^*\Pi$. Since $\mathscr{A} - BR^{-1}B^*\Pi$ has compact resolvent, both the algebraic and geometric multiplicities of λ as an eigenvalue of $\mathscr{A} - BR^{-1}B^*\Pi$ are finite.

For the equivalence of the algebraic multiplicities of λ as an eigenvalue of $\hat{\mathscr{A}}$ and of $\mathscr{A} - BR^{-1}B^*\Pi$, we must extend the foregoing arguments to generalized eigenvectors of rank $k = 2, 3, \cdots$.

Suppose $(z_k, p_k) \neq 0$, $k = 1, 2$, and $(\hat{\mathscr{A}} - \lambda)(z_1, p_1) = 0$ and $(\hat{\mathscr{A}} - \lambda)(z_2, p_2) = (z_1, p_1)$. Define $(z(t), p(t)) = \operatorname{Re}(e^{\lambda t}[(1+t)(z_1, p_1) + (z_2, p_2)])$ and $u(t) = -R^{-1}B^*p(t)$. Again we have $(\dot{z}(t), \dot{p}(t)) = \hat{\mathscr{A}}(z(t), p(t))$, and $u(t)$ is an admissible control for $z(0) = \operatorname{Re}(z_1 + z_2)$. We can proceed exactly as before to obtain (4.37)–(4.39) for $(z, p) = (z_1 + z_2, p_1 + p_2)$, and, since we already know that (4.37)–(4.39) hold for $(z, p) = (z_1, p_1)$, these equations must also hold for $(z, p) = (z_2, p_2)$. Then, since $p_2 = \Pi z_2$ and $(\hat{\mathscr{A}} - \lambda)(z_2, p_2) = (z_1, p_1)$,

$$(\mathscr{A} - BR^{-1}B^*\Pi - \lambda)z_2 = \mathscr{A}z_2 - BR^{-1}B^*p_2 - \lambda z_2 = z_1. \tag{4.43}$$

Since (4.40) holds for $(z, p) = (z_1, p_1)$, (4.43) says that z_2 is a generalized eigenvector of rank 2 of $\mathscr{A} - BR^{-1}B^*\Pi$, corresponding to the eigenvalue λ.

On the other hand, if $(\mathscr{A} - BR^{-1}B^*\Pi - \lambda)z_1 = 0$ and $(\mathscr{A} - BR^{-1}B^*\Pi - \lambda)z_2 = z_1$, letting $p_1 = \Pi z_1$ and $p_2 = \Pi z_2$, we can easily use (4.2) to show $(\hat{\mathscr{A}} - \lambda)(z_2, p_2) = (z_1, p_1)$. From here, a straightforward induction establishes a one-to-one correspondence between chains of generalized eigenvectors of $\mathscr{A} - BR^{-1}B^*\Pi$ and $\hat{\mathscr{A}}$, respectively. □

COROLLARY 4.3. *Let λ be an eigenvalue of $\mathscr{A} - BR^{-1}B^*\Pi$ with negative real part, z be in $D(\mathscr{A})$, and k a positive integer. Then $(\lambda - (\mathscr{A} - BR^{-1}B^*\Pi))^k z = 0$ if and only if $(\lambda - \hat{\mathscr{A}})^k(z, \Pi z) = 0$. Also, $(\lambda - \hat{\mathscr{A}})^k(z, p) = 0$ if and only if $(\lambda - (\mathscr{A} - BR^{-1}B^*\Pi))^k z = 0$ and $p = \Pi z$.* □

The next corollary follows from Theorem 4.7 and its proof, Theorem 4.9 and Corollary 4.3.

COROLLARY 4.4. *If λ is an eigenvalue of $\hat{\mathscr{A}}$ and $\mathscr{A} - BR^{-1}B^*\Pi$, then $(z, p) = ((x, \phi), (y, \psi))$ is an eigenvector of $\hat{\mathscr{A}}$ corresponding to λ if and only if x and y satisfy* (4.26a) *and* (4.26b),

$$\phi(\theta) = e^{\lambda\theta}x, \qquad -r \leqq \theta \leqq 0, \tag{4.44}$$

and ψ satisfies (2.7), (2.8) *and* (4.25d).

5. The traces of the operators $\Pi(t)$ and Π. Before discussing the traces of the solutions to the Riccati equations of the preceding sections, we should review some of the standard results for the trace of an infinite dimensional operator. For the trace results used here, see [3], [17], [22].

Let Z be a separable Hilbert space and T a compact linear operator on Z. Denoting by μ_i the eigenvalues of $(T^*T)^{1/2}$, repeated according to multiplicity, we have a sequence of nonnegative real μ_i's, each of finite multiplicity. The operator T

is said to be of *trace class* (or *nuclear*) if

$$\|T\|_1 = \sum_{i=1}^{\infty} \mu_i < \infty. \tag{5.1}$$

When the summation is finite, (5.1) defines the *trace norm* of T, and for any orthonormal basis $\{z_i\}$ in Z we have

$$\|T\|_1 = \sum_{i=1}^{\infty} \langle (T^*T)^{1/2} z_i, z_i \rangle_Z. \tag{5.2}$$

We denote the space of all trace class operators on Z by $\mathscr{L}_1(Z, Z)$. With the norm $\|\cdot\|_1$, $\mathscr{L}_1(Z, Z)$ is a Banach space and a two-sided ideal in $\mathscr{L}(Z, Z)$.

Any $T \in \mathscr{L}_1(Z, Z)$ has finite trace defined by

$$\operatorname{tr} T = \sum_{i=1}^{\infty} \langle T z_i, z_i \rangle_Z, \tag{5.3}$$

$\{z_i\}$ an arbitrary orthonormal basis in Z, the value of tr T being independent of the orthonormal basis chosen for (5.3).

Property 5.1. If $T \in \mathscr{L}_1(Z, Z)$, tr T is the absolutely convergent sum of the diagonal elements of the matrix representation of T referred to any complete family of mutually orthogonal elements of Z. If $\dim(Z) < \infty$, tr T is the diagonal sum of the matrix representation of T referred to any complete linearly independent family.

If $T \in \mathscr{L}_1(Z, Z)$ is self-adjoint, $\|T\|_1$ is just the sum of the absolute values of the repeated eigenvalues of T, and, if T is self-adjoint and nonnegative, $\|T\|_1 = \operatorname{tr} T$.

The following rather elementary property of trace class operators will be especially useful to us.

Property 5.2. If $T \in \mathscr{L}(Z, Z)$ has finite rank k, then $T \in \mathscr{L}_1(Z, Z)$ and $\|T\|_1 \leqq k\|T\|$.

We should note that, for $1 \leqq p < \infty$, we have the Banach algebra $\mathscr{L}_p(Z, Z)$, consisting of compact operators T for which the norm

$$\|T\|_p = \left(\sum_{i=1}^{\infty} \mu_i^p \right)^{1/p} \tag{5.4}$$

is finite. For $1 \leqq p \leqq q \leqq \infty$, $\mathscr{L}_q(Z, Z)$ contains $\mathscr{L}_p(Z, Z)$ algebraically and topologically, where $\mathscr{L}_\infty(Z, Z) = \mathscr{L}(Z, Z)$. In particular,

$$\|T\|_\infty = \|T\| \leqq \|T\|_1 \tag{5.5}$$

for $T \in \mathscr{L}_1(Z, Z)$. Also, $\mathscr{L}_2(Z, Z)$ is the space of Hilbert–Schmidt operators, and $T \in \mathscr{L}_1(Z, Z)$ if and only if T is the product of two Hilbert–Schmidt operators.

LEMMA 5.1. *For $t \geqq r$, $T(t)$ is Hilbert–Schmidt.*

Proof. For $u = 0$ in (2.1), define $\tilde{L}: Z \to L_2(-r, 0; R^n)$ by $(\tilde{L}(x(0), x_0))(\theta) = L x_{r+\theta}$, $-r < \theta < 0$. Since, for $u = 0$, $(x(t), x_t) = T(t)(x(0), x_0)$, it is not difficult to see that $\tilde{L} \in \mathscr{L}(Z, L_2(-r, 0; R^n))$. From (2.1), we have

$$x_r(\theta) = x(r+\theta) = x(0) + \int_{-r}^{\theta} (\tilde{L}(x(0), x_0))(\eta)\, d\eta, \qquad -r < \theta < 0.$$

Thus $T(r)$ can be written as the sum of operators of finite rank and the product of an integral operator on $L_2(-r, 0; R^n)$ with $\tilde{L}$. Therefore, $T(t) = T(t-r)T(r)$ is Hilbert–Schmidt for $t \geqq r$ because operators of finite rank are Hilbert–Schmidt, integral operators with L_2 kernals on $L_2(-r, 0; R^n)$ are Hilbert–Schmidt, and the product of a bounded operator with a Hilbert–Schmidt operator is Hilbert–Schmidt. □

THEOREM 5.1. *Let* $\Pi(\cdot)$ *be the solution of the Riccati integral equations* (3.14) *and* (3.24). *Then* $\Pi(t) \in \mathscr{L}_1(Z, Z)$ *for* $t \leqq t_f - r$. *If* $G \in \mathscr{L}_1(Z, Z)$, $\Pi(t) \in \mathscr{L}_1(Z, Z)$ *for* $t \leqq t_f$. *Furthermore,* $\Pi(\cdot)$ *satisfies*

$$\Pi(t) = T^*(t_f - t)GT(t_f - t) + \int_t^{t_f} T^*(\eta - t)[Q - \Pi(\eta)BR^{-1}B^*\Pi(\eta)]T(\eta - t)\, d\eta, \qquad t \leqq t_f, \tag{5.6}$$

where the integral is a Riemann integral convergent in $\mathscr{L}_1(Z, Z)$.

Proof. Since $T(s)$ is Hilbert–Schmidt for $s \geqq r$, $T^*(s)$ and $GT(s)$ are Hilbert–Schmidt for $s \geqq r$. Therefore $T^*(t_f - t)GT(t_f - t)$ is of trace class for $t_f - t \geqq r$. Since $\mathscr{L}_1(Z, Z)$ is a two-sided ideal in $\mathscr{L}(Z, Z)$, if $G \in \mathscr{L}_1(Z, Z)$, so is $T^*(t_f - t)GT(t_f - t)$ for $t_f - t \geqq 0$.

Now consider the integral on the right side of (5.6). For $z = (x, \phi) \in Z$, $T^*(\eta - t)Q^{1/2}z = T^*(\eta - t)(Q_0^{1/2}x, 0)$, and, for $u \in R^m$, $T^*(\eta - t)\Pi(\eta)Bu = T^*(\eta - t)\Pi(\eta)(B_0u, 0)$. Hence, since $T^*(\cdot)$ and $\Pi(\cdot)$ are strongly continuous and x and u have finite dimension, $T^*(\eta - t)Q^{1/2}$ is continuous in η and t in the uniform norm topology of $\mathscr{L}(Z, Z)$ and $T^*(\eta - t)\Pi(\eta)B$ is continuous in η and t in the uniform norm topology of $\mathscr{L}(R^m, Z)$. And, since the norm of an operator is equal to the norm of its adjoint, $Q^{1/2}T(\eta - t)$ and $B^*\Pi(\eta)T(\eta - t)$ are similarly continuous in the uniform norm topologies of $\mathscr{L}(Z, Z)$ and $\mathscr{L}(Z, R^m)$, respectively. Thus the integrand of the integral in (5.6) is continuous in η and t in the uniform norm topology of $\mathscr{L}(Z, Z)$.

For all t and η, the integrand of (5.6) has rank less than $m + n$. Therefore, Property 5.2 implies that the integrand is continuous in η and t in the trace norm, so that the integral in (5.6) converges in the Banach space $\mathscr{L}_1(Z, Z)$. Given $\Pi(\cdot)$, the right sides of (3.14) and (5.6) obviously define the same operator function of t. □

Note that $G \in \mathscr{L}_1(Z, Z)$ if G is defined by $G(x, \phi) = (G_0x, 0)$ where G_0 is a nonnegative, symmetric $n \times n$ matrix.

THEOREM 5.2. *The solution of the Riccati algebraic equation* (4.2) *is of trace class and satisfies*

$$\Pi = \int_0^{\infty} S^*(\eta)[Q + \Pi BR^{-1}B^*\Pi]S(\eta)\, d\eta, \tag{5.7}$$

where the integral is a Riemann integral absolutely convergent in $\mathscr{L}_1(Z, Z)$.

Proof. Recall that our hypothesis that any admissible control for the infinite interval drives the state to zero results in uniform exponential stability of the semigroup $S(\cdot)$. As in the proof of Theorem. 5.1, we see that the integrand in (5.7) has finite rank and is continuous in η in the trace norm. Also, since the integrand has rank $\leqq m + n$ and since its $\mathscr{L}(Z, Z)$-norm approaches zero exponentially as $\eta \to \infty$, by Property 5.2, the trace norm of the integrand approaches zero exponentially as $\eta \to \infty$. Therefore, the integral in (5.7) converges absolutely in $\mathscr{L}_1(Z, Z)$, and, by Corollary 4.2, the Π of (5.7) is the solution of (4.2). □

Since the trace norm is the strongest of the operator norms in (5.4) and is stronger than the usual supremum norm, it seems natural to ask whether finite dimensional approximations to the solutions of our infinite dimensional Riccati equations converge in trace norm. We will address this question in the next section, but first we should mention another use for $\operatorname{tr} \Pi(\cdot)$. For simplicity, let us discuss control on the infinite interval only.

For parameter optimization and sensitivity analysis in engineering design, it is often desired to compute an average or expected value of a quadratic performance

index of the form (4.1) based on a statistical distribution for the initial conditions [2, Chapt. 13], [28]. In a finite dimensional version of our problem, this expected value would be

$$E(J) = \operatorname{tr}(K\Pi), \tag{5.8}$$

where the real nonnegative, symmetric matrix K is the covariance of the initial conditions and Π is the solution of the appropriate Riccati equation. According to Theorem 5.2, we can define a similar performance measure for our hereditary problem:

$$\tilde{E}(J) = \operatorname{tr}(K\Pi), \tag{5.9}$$

where Π is the solution of the Riccati algebraic equation (4.2) and K is any nonnegative, self-adjoint bounded linear operator on Z.

Since $\mathscr{L}_1(Z, Z)$ is a two-sided ideal in $\mathscr{L}(Z, Z)$, $\operatorname{tr}(K\Pi)$, defined as in (5.3), is finite. From (4.2) and (5.3) we see that $\tilde{E}(J)$ is equal to the minimum value of the performance index in (4.1) summed over an arbitrary complete orthonormal set of initial conditions weighted according to the operator K, which may be taken as the covariance operator of the Z-valued random variable $z(0)$. (For Hilbert space-valued random variables and definitions of expectation and covariance, see [3].)

6. Abstract approximation theory. The operators $T(\cdot)$, B, G, Q and R are as in § 3. We assume that there exist a sequence of strongly continuous semigroups $T_N(\cdot)$ on Z and positive constants M and β such that

$$\|T_N(t)\| \leqq M e^{\beta t}, \qquad t \geqq 0, \quad N \geqq 1, \tag{6.1}$$

and that

$$T_N(t) \to T(t) \text{ strongly as } N \to \infty, \quad t \geqq 0. \tag{6.2}$$[5]

Because of (6.1), the convergence in (6.2) is necessarily uniform in t for t in bounded intervals (see [22, Thm. 2.16, p. 504]); i.e., for each $z \in Z$, $T_N(t)z \to T(t)z$ uniformly for t in bounded intervals. We denote the generator of $T_N(\cdot)$ by $\mathscr{A}_N$. Also, we assume the existence of a sequence of linear operators B_N from R^m to Z and sequences of nonnegative, self-adjoint bounded linear operators G_N and Q_N on Z such that

$$B_N \to B \text{ strongly}, \tag{6.3}$$

$$G_N \to G \text{ strongly}, \tag{6.4}$$

$$Q_N \to Q \text{ strongly}. \tag{6.5}$$

Since $\dim(R^m) < \infty$, (6.3) implies

$$\|B_N - B\| = \|B_N^* - B^*\| \to 0. \tag{6.6}$$

While (6.2) implies only

$$T_N^*(t) \to T^*(t) \text{ weakly}, \qquad t \geqq 0, \tag{6.7}$$

(6.2) and (6.3) imply

$$B_N^* T_N^*(t) = (T_N(t)B_N)^* \to B^* T^*(t) \text{ in norm}, \qquad t \geqq 0, \tag{6.8}$$

because $\dim(R^m) < \infty$. Also note that (6.3)–(6.5) imply that $\|B_N\|$, $\|G_N\|$, and $\|Q_N\|$ are uniformly bounded in N.

[5] As usual, we say that a sequence of operators T_N converges strongly (weakly) to an operator T if $T_N z$ converges strongly (weakly) to Tz for each z.

Remark 6.1. Until now, this paper has dealt with optimal control problems for hereditary differential systems only. However, for the approximation theory developed in this section, we must deal with sequences of optimal control problems defined in terms of the operators $T_N(\cdot)$, B_N, Q_N, G_N, and the approximating systems are not themselves hereditary differential systems. The analysis of [19], from which we have taken the basic results for infinite dimensional regulator problems and Riccati equations, covers the more general situation where $T(\cdot)$ is an arbitrary C_0-semigroup on a real Hilbert space Z, B is an arbitrary linear operator from R^m to Z, and Q and G are arbitrary nonnegative, self-adjoint operators on Z. The state $z(\cdot)$ is defined as in (2.5), and the performance indices are given by the right side of (3.2) and by (4.1). The equations and results of § 3 up to and including Theorem 3.2 hold for the general problem on a finite interval, and the equations and results of § 4 up to and including Theorem 4.3, excluding Corollary 4.1, hold for the general problem on the infinite interval. We exclude Corollary 4.1 to avoid entanglement, unnecessary here, in the difference between strong asymptotic stability and uniform exponential stability for general C_0 semigroups. Also, the results of § 5 depend only on rank Q and rank G being finite.

In considering approximation on finite time intervals, we will refer to the optimal control problem of § 3 as the "original optimal control problem" or "original problem," and to the optimal control problem corresponding to $T_N(\cdot)$, B_N and Q_N as the "Nth approximate problem." ($z(t_0)$ and R are the same for all the control problems.) Since the integrands of all the integrals encountered in this section are continuous, the convergence we need will follow from repeated application of the following lemma, whose proof is an easy exercise.

LEMMA 6.1. *Let X and Y be Banach spaces and let Ω be a compact subset of R^n. Suppose $\mathcal{A}(\cdot):\Omega\to\mathcal{L}(X, Y)$ and, for $i\geqq 1$, $\mathcal{A}_i(\cdot):\Omega\to\mathcal{L}(X, Y)$. Suppose also that $\|\mathcal{A}_i(\xi)\|$ is bounded uniformly in i and ξ, and that for each $x\in X$, $\mathcal{A}_i(\xi)x$ converges strongly (weakly) to $\mathcal{A}(\xi)x$ uniformly in ξ. Let $g(\cdot):\Omega\to X$ be continuous and suppose there is a sequence of functions $g_i(\cdot):\Omega\to X$ which converge uniformly to $g(\cdot)$. Then $\mathcal{A}_i(\cdot)g_i(\cdot)$ converges strongly (weakly) uniformly to $\mathcal{A}(\cdot)g(\cdot)$.*

Referring to (3.8) and (3.9), we write $\tilde{R}_{sN}$ and $\tilde{B}^*_{sN}$ for the corresponding operators in the Nth problem with $t_0=s$ and note that $\|\tilde{R}_{sN}\|$, $\|\tilde{R}^{-1}_{sN}\|$ and $\|\tilde{B}^*_{sN}\|$ are bounded uniformly in s and N for s in bounded intervals ($s\leqq t_f$). In particular $\|\tilde{R}^{-1}_{sN}\|\leqq\|R^{-1}\|$. Dominated convergence, with (3.3)–(3.5), (3.8)–(3.11), and (6.2)–(6.8), implies

$$\tilde{R}_{sN}v\to\tilde{R}_sv,\qquad v\in L_2(s, t_f; R^m) \tag{6.9}$$

and

$$\tilde{B}^*_{sN}z\to\tilde{B}^*_sz,\qquad z\in Z, \tag{6.10}$$

the convergence in both (6.9) and (6.10) being in $L_2(s, t_f; R^m)$ for $s\leqq t_f$. The identity

$$\tilde{R}^{-1}_{sN}-\tilde{R}^{-1}_s=\tilde{R}^{-1}_{sN}(\tilde{R}_s-\tilde{R}_{sN})\tilde{R}^{-1}_s \tag{6.11}$$

and (6.9) and (6.10) imply

$$\tilde{R}^{-1}_{sN}\tilde{B}^*_{sN}z\to\tilde{R}^{-1}_s\tilde{B}^*_sz\quad\text{in } L_2(s, t_f; R^m),\quad z\in Z. \tag{6.12}$$

Denoting the optimal control for the original problem by u and the optimal control for the Nth approximate optimal control problem by u_N, we see from (3.12) and (6.12) that

$$u_N\to u\quad\text{in } L_2(t_0, t_f; R^m). \tag{6.13}$$

Denoting by $\Pi_N(\cdot)$ the solution of the Riccati integral equations for the Nth approximate problem and by $S_N(\cdot,\cdot)$ the perturbed evolution operator corresponding to the perturbation of $T_N(\cdot)$ by $-B_N R^{-1} B_N^* \Pi_N$, we see from (2.5), (3.22) and (6.13) that

$$S_N(t,s)z \to S(t,s)z, \qquad z \in Z, \quad t_0 \leqq s \leqq t \leqq t_f, \tag{6.14}$$

where the convergence is uniform in t and s. ($S(\cdot,\cdot)$ is the evolution operator of (3.21)–(3.24).) Then (3.23), (6.4), (6.5), (6.7) and (6.14) imply

$$\Pi_N(t) \to \Pi(t) \text{ weakly}, \qquad t_0 \leqq t \leqq t_f, \tag{6.15}$$

where the convergence is uniform in t. Since $B_N^* \Pi_N(t) = (\Pi_N(t) B_N)^*$ and $\dim(R^m) < \infty$, (3.13), (6.3) and (6.15) show

$$\|u_N(t) - u(t)\|_{R^m} \to 0 \tag{6.16}$$

uniformly for $t_0 \leqq t \leqq t_f$. If we have strong convergence in (6.7), $\Pi_N(t)$ converges strongly to $\Pi(t)$.

Remark 6.2. Since the initial time t_0 is arbitrary, any time we said above that convergence was uniform in t for $t_0 \leqq t \leqq t_f$, we could just as well have said that convergence was uniform in t for t in bounded intervals.

Thus we have the following about the solutions $\Pi_N(\cdot)$ of the sequences of approximating Riccati integral equations:

THEOREM 6.1. *If the boundedness and convergence hypotheses in* (6.1)–(6.5) *hold, then, for* $t \leqq t_f$, $\Pi_N(t)$ *converges weakly to* $\Pi(t)$, *the solution of the Riccati integral equations for the original problem, and the convergence is uniform in t for t in bounded intervals. If also* $T_N^*(t)$ *converges strongly to* $T^*(t)$ *for* $t \geqq 0$, *then* $\Pi_N(t)$ *converges strongly to* $\Pi(t)$, *and the convergence is uniform in t for t in bounded intervals.*

The significance of strong convergence for $\Pi_N(t)$ lies in the following theorem.

THEOREM 6.2. *Suppose that* $\Pi_N(t)$ *converges strongly to* $\Pi(t)$ *and that the convergence is uniform in t for t in bounded intervals. Then* $B_N^* \Pi_N(t)$ *converges in norm to* $B^* \Pi(t)$, *i.e.,*

$$\|B_N^* \Pi_N(t) - B^* \Pi(t)\| \to 0 \tag{6.17}$$

uniformly in bounded t-intervals; and, for $\varepsilon > 0$, *there exists a nonincreasing function* $N_\varepsilon(\cdot) : (-\infty; t_f) \to R_+$ *such that, for* $N \geqq N_\varepsilon(s)$,

$$\begin{aligned} J(s, z(s), -R^{-1} B_N^* \Pi_N(\cdot) \tilde{z}_N(\cdot)) &\leqq J(s, z(s), -R^{-1} B^* \Pi(\cdot) z(\cdot)) + \varepsilon \|z(s)\|^2 \\ &= \langle \Pi(s) z(s), z(s) \rangle_Z + \varepsilon \|z(s)\|^2, \qquad z(s) \in Z, \end{aligned} \tag{6.18}$$

where $J(s, z(s), -R^{-1} B_N^* \Pi_N(\cdot) \tilde{z}_N(\cdot))$ *is the value of the performance index in* (3.2) *corresponding to the initial time s, the initial state* $z(s)$, *and the feedback control* $\tilde{u}_N(\cdot) = -R^{-1} B_N^* \Pi_N(\cdot) \tilde{z}_N(\cdot)$, *and* $\tilde{z}_N(\cdot)$ *is the corresponding solution of* (2.4) *and* (2.5); *i.e., the* Nth *feedback control law is applied to the original hereditary system.*

Proof. Since $\Pi_N(t)$ and $\Pi(t)$ are self-adjoint, $\|B_N^* \Pi_N(t) - B^* \Pi(t)\| = \|\Pi_N(t) B_N - \Pi(t) B\|$. Since $\Pi_N(t)$ and B_N converge strongly, so does $\Pi_N(t) B_N$, and the finite dimensionality of R^m then implies $\|\Pi_N(t) B_N - \Pi(t) B\| \to 0$.

For (6.18), let $\tilde{S}_N(\cdot,\cdot)$ be the perturbed evolution operator corresponding to the perturbation of $T(\cdot)$ by $-BR^{-1} B_N^* \Pi_N(\cdot)$, and recall the definition of $S(\cdot,\cdot)$ in 3.21). Then (3.17), (3.18) and (3.20) along with Gronwall's lemma and the fact that $\|B_N^* \Pi_N(t) - B^* \Pi(t)\| \to 0$ uniformly for $s \leqq t \leqq t_f$ imply that $\|\tilde{S}_N(t,s) - S(t,s)\|$ goes to zero uniformly for $s \leqq t \leqq t_f$ as $\|BR^{-1} B_N^* \Pi_N(\cdot) - BR^{-1} B^* \Pi(\cdot)\|_{\mathcal{B}_\infty(s,t_f;Z,Z)}$ goes to

zero. The existence of $N_\varepsilon(\cdot)$ and (6.18) then follow from (3.15), (3.24) and

$$
\begin{aligned}
&J(s, z(s), -R^{-1}B_N^*\Pi_N(\cdot)\tilde{z}_N(\cdot)) \\
&\quad = \langle \tilde{S}_N^*(t_f, s)G\tilde{S}_N(t_f, s)z(s), z(s)\rangle_Z \\
&\quad\quad + \left\langle \int_s^{t_f} \tilde{S}_N^*(\eta, t)[Q + \Pi_N(\eta)B_NR^{-1}B_N^*\Pi_N(\eta)]\tilde{S}_N(\eta, t)z(s)\, d\eta, z(s)\right\rangle_Z . \qquad \square
\end{aligned} \tag{6.19}
$$

When $T^*(t)$ converges strongly, we have even more:

THEOREM 6.3. *Suppose that, for all N, $Q_N = Q$ and $G_N = G$ is given by $G(x, \phi) = (G_0x, 0)$. Then, if $T_N^*(t)$ converges strongly to $T^*(t)$ for $t \geqq 0$, $\Pi_N(t)$ converges in trace norm to $\Pi(t)$ for $t \leqq t_f$ and the convergence is uniform for t in bounded intervals.*

Proof. We will use (3.14), the first Riccati integral equation. Since

$$
\dim(R^n) < \infty, \qquad \|(T_N^*(t_f - t) - T^*(t_f - t))G^{1/2}\| \to 0
$$

uniformly in t for t in bounded intervals. Since

$$
\|G_N^{1/2}(T_N(t_f - t) - T(t_f - t))\| = \|(T_N^*(t_f - t) - T^*(t_f - t))G^{1/2}\|
$$

and rank

$$
G^{1/2} \leqq n, \qquad \|T_N^*(t_p - t)GT_N(t_f - t) - T^*(t_f - t)GT(t_f - t)\|_1 \to 0
$$

by Property 5.2. Similarly,

$$
\|T_N^*(\eta - t)QT_N(\eta - t) - T^*(\eta - t)QT(\eta - t)\|_1 \to 0.
$$

Finally, by Theorem. 6.1, $\Pi_N(\eta)$ converges strongly to $\Pi(\eta)$, so that

$$
\|T_N^*(\eta - t)\Pi_N(\eta)B_N - T^*(\eta - t)\Pi(\eta)B\| \to 0.
$$

Hence, by Property 5.2, since rank

$$
\begin{aligned}
B_N \leqq m, \qquad &\|T_N^*(\eta - t)\Pi_N(\eta)B_NR^{-1}B_N^*\Pi_N(\eta)T_N(\eta - t) \\
&- T^*(\eta - t)\Pi(\eta)BR^{-1}B^*\Pi(\eta)T(\eta - t)\|_1 \to 0.
\end{aligned}
$$

All convergence here is uniform for t and η in bounded intervals, and the theorem follows from (3.14). $\square$

COROLLARY 6.1. *Under the hypothesis of Theorem* 6.3, $\operatorname{tr}\Pi_N(t) = \|\Pi_N(t)\|_1$ *converges to* $\operatorname{tr}\Pi(t) = \|\Pi(t)\|_1$ *uniformly for t in bounded intervals.*

Note that, until Theorem 6.3, we did not require Q_N and G_N to have finite rank. This generally will be the case in numerical approximations, so that each $\Pi_N(t)$ will have finite trace.

With only weak convergence of $\Pi_N(t)$, we cannot guarantee that $\operatorname{tr}\Pi_N(t)$ converges to $\operatorname{tr}\Pi(t)$; all we can say is the following, which follows from (5.3).

THEOREM 6.4. *Let Π_N be a sequence of nonnegative, self-adjoint bounded linear operators on a Hilbert space Z, and suppose that Π_N converges weakly to $\Pi \in \mathcal{L}(Z, Z)$. (Necessarily, $\Pi = \Pi^* \geqq 0$.) If Π and each Π_N have finite trace, then* $\liminf \operatorname{tr}\Pi_N \geqq \operatorname{tr}\Pi$.

Also, changing the hypothesis to strong convergence for Π_N would not enable us to say that $\lim \operatorname{tr}\Pi_N = \operatorname{tr}\Pi$.

Next we consider approximation on the infinite interval, letting $t_0 = 0$ and $t_f = \infty$. We will refer to the problem of § 4 as the "original optimal control problem on the infinite interval" or "original problem," and to the infinite-interval optimal control problem corresponding to $T_N(\cdot)$, B_N and Q_N as the "Nth approximate optimal control problem on the infinite interval" or "Nth approximate problem." The following two theorems are essential to the development here.

THEOREM 6.5. *If $\{\Pi_i\}$ is a uniformly bounded sequence of bounded linear operators on a separable Hilbert space Z, there exists a subsequence which converges weakly to some $\Pi \in \mathscr{L}(Z, Z)$. If each Π_i is nonnegative and self-adjoint, so is Π.*

Proof. Let $\{z_j\}$ be a basis for Z. Since $\|\Pi_i z_1\|$ is bounded in i, there is a subsequence $\{\Pi_i^{(1)}\}$ such that $\{\Pi_i^{(1)} z_1\}$ converges weakly. From $\{\Pi_i^{(1)}\}$ we can extract a subsequence $\{\Pi_i^{(2)}\}$ such that $\{\Pi_i^{(2)} z_2\}$ converges weakly. Proceeding in this way, we construct a sequence of subsequences $\{\Pi_i^{(j)}\}$ such that $\Pi_i^{(j)} z_k$ converges weakly (as $i \to \infty$) for $1 \leqq k \leqq j$. Next we show that the diagonal sequence $\{\tilde{\Pi}_j = \Pi_j^{(j)}\}$ is a weakly convergent subsequence of the original sequence. Clearly, for each i, the sequence $\tilde{\Pi}_j z_i$ converges weakly to some y_i. The set $\tilde{Z}$ of finite linear combinations of z_j's is dense in Z, and, for $\tilde{z} \in \tilde{Z}$, $\tilde{\Pi}_j \tilde{z}$ converges weakly to $\tilde{\Pi}\tilde{z}$, where $\tilde{\Pi}$ is a uniquely determined bounded linear operator from $\tilde{Z}$ to Z. Therefore, $\tilde{\Pi}_j$ converges weakly to Π, where Π is the extension of $\tilde{\Pi}$ to all of Z. It is easy to show that, if each Π_i is non-negative and self-adjoint, so is Π. □

Of course, the standard proof that a Hilbert space is weakly sequentially compact [3], [16], [22] suggested the use of the diagonal subsequence to prove Theorem 6.5; however, the present author has been unable to find this result in the previous literature.

THEOREM 6.6. *Let $\{C_N\}$ be a sequence of bounded linear operators on Z, and let $S_N(\cdot)$ be the semigroup generated by $\mathscr{A}_N + C_N$, where $\mathscr{A}_N$ is the generator of the semigroup $T_N(\cdot)$ in (6.2). If C_N converges strongly to a bounded linear operator C, then, for $t \geqq 0$, $S_N(t)$ converges strongly to $S(t)$, the semigroup generated by $\mathscr{A} + C$, and the convergence is uniform in t for t in bounded intervals.*

Proof. We will use a standard series to construct the perturbed semigroups (see [21], [22, pp. 497–98]). We have

$$S_N(t) = \sum_{j=0}^{\infty} S_N^{(j)}(t), \qquad t \geqq 0, \tag{6.20}$$

where $S_N^{(0)}(t) = T_N(t)$ and

$$S_N^{(j+1)}(t)z = \int_0^t T_N(t-s) C_N S_N^{(j)}(s) z \, ds, \qquad z \in Z, \quad t \geqq 0, \quad j \geqq 0. \tag{6.21}$$

It can be shown (see [21], [22]) that

$$\|S_N^{(j)}(t)\| \leqq M^{j+1} \|C_N\|^j \frac{e^{\beta t} t^j}{j!}, \qquad j \geqq 0. \tag{6.22}$$

Since, by the principle of uniform boundedness, $\|C_N\|$ is bounded in N, we see from (6.22) that the series in (6.20) converges absolutely in $\mathscr{L}(Z, Z)$, and uniformly in N and t for t in bounded intervals. Also,

$$S(t) = \sum_{j=0}^{\infty} S^{(j)}(t), \qquad t \geqq 0, \tag{6.23}$$

where $S^{(0)}(t) = T(t)$ and $\{S^{(j)}(\cdot)\}$ is defined in the same way as $\{S_N^{(j)}(\cdot)\}$ was in (6.21). The estimate for $\|S^{(j)}(\cdot)\|$, like (6.22), shows that the series in (6.23) converges absolutely, and uniformly for t in bounded intervals.

Since $T_N(\cdot)$ and C_N converge strongly to $T(\cdot)$ and C, from (6.21) we see

$$\lim_{N \to \infty} S_N^{(j)}(t)z = S^{(j)}(t)z, \qquad z \in Z, \quad t \geqq 0, \quad j \geqq 0, \tag{6.24}$$

the convergence being uniform in t for t in bounded intervals. From (6.23) and (6.24), we have

$$\lim_{J \to \infty} \lim_{N \to \infty} \sum_{j=0}^{J} S_N^{(j)}(t)z = S(t)z. \tag{6.25}$$

Because the series in (6.20) converges uniformly in N, we can reverse the order of the limits in (6.25) to obtain

$$\lim_{N\to\infty} S_N(t)z = S(t)z, \qquad z\in Z, \quad t\geqq 0, \tag{6.26}$$

which was to be proved. Of course, the convergence in (6.26) is uniform in t for t in bounded intervals. □

Hypothesis 6.1. From here on, we assume that, for each N, there exists a unique nonnegative, self-adjoint solution Π_N of the Riccati algebraic equation corresponding to the Nth approximate problem.

$S_N(\cdot)$ is the semigroup generated by $\mathcal{A}_N - B_N R^{-1} B_N^* \Pi_N$.

THEOREM 6.7. *If $\|\Pi_N\|$ is bounded in N, then there exists a nonnegative, self-adjoint solution Π of the Riccati algebraic equation* (4.2), *and Π_N converges weakly to Π.*

Proof. By Theorem 6.5, there is a subsequence $\{\Pi_{N_j}\}$ which converges weakly to some nonnegative, self-adjoint $\Pi\in\mathcal{L}(Z,Z)$. We must show that Π is a solution of (4.2).

From Theorem 4.2, we have

$$\Pi_{N_j} z = S_{N_j}^*(t)\Pi_{N_j} S_{N_j}(t)z + \int_0^t S_{N_j}^*(\eta)[Q_{N_j} + \Pi_{N_j} B_{N_j} R^{-1} B_{N_j}^* \Pi_{N_j}] S_{N_j}(\eta) z \, d\eta, \tag{6.27}$$
$$0\leqq t<\infty, \quad z\in Z, \quad N_j\geqq 1.$$

Since $\dim(R^m)<\infty$, $B_{N_j}^*\Pi_{N_j}$ converges strongly to $B^*\Pi$, so that $B_{N_j}R^{-1}B_{N_j}^*\Pi_{N_j}$ converges strongly to $BR^{-1}B^*\Pi$. By Theorem 6.6 then, $S_{N_j}(t)$ converges strongly to $S(t)$ for $t\geqq 0$, where $S(\cdot)$ is the semigroup generated by $\mathcal{A}-BR^{-1}B^*\Pi$. Since $\Pi_{N_j}S_{N_j}(t)$ converges weakly to $\Pi S(t)$, $S_{N_j}^*(t)\Pi_{N_j}S_{N_j}(t) = (\Pi_{N_j}S_{N_j}(t))^* S_{N_j}(t)$ converges weakly to $S^*(t)\Pi S(t)$. Since $B_{N_j}^*\Pi_{N_j}$ and $S_{N_j}(t)$ converge strongly, $B_{N_j}^*\Pi_{N_j}S_{N_j}(t)$ converges strongly to $B^*\Pi S(t)$ and $S_{N_j}^*\Pi_{N_j}B_{N_j} = (B_{N_j}^*\Pi_{N_j}S_{N_j}(t))^*$ converges weakly to $S(t)\Pi B$. Of course, $S_{N_j}^*(t)$ converges weakly to $S^*(t)$, and Q_{N_j} converges strongly to Q. All convergence here is uniform in t for t in bounded intervals. Thus, for $t\geqq 0$ and $z\in Z$, the integral on the right side of (6.27) converges weakly to $\int_0^t S^*(\eta)[Q+BR^{-1}B^*\Pi]S(\eta)z\,d\eta$, and we see that Π is a solution of the Riccati algebraic equation for the original problem.

To see that the original sequence $\{\Pi_N\}$ converges weakly to Π, note that the argument we have just given shows that any subsequence of $\{\Pi_N\}$ in turn contains a subsequence which converges to a nonnegative, self-adjoint solution of the Riccati algebraic equation for the original problem, and recall that Theorem 4.1 says that such a solution is unique. □

The following theorem is proved in the same way as Theorem 6.2.

THEOREM 6.8. *Suppose that Π_N converges strongly to Π. Then $B_N^*\Pi_N$ converges in norm to $B^*\Pi$, and the semigroup $\tilde{S}_N(\cdot)$, generated by $\mathcal{A}-BR^{-1}B_N^*\Pi_N$, is uniformly exponentially stable for N sufficiently large. Also, for $\varepsilon>0$, there exists $N_\varepsilon>0$ such that, for $N\geqq N_\varepsilon$,*

$$\begin{aligned} J(z,-R^{-1}B_N^*\Pi_N z(\cdot)) &< J(z,-R^{-1}B^*\Pi z(\cdot)) + \varepsilon\|z\|^2 \\ &= \langle \Pi z, z\rangle_Z + \varepsilon\|z\|^2, \qquad z\in Z, \end{aligned} \tag{6.28}$$

where $J(z,-R^{-1}B_N^*\Pi_N z(\cdot))$ is the value of the performance index in (4.1) *corresponding to the initial state z and the feedback control $\tilde{u}_N = -R^{-1}B_N^*\Pi_N z(\cdot)$.*

THEOREM 6.9. *Assume that $Q_N = Q$ for all N and that $T_N^*(t)$ converges strongly to $T^*(t)$ for $t\geqq 0$. Furthermore, suppose there exist positive constants M and β such that*

$$\|S_N(t)\| \leqq M e^{-\beta t}, \qquad t\geqq 0, \quad N\geqq 1 \tag{6.29}$$

and

$$\Pi_N \leqq M, \qquad N \geqq 1. \tag{6.30}$$

Then there exists a nonnegative, self-adjoint solution Π *of the Riccati algebraic equation for the original problem,* Π_N *converges in trace norm to* Π, *and*

$$\|S(t)\| \leqq M e^{-\beta t}, \qquad t \geqq 0, \tag{6.31}$$

where $S(\cdot)$ *is the semigroup generated by* $\mathcal{A} - BR^{-1}B^*\Pi$.

Proof. We know from Theorem 6.7 that Π_N converges weakly to the required Π, and this is sufficient for $S_N(t)$ to converge strongly to $S(t)$ (see the proof of Theorem 6.7). Hence (6.29) implies (6.31). Also, for each z, since $S_N(t)z \to S(t)z$ uniformly in bounded t-intervals, (6.29) then implies that $S_N(t)z \to S(t)z$ uniformly for $t \geqq 0$.

Next we show that Π_N converges strongly to Π. We compute Π and Π_N according to Theorem 4.3 (see Remark 6.1). We take $G = MI$, where I is the identity, so that $G \geqq \Pi_N$, $N \geqq 1$, and $G \geqq \Pi$ (since Π_N converges weakly to Π). For $\varepsilon > 0$, by (6.29) and (6.31), we can choose s according to (4.10) such that

$$\|\Pi_N(s) - \Pi_N\| < \varepsilon, \qquad N \geqq 1 \tag{6.32}$$

and

$$\|\Pi(s) - \Pi\| < \varepsilon. \tag{6.33}$$

Holding s fixed, for $z \in Z$, we can choose N according to Theorem 6.3 such that

(6.34)

$$\|\Pi_N z - \Pi z\| \leqq \|\Pi_N z - \Pi_N(s)z\| + \|\Pi_N(s)z - \Pi(s)z\| + \|\Pi(s)z - \Pi z\| < 3\varepsilon, \qquad N \geqq \bar{N}.$$

Since $\Pi_N B_N R^{-1} B_N^*$ and $T_N^*(t)$ converge strongly, so does $S_N^*(t)$, and, as with $S_N(t)$, (6.29) implies that the convergence is uniform in t for $t \geqq 0$. From Theorem 5.2 (and Remark 6.1), we have

$$\Pi_N = \int_0^\infty S_N^*(\eta)[Q + \Pi_N B_N R^{-1} B_N^* \Pi_N] S_N(\eta)\, d\eta, \tag{6.35}$$

and the remaining argument is quite similar to the proof of Theorem 6.3. Because of the finite dimensionality of R^n and R^m and the strong convergence we have already,

$$\|S_N^*(\eta)Q^{1/2} - S^*(\eta)Q^{1/2}\| = \|Q^{1/2}S_N(\eta) - Q^{1/2}S(\eta)\| \to 0$$

and

$$\|S_N^*(\eta)\Pi_N B_N - S^*(\eta)\Pi B\| = \|B_N^*\Pi_N S_N(\eta) - B\Pi S(\eta)\| \to 0,$$

uniformly for $\eta \geqq 0$. Then Property 5.2 implies that the integrand in (6.35) converges in trace norm, uniformly in η, to the integrand in (5.7), and this convergence along with (6.29) and (6.31) implies $\|\Pi_N - \Pi\|_1 \to 0$. □

Remark 6.3. In Theorems 6.3 and 6.9, we could let $Q_N = P_N Q P_N$ and $G_N = P_N G P_N$ where P_N is a sequence of continuous projections on Z which converge strongly to the identity. We would need only to replace $Q^{1/2}$ by $P_N Q^{1/2}$ or $Q^{1/2}P_N$ in the appropriate places in the proofs, and similarly for $G^{1/2}$. It has been convenient to avoid discussing projections explicitly in this section, and the versions of the theorems we have suffice for the approximation scheme of the next section. However, to apply the results of this section to a scheme where Z is decomposed with the eigenvectors of $\mathcal{A}$ (see [5], [20], [27]), we would take $Q_N = P_N Q P_N$ and $G_N = P_N G P_N$ where P_N is the projection onto the span of a finite number of such eigenvectors.

7. An approximation scheme.

7.1. Approximation of the semigroups $T(\cdot)$ and $T^*(\cdot)$. The convergence results for the sequences of semigroups in this section follow from the Trotter–Kato semigroup approximation theorem ([22], [33], [40]). We will give a convenient version of this theorem, which is equivalent to the version used in [4].

THEOREM 7.1. *Let $\mathcal{A}$ generate a C_0-semigroup $T(\cdot)$ on a Hilbert space Z and suppose there is a sequence of linear operators $\mathcal{A}_N$ each of which generates a C_0-semigroup $T_N(\cdot)$ on Z. Suppose there exist constants M and β such that* (6.1) *holds and a dense subset $\mathcal{D}$ of Z such that, for* Re λ *sufficiently large, $(\mathcal{A}-\lambda)\mathcal{D}$ is dense in Z. Then, if*

$$\mathcal{A}_N z \to \mathcal{A} z \quad \text{for } z \in \mathcal{D}, \tag{7.1}$$

$T_N(t)z \to T(t)z$ for $t \geqq 0$ and $z \in Z$, and the convergence is uniform in t for t in bounded intervals. Furthermore, (6.1) *holds with $T_N(\cdot)$ replaced by $T(\cdot)$.*

From here on, we will restrict most of our analysis to the system

$$\dot{x}(t) = A_0 x(t) + A_1 x(t-r) + \int_{-r}^{0} D(\theta) x(t+\theta)\, d\theta + B_0 u(t), \qquad t \geqq t_0. \tag{7.2}$$

For any positive integer N, let χ_j^N denote the characteristic function of $[-jr/N, -(j-1)r/N]$ for $2 \leqq j \leqq N$ and χ_1^N denote the characteristic function of $[-r/N, 0]$. Define the finite dimensional subspace Z_N of Z by

$$Z_N = \left\{ (\psi_0, \psi) \in Z : \psi = \sum_{j=1}^{N} \psi_j \chi_j^N ;\ \psi_j \in R^n,\ 0 \leqq j \leqq N \right\}. \tag{7.3}$$

Note that the orthogonal projection $\mathcal{P}_N$ of Z onto Z_N is given by

$$\mathcal{P}_N(x, \psi) = \left(x, \sum_{j=1}^{N} \psi_j \chi_j^N \right) \tag{7.4}$$

where

$$\psi_j = \frac{N}{r} \int_{-jr/N}^{(j-1)r/N} \psi(s)\, ds, \qquad 1 \leqq j \leqq N. \tag{7.5}$$

Next we define a sequence of operators $\mathcal{A}_N : Z \to Z_N$. For $(\psi_0, \psi) \in Z_N$,

$$\mathcal{A}_N(\psi_0, \psi) = \left(A_0\psi_0 + A_1\psi_N + \sum_{j=1}^{N} \frac{r}{N} D_j^N \psi_j,\ \sum_{j=1}^{N} \frac{N}{r} (\psi_{j-1} - \psi_j) \chi_j^N \right), \tag{7.6}$$

where

$$D_j^N = \frac{N}{r} \int_{-jr/N}^{-(j-1)r/N} D(s)\, ds, \qquad 1 \leqq j \leqq N. \tag{7.7}$$

For $(x, \psi) \in Z$, we take $\mathcal{A}_N(x, \psi) = \mathcal{A}_N \mathcal{P}_N(x, \psi)$.

For the approximation scheme of this section, we take $T_N(t) = e^{\mathcal{A}_N t}$. Note that Z_N and $Z_N^{\perp}$ reduce $\mathcal{A}_N$ and $T_N(t)$. According to the next two lemmas, which Banks and Burns proved in [4], this sequence of approximating semigroups satisfies the hypotheses of Theorem 7.1.

LEMMA 7.1 ([4, Lemmas 3.2 and 3.3]). *The set $\mathcal{D} = \{(\psi(0), \psi) \in Z : \psi$ is continuously differentiable on $[-r, 0]\}$ is dense in Z, and $\mathcal{A}_N z \to \mathcal{A} z$ for $z \in \mathcal{D}$. Also, for* Re λ *sufficiently large, $(\mathcal{A}-\lambda)\mathcal{D}$ is dense in Z.*

LEMMA 7.2. ([4, Lemma 3.6]). *There exist constants M and β such that* (6.1) *holds for the approximation scheme of this section.*

We follow [4] and define the vectors $e_0^N, e_1^N, \cdots, e_N^N$ by

$$e_0^N = (1, 0), \qquad e_j^N = (0, \chi_j^N), \qquad 1 \leqq j \leqq N, \tag{7.8}$$

so that each $z \in Z_N$ can be written as

$$z = \sum_{j=0}^{N} w_j e_j^N, \tag{7.9}$$

where each w_j is an n-vector. Thus each $z \in Z$ corresponds to a unique $n(N+1)$ vector of scalars. When the elements of Z_N are represented in this way, the restriction of $\mathscr{A}_N$ to Z_N has the matrix representation

$$A^N = \begin{bmatrix} A_0 & \frac{r}{N}D_1^N & \frac{r}{N}D_2^N & \cdots & A_1+\frac{r}{N}D_N^N \\ \frac{N}{r}I & -\frac{N}{r}I & 0 & & 0 \\ 0 & \frac{N}{r}I & -\frac{N}{r}I & \vdots & \vdots \\ \vdots & & \ddots & & \\ 0 & \cdots & & \frac{N}{r}I & -\frac{N}{r}I \end{bmatrix}, \tag{7.10}$$

where I is the $n \times n$ identity matrix.

For $1 \leqq j \leqq n(N+1)$, define $z_j^N \in Z_N$ such that the jth component of the $n(N+1)$ vector representing z_j^N is 1 and all other components are zero. In this section, all matrix representations of linear operators on Z_N, including A^N, are the matrix representations of the operators referred to the basis vectors z_j^N. Let W^N be the $n(N+1) \times n(N+1)$ matrix whose elements are $\langle z_i^N, z_j^N \rangle_Z$. Then

$$W^N = \begin{bmatrix} I & 0 & \cdots & 0 \\ 0 & \frac{r}{N}I & 0 \cdots & 0 \\ \vdots & 0 & \frac{r}{N}I & \vdots \\ & \vdots & \ddots & 0 \\ 0 & 0 & \cdots 0 & \frac{r}{N}I \end{bmatrix}. \tag{7.11}$$

It is an elementary exercise to show that the restriction of $\mathscr{A}_N^*$ to Z_N has the matrix representation

$$A^{*N} = W^{N-1}A^{NT}W^N = \begin{bmatrix} A_0^T & I & 0 & \cdots & 0 \\ D_1^{NT} & -\frac{N}{r}I & \frac{N}{r}I & 0 \cdots & \vdots \\ \vdots & 0 & -\frac{N}{r}I & \frac{N}{r}I \cdots & \\ D_{N-1}^{NT} & \vdots & & \ddots & \frac{N}{r}I \\ \frac{N}{r}A^T + D_N^{NT} & & 0 & \cdots 0 & -\frac{N}{r}I \end{bmatrix}. \tag{7.12}$$

THEOREM 7.2. *The set* $\tilde{\mathscr{D}} = \{(y, \psi) \in D(\mathscr{A}^*)$: ψ *is continuously differentiable on* $[-r, 0]\}$ *is dense in* Z *and* $\mathscr{A}_N^* z \to \mathscr{A}^* z$ *for* $z \in \tilde{\mathscr{D}}$. *For* $\operatorname{Re}\lambda$ *sufficiently large,* $(\mathscr{A}^* - \lambda)\tilde{\mathscr{D}}$ *is dense in* Z.

Proof. From Theorem 2.1, it is easy to see that $\tilde{\mathscr{D}}$ is dense in Z. Since $\mathscr{A}_N = \mathscr{A}_N \mathscr{P}_N : Z \to Z_N$, $\mathscr{A}_N = \mathscr{P}_N \mathscr{A}_N \mathscr{P}_N$ and $\mathscr{A}_N^* = \mathscr{P}_N \mathscr{A}_N^* \mathscr{P}_N$. Then, since $\mathscr{P}_N \to I$ strongly, we need only show that $\mathscr{A}_N^* \mathscr{P}_N z \to \mathscr{A}^* z$ for $z \in \tilde{\mathscr{D}}$.

First consider the case $D = 0$. Because of the boundary condition (2.8), the statement that $\mathscr{A}_N^* \mathscr{P}_N (y, \psi) \to \mathscr{A}^*(y, \psi)$ for all (y, ψ) in $\tilde{\mathscr{D}}$ is equivalent to the statement that $\mathscr{A}_N(\psi(0), \psi) = \mathscr{A}_N \mathscr{P}_N(\psi(0), \psi) \to \mathscr{A}(\psi(0), \psi)$ for all $(\psi(0), \psi)$ in the $\mathscr{D}$ of Lemma 7.1. This can be seen by recalling (7.4)–(7.6), comparing (7.10) and (7.12), and noting how nicely the term $(N/r)A_1^T$ in A^{*N} works out in light of (2.8) and how the $n \times n$ identity matrix in the first row (n rows actually) of A^{*N} yields the $\psi(0)$ in (2.9).

For $D \neq 0$, the matrix function D defines a bounded operator $\hat{D}: Z \to Z$ according to $\hat{D}(x, \phi) = (\int_{-r}^{0} D(\theta)\phi(\theta)\, d\theta, 0)$. In the approximating operators $\mathscr{A}_N$, we have approximated $\hat{D}$ by $\hat{D}\mathscr{P}_N$, and the corresponding term in $\mathscr{A}_N^*$ is $(\hat{D}\mathscr{P}_N)^* = \mathscr{P}_N \hat{D}^*$. Since $\mathscr{P}_N \to I$ strongly, $(\hat{D}\mathscr{P}_N)^* \to \hat{D}^*$ strongly. Therefore, for $D \neq 0$, $\mathscr{A}_N^* z \to \mathscr{A}^* z$ for $z \in \tilde{\mathscr{D}}$. Also, it is instructive to compare the term $D^T y$ in (2.9) with the first column of A^{*N}.

Finally, since $\mathscr{A}^*$ generates a C_0-semigroup, for $\operatorname{Re}\lambda$ sufficiently large, λ is in the resolvent set of $\mathscr{A}^*$ so that range $(\mathscr{A}^* - \lambda) = Z$. Then, if D is continuous, (2.9) shows that $(\mathscr{A}^* - \lambda)\tilde{\mathscr{D}} = \{(x, \phi) \in Z: \phi \text{ is continuous}\}$, which is dense in Z. For D only L_2, we can approximate D by a sequence of continuous D_c's, each of which results in an $\mathscr{A}_c^*$, and $\|\mathscr{A}_c^* - \mathscr{A}^*\|_{\mathscr{L}(Z,Z)} = \|D_c - D\|_{L_2}$. Hence, for λ in the resolvent set of $\mathscr{A}^*$ and $\|D_c - D\|_{L_2}$ sufficiently small, λ is in the resolvent set of $\mathscr{A}_c^*$ and $(\mathscr{A}_c^* - \lambda)\tilde{\mathscr{D}}$ is dense. Now it is easy to show that $(\mathscr{A}^* - \lambda)\tilde{\mathscr{D}}$ is dense by letting $\|D_c - D\|_{L_2} \to 0$. □

Since $\|T_N^*(t)\| = \|T_N(t)\|$, Theorem 7.2 with Theorem 7.1 and Lemma 7.2 yields the following.

THEOREM 7.3. *For* $t \geqq 0$ *and* $z \in Z$, $T_N^*(t)z \to T^*(t)z$ *uniformly in* t *for* t *in bounded intervals.*

In [4], Banks and Burns have indicated the modifications of the foregoing formulas and arguments needed to obtain a strongly convergent sequence of approximating semigroups $T_N(\cdot)$ for the general hereditary system (2.1). After such modifications, our arguments can be modified to show that the adjoint semigroups $T_N^*(\cdot)$ also converge strongly for the general problem. In particular, if (y, ψ) is in the set $\tilde{\mathscr{D}}$ of Theorem 7.2, then ψ must have the required jumps at the points $-h_i$ (see (2.7)).

7.2. Approximation for optimal control on a finite interval. For our Nth optimal control problem, we define

$$B_N = \mathscr{P}_N B = B, \tag{7.13}$$

$$Q_N = \mathscr{P}_N Q \mathscr{P}_N = Q, \tag{7.14}$$

$$G_N = \mathscr{P}_N G \mathscr{P}_N. \tag{7.15}$$

If G has the form $G(x, \phi) = (G_0 x, 0)$, then $G_N = G$. For B_N and Q_N, we have the respective matrix representations B^N and Q^N:

$$B^N = \begin{bmatrix} B_0 \\ 0 \\ \vdots \\ \vdots \end{bmatrix}, \qquad Q^N = \begin{bmatrix} Q_0 & 0 & \cdots \\ 0 & & 0 \\ \vdots & & \ddots \\ \vdots & & \end{bmatrix}. \tag{7.16}$$

Also, we denote the matrix representation of G_N by $\tilde{G}^N$, and, if $G(x, \phi) = (G_0 x, 0)$, then $\tilde{G}^N$ looks like Q^N with Q_0 replaced by G_0.

As in § 6, we denote the solution of the Riccati integral equation for the Nth approximate problem by $\Pi_N(\cdot)$. From the Nth versions of (3.22) and (3.24), we see that the minimum value of the cost functional for the Nth problem with initial time s and initial state $z(s)$ is $\langle \Pi_N(s)z(s), z(s)\rangle_Z$. Then, from (7.14) and (7.15) and the fact that Z_N and $Z_N^\perp$ reduce $T_N(\cdot)$, we see that for $z(s) \in Z_N^\perp$ the control $u = 0$ is optimal, since it yields zero cost. Hence the null space of $\Pi_N(s)$ contains $Z_N^\perp$ for $s \leqq t_f$.

Since $\mathscr{A}_N$ and $\mathscr{A}_N^*$ are bounded, we can differentiate the Nth version of either (3.14) or (3.24) to obtain the Riccati differential equation

$$\begin{aligned} &\dot{\Pi}_N(t) = -\mathscr{A}_N^* \Pi_N(t) - \Pi_N(t)\mathscr{A}_N + \Pi_N(t) B_N R^{-1} B_N^* \Pi_N(t) - Q_N, \qquad t \leqq t_f, \\ &\Pi_N(t_f) = G_N. \end{aligned} \tag{7.17}$$

We emphasize that (7.17) involves operators on the infinite dimensional space Z. For numerical computation, we need a Riccati differential equation involving finite dimensional matrices. To this end, we denote the matrix representation of $\Pi_N(t)|_{Z_N}$ by $\tilde{P}^N(t)$, and note that the matrix representation of B_N^* is B^{N^T}. Then (7.17) yields

$$\begin{aligned} &\dot{\tilde{P}}^N(t) = -A^{*N}\tilde{P}^N(t) - \tilde{P}^N(t)A^N + \tilde{P}^N(t)B^N R^{-1} B^{N^T}\tilde{P}^N(t) - Q^N, \qquad t \leqq t_f, \\ &\tilde{P}^N(t_f) = \tilde{G}^N. \end{aligned} \tag{7.18}$$

Remark 7.1. The operator $\Pi_N(t)$ is self-adjoint; however, in general the matrix $\tilde{P}^N(t)$ is not symmetric.

The matrix representation of $\Pi_N^*(t)$ is $W^{N^{-1}}\tilde{P}^{N^T}(t)W^N$, and, since $\Pi_N^*(t) = \Pi_N(t)$,

$$W^{N^{-1}}\tilde{P}^{N^T}(t)W^N = \tilde{P}^N(t). \tag{7.19}$$

From (7.19), we obtain

$$(W^N\tilde{P}^N(t))^T = \tilde{P}^{N^T}(t)W^N = W^N\tilde{P}^N(t). \tag{7.20}$$

For $t \leqq t_f$, we define

$$P^N(t) = W^N\tilde{P}^N(t), \tag{7.21}$$

which (7.20) shows to be symmetric. Also $P^N(t)$ is nonnegative because $\Pi_N(t)$ is nonnegative. Upon multiplying (7.18) on the left by W^N and noting $W^N A^{*N} = A^{N^T}W^N$, $B^{N^T}W^{N^{-1}} = B^{N^T}$ and $W^N Q^N = Q^N$, we obtain the Riccati differential equation to be solved for the symmetric $n(N+1) \times n(N+1)$ matrix $P^N(t)$:

$$\begin{aligned} &\dot{P}^N(t) = -A^{N^T}P^N(t) - P^N(t)A^N + P^N B^N R^{-1} B^{N^T} P^N - Q^N, \qquad t \leqq t_f, \\ &P^N(t_f) = G^N, \end{aligned} \tag{7.22}$$

where $G^N = W^N\tilde{G}^N$ is symmetric and nonnegative.

Next we will consider the procedure for computing the Nth feedback control law for the hereditary system once (7.22) has been solved for the matrix $P^N(t)$. The operator $\Pi_N(t)$ has the form (see (3.25))

$$\Pi_N(t) = \begin{bmatrix} \Pi_N^{00}(t) & \Pi_N^{01}(t) \\ \Pi_N^{10}(t) & \Pi_N^{11}(t) \end{bmatrix}, \qquad t \leqq t_f, \tag{7.23}$$

where $\Pi_N^{00}(t)$ is a real nonnegative, symmetric $n \times n$ matrix, $\Pi_N^{10}(t)$ is a real $n \times n$ matrix function $\Pi_N^{10}(t, \cdot)$ on $[-r, 0]$, $\Pi_N^{01}(t) = \Pi_N^{10}(t)^*$ and

$$\Pi_N^{01}(t)\phi = \int_{-r}^{0} \Pi_N^{10}(t, \theta)^T \phi(\theta)\, d\theta, \qquad \phi \in L_2(-r, 0; R^n), \tag{7.24}$$

and $\Pi_N^{11}(t)$ is a nonnegative, self-adjoint bounded linear operator on $L_2(-r, 0; R^n)$. Let us write

$$\tilde{P}^N(t) = \begin{bmatrix} \tilde{P}_{00}^N(t) & P_{01}^N(t) & \cdots & \tilde{P}_{0N}^N(t) \\ \tilde{P}_{10}^N(t) & \tilde{P}_{11}^N(t) & & \vdots \\ \vdots & & \ddots & \\ \tilde{P}_{N0}^N(t) & & \cdots & \tilde{P}_{NN}^N(t) \end{bmatrix}, \tag{7.25}$$

where each $\tilde{P}_{ij}^N$ is an $n \times n$ matrix. Similarly, we write

$$P^N(t) = \begin{bmatrix} P_{00}^N(t) & P_{01}^N(t) & \cdots & P_{0N}^N(t) \\ P_{10}^N(t) & P_{11}^N(t) & & \vdots \\ \vdots & & \ddots & \\ P_{N0}^N(t) & & \cdots & P_{NN}^N(t) \end{bmatrix}. \tag{7.26}$$

Since $\tilde{P}^N(t)$ is the matrix representation of $\Pi_N(t)$ and $P^N(t) = W^N \tilde{P}^N(t)$, we have

$$\Pi_N^{00}(t) = \tilde{P}_{00}^N(t) = P_{00}^N(t), \qquad t \leqq t_f. \tag{7.27}$$

Of course, $\Pi_N^{00}(t)$ is symmetric and nonnegative. Recalling (7.8), we see

$$\Pi_N^{10}(t) = \sum_{j=1}^{N} \tilde{P}_{j0}^N(t) \chi_j^N, \qquad t \leqq t_f. \tag{7.28}$$

Then, using $\tilde{P}^N(t) = W^{N-1} P^N(t)$ and taking transposes in (7.28), we obtain the kernel for (7.24):

$$\Pi_N^{10}(t)^T = \sum_{j=1}^{N} \frac{N}{r} P_{0j}^N(t) \chi_j^N, \qquad t \leqq t_f. \tag{7.29}$$

The Nth feedback control law for the hereditary system is then

$$\begin{aligned} u_N(t) &= -R^{-1} B_N^* \Pi_N(t) z(t) \\ &= -R^{-1} B_0^T \left[P_{00}^N(t) x(t) + \int_{-r}^{0} \left(\sum_{j=1}^{N} \frac{N}{r} P_{0j}^N(t) \chi_j^N(\theta) \right) x(t+\theta)\, d\theta \right], \qquad t \leqq t_f. \end{aligned} \tag{7.30}$$

By Theorem 6.1, $\Pi_N^{00}(t) = P_{00}^N(t)$ converges to $\Pi^{00}(t)$ uniformly in bounded t-intervals, and the elements of $\Pi_N^{10}(t)$ converge in $L_2(-r, 0)$ to the elements of $\Pi^{10}(t)$, uniformly in bounded t-intervals. Also, Theorem 6.2 applies to the control law of (7.30).

Now suppose $G(x, \phi) = (G_0 x, 0)$, so that

$$\tilde{G}^N = G^N = \begin{bmatrix} G_0 & 0 & \cdots \\ 0 & 0 & \\ \vdots & & \ddots \end{bmatrix}. \tag{7.31}$$

Then, by Theorem 5.1, $\operatorname{tr} \Pi(t) < \infty$ for $t \leqq t_f$, and, by Theorem 6.3 and Corollary 6.1, $\Pi_N(t)$ converges in trace norm to $\Pi(t)$ and $\operatorname{tr} \Pi_N(t)$ converges to $\operatorname{tr} \Pi(t)$. Since $Z_N^{\perp} \subset \mathcal{N}(\Pi_N(t))$, $\operatorname{tr} \Pi_N(t) = \operatorname{tr} (\Pi_N(t)|_{Z_N})$, and, from property 5.1, we see

$$\operatorname{tr} \Pi_N(t) = \operatorname{tr} \tilde{P}^N(t) = \operatorname{tr} (W^{N-1} P^N(t)), \tag{7.32}$$

where $\operatorname{tr} \tilde{P}^N(t)$ is, as usual, the sum of the diagonal elements of the matrix $\tilde{P}^N(t) = W^{N-1} P^N(t)$.

7.3. Approximation for optimal control on the infinite interval. Now we consider the sequence of Riccati algebraic equations

$$\mathcal{A}_N^*\Pi_N + \Pi_N\mathcal{A}_N - \Pi_N B_N R^{-1} B_N^*\Pi_N + Q_N = 0 \tag{7.33}$$

corresponding to the sequence of approximate optimal control problems. Recall (7.13) and (7.14), and that, when a nonnegative, self-adjoint solution of (7.33) exists, $S_N(\cdot)$ denotes the semigroup generated by $\mathcal{A}_N - BR^{-1}B^*\Pi_N$. We would like to say that, if there exists a nonnegative, self-adjoint solution Π of the Riccati algebraic equation for the hereditary system, then, for N sufficiently large, there exists a nonnegative, self-adjoint solution Π_N of the Riccati algebraic equation for the Nth approximate problem and there exist positive constants M and β (independent of N) for which (6.29) and (6.30) hold. To make this statement, we must assume the following conjecture about the semigroup approximation scheme of this section.

Conjecture 7.1. If the semigroup $T(\cdot)$ is uniformly exponentially stable, then there exist positive constants M' and β' such that, for N sufficiently large,

$$\|T_N(t)|_{Z_N}\| \leqq M' e^{-\beta' t}, \qquad t \geqq 0. \tag{7.34}$$

Remark 7.2. The present author has tried repeatedly but unsuccessfully to prove this conjecture–even for the no-delay case where $A_1 = A_2 = \cdots = A_\nu = 0$, $D = 0$, but $r > 0$. While not all semigroup approximation schemes have this property, the present scheme certainly seems to. Indeed, considering the present approximation scheme from the point of view of approximating a delay differential equation by a chain of ordinary differential equations, Repin [36] gave decay rates, uniform in N, for the approximating systems. While Repin did not work with semigroups or in the state space Z, his Theorem 3.1 seems to yield Conjecture 7.1. However, in several attempts, the present author has not been able to decipher the English translation of Repin's paper sufficiently to say unequivocally that Conjecture 7.1 has been proved. Even so, *for the rest of this section we will assume that Conjecture* 7.1 *holds.*

There exists a nonnegative, self-adjoint solution Π_N of the Riccati algebraic equation (7.33) for the Nth approximate problem if and only if there exists an admissible control for each initial condition $z \in Z$. (See Definitions 4.1 and 4.2, Theorem 4.1, Remark 6.1 and [19].) Such a solution could be defined arbitrarily on $Z_N^\perp$ because Z_N and $Z_N^\perp$ reduce $\mathcal{A}_N$, $B_N^*R^{-1}B_N$ and Q_N and $\mathcal{A}_N|_{Z_N^\perp} = \mathcal{A}_N^*|_{Z_N^\perp} = B_N^*R^{-1}B_N|_{Z_N^\perp} = Q_N|_{Z_N^\perp} = 0$. However, the appropriate solution for the Nth approximate optimal control problem satisfies $\Pi_N|_{Z_N^\perp} = 0$, since $u = 0$ is an admissible control for $z \in Z_N^\perp$ and we want the minimum value of the performance index to be $\langle \Pi_N z, z\rangle_Z$. If we replaced $\mathcal{A}_N$ by $\mathcal{A}_N - \beta(I - \mathcal{P}_N)$, then the only solution of (7.33) in $\mathcal{L}(Z_N^\perp, Z_N^\perp)$ would be 0, and the problem on Z_N would be unchanged. From here on $\Pi_N|_{Z_N^\perp} = 0$ will be implicit in all references to Π_N.

The only significant questions of existence and uniqueness of solutions to (7.33) pertain then to the optimal control problem on Z_N. From finite dimensional control theory, we know that a nonnegative, self-adjoint solution of (7.33) in $\mathcal{L}(Z_N, Z_N)$ is unique if the pair $(Q^{N^{1/2}}, A^N)$ is observable. Also, if $(Q^{N^{1/2}}, A^N)$ is observable, there exists a nonnegative solution of (7.33) in $\mathcal{L}(Z_N, Z_N)$ if and only if (A^N, B^N) is stabilizable, and such a solution is positive definite.

Suppose now that there exists a nonnegative, self-adjoint solution of the Riccati algebraic equation (4.2) for the hereditary control problem, or equivalently (Theorem 4.1 and Corollary 4.1) the hereditary system (2.1) is stabilizable. As in the previous sections, $S(\cdot)$ is the semigroup generated by $\mathcal{A} - BR^{-1}B^*\Pi$, and, according to Theorem 4.1, $S(\cdot)$ is uniformly exponentially stable.

Let us approximate $S(\cdot)$ in the same way we have approximated $T(\cdot)$. In (7.6) and (7.7), A_0 is replaced by $A_0 - B_0R^{-1}B_0^T\Pi^{00}$ and $D(\cdot)$ is replaced by $D(\cdot) - B_0R^{-1}B_0^T\Pi^{10}(\cdot)^T$. The result is an operator $\bar{\mathscr{A}}_N$, and, in view of (7.4), we have

$$\bar{\mathscr{A}}_N = \mathscr{A}_N - BR^{-1}B^*\Pi\mathscr{P}_N. \tag{7.35}$$

Thus, we define $\bar{S}_N(t) = e^{\bar{\mathscr{A}}_N t}$. Since, according to Theorem 4.1, $S(\cdot)$ is uniformly exponentially stable, Conjecture 7.1 implies the existence of positive constants M and β such that, for N sufficiently large,

$$\|\bar{S}_N(t)|_{Z_N}\| \leqq M e^{-\beta t}, \qquad t \geqq 0. \tag{7.36}$$

THEOREM 7.4. *If the hereditary system is stabilizable, then, for N sufficiently large, there exists a nonnegative, self-adjoint solution* Π_N *of* (7.33) *and we have*

$$\langle \Pi_N z, z\rangle_Z \leqq \frac{M^2}{2\beta}(\|Q_0\| + \|B_0\|^2\|\Pi\|^2\|R\|)\|z\|^2, \qquad z \in Z, \tag{7.37}$$

where M and β *are the constants in* (7.36).

Proof. For the Nth optimal control problem with initial condition z, take the feedback control $\bar{u}_N(t) = -R^{-1}B^*\Pi\mathscr{P}_N z(t)$. The resulting state then is given by $\bar{z}_N(t) = \bar{S}_N(t)z$, and the value of the performance index is

$$J(z, \bar{u}_N) = \int_0^\infty (\langle Q\bar{z}_N(t), \bar{z}_N(t)\rangle_Z + \langle R\bar{u}_N(t), \bar{u}_N(t)\rangle_{R^m})\, dt, \tag{7.38}$$

which is not greater than the right side of (7.37). Since the minimum value of the performance index is $\langle \Pi_N z, z\rangle_Z$, (7.37) follows. □

Recalling Theorem 6.7 (and assuming Conjecture 7.1), we can now say that, if there exists a nonnegative, self-adjoint solution Π of the Riccati algebraic equation for the hereditary system, then Π_N exists for N sufficiently large and Π_N converges weakly to Π. To prove that Π_N converges strongly and in trace norm requires more work. We will need the following lemmas.

LEMMA 7.3. *For* $f \in L_2(0, \infty)$ *and* $\alpha > 0$, *define* $g : [0, \infty) \to R$ *by*

$$g(t) = \int_0^t e^{-\alpha(t-s)} f(s)\, ds. \tag{7.39}$$

Then $g \in L_2(0, \infty)$ *and*

$$\int_0^\infty g^2(t)\, dt \leqq \frac{1}{\alpha^2}\int_0^\infty f^2(t)\, dt. \tag{7.40}$$

Proof. This is a special case of a more general result on convolutions (see [17, p. 951]). □

The next lemma is a special case of a result proved by Datko [8] for linear evolution operators on Banach spaces.

LEMMA 7.4. *Suppose that* $T(\cdot)$ *is a* C_0*-semigroup on a Hilbert space Z, and that there exist positive constants* M_1, M_2 *and* α_1 *such that*

$$\|T(t)\| \leqq M_1 e^{\alpha_1 t}, \qquad t \geqq 0 \tag{7.41}$$

and

$$\int_0^\infty \|T(t)z\|^2\, dt \leqq M_2\|z\|^2 \tag{7.42}$$

for each $z \in Z$. Then

$$\|T(t)\| \leqq M_3 e^{-\alpha_2 t}, \qquad t \geqq 0, \tag{7.43}$$

where

$$M_3 = 4M_1(M_2+1)(\alpha_1+1) \quad \textit{and} \quad \alpha_2 = (32M_1^2(M_2+1)^2(\alpha_1+1)^2)^{-1}. \tag{7.44}$$

Datko did not give M_3 and α_2 explicitly in terms of M_1, M_2 and α_1, but a rephrasing of his proof shows that (7.43) holds with the M_3 and α_2 of (7.44). While this M_3 and α_2 are hardly the sharpest possible bounds, they serve our purpose here, which is to give a decay rate for $T(\cdot)$ in terms of M_1, α_1 and M_2 only. This is essential to establishing a uniform decay rate for the semigroups $S_N(\cdot)$.

Consider the case of our hereditary differential equation in which $A_0 = -I$, $A_1 = A_2 = \cdots = A_\nu = 0$, and $D = 0$. Since $\{(0, \phi): \phi \in L_2(-r, 0; R^n)\}$ is invariant under each $\mathcal{A}_N$, we can define a sequence of semigroups $V_N(\cdot)$ on $L_2(-r, 0; R^n)$ by

$$(0, V_N(t)\phi) = T_N(t)(0, \phi), \qquad t \geqq 0, \quad N \geqq 1. \tag{7.45}$$

The matrix representing the generator of $V_N(\cdot)$ is just A^N with the first row and column deleted. Clearly, the semigroups $T(\cdot)$ and $T_N(\cdot)$, $N \geqq 1$, are uniformly exponentially stable, so that Conjecture 7.1 implies the existence of positive constants M' and β' such that

$$\|V_N(t)\| \leqq M' e^{-\beta' t}, \qquad t \geqq 0, \quad N \geqq 1. \tag{7.46}$$

THEOREM 7.5. *If the hereditary system is stabilizable and Q_0 is positive definite, then there exist positive constants M and β such that, for N sufficiently large,*

$$\|S_N(t)|_{Z_N}\| \leqq M e^{-\beta t}, \qquad t \geqq 0. \tag{7.47}$$

Proof. For $z \in Z_N$, let $z_N(t) = (w_0^N(t), \phi_N(t)) = S_N(t)z$, $t \geqq 0$, where $\phi_N(t) = \sum_{j=1}^N w_j^N(t)\chi_j$ (see 7.8) and $w_j^N(t) \in R^n$, $0 \leqq j \leqq N$; i.e., as in § 6, $z_N(t)$ is the optimal trajectory in the Nth approximate control problem. We emphasize that $z_N(t) \in Z_N \subset Z$, $\phi_N(t) \in L_2(-r, 0; R^n)$ and $w_j^N(t) \in R^n$. In every instance below, $\|\cdot\|$ indicates the appropriate norm. We then have

$$\begin{aligned} \int_0^\infty \|w_0^N(t)\|^2 \, dt &\leqq \mu^{-1} \int_0^\infty \langle Q_0 w_0^N(t), w_0^N(t)\rangle_{R^n} \, dt \\ &\leqq \mu^{-1} \int_0^\infty \langle (Q + \Pi_N B R^{-1} B^* \Pi_N) z_N(t), z_N(t)\rangle_Z \, dt \\ &\leqq \mu^{-1} \langle \Pi_N z, z\rangle_Z, \end{aligned} \tag{7.48}$$

where μ is the minimum eigenvalue of Q_0.

We can write $\phi_N(t)$ as

$$\phi_N(t) = V_N(t)\phi_N(0) + \int_0^t V_N(t-s) \frac{N}{r} w_0^N(s)\chi_1 \, ds, \qquad t \geqq 0. \tag{7.49}$$

If we write the integral in (7.49) as

$$\int_0^t V_N(t-s) \frac{N}{r} w_0^N(s)\chi_1 \, ds = \sum_{j=1}^N \hat{w}_j^N(t)\chi_j, \qquad t \geqq 0, \tag{7.50}$$

we have (differentiate the integral in (7.50) and look at the generator of $V_N(\cdot)$)

$$\dot{\hat{w}}_j^N(t) = -\frac{N}{r} \hat{w}_j^N(t) + \frac{N}{r} \hat{w}_{j-1}^N(t), \qquad t \geqq 0, \quad 1 \leqq j \leqq N, \tag{7.51}$$

where $\hat{w}_0^N(\cdot) = w_0^N(\cdot)$ and

$$\hat{w}_j^N(0) = 0, \qquad 1 \leqq j \leqq N. \tag{7.52}$$

Hence,

$$\hat{w}_j^N(t) = \int_0^t e^{-N(t-s)/r} \frac{N}{r} \hat{w}_{j-1}^N(s)\, ds, \qquad t \geqq 0, \quad 1 \leqq j \leqq N, \tag{7.53}$$

and

$$\|\hat{w}_j^N(t)\| \leqq \int_0^t e^{-N(t-s)/r} \frac{N}{r} \|\hat{w}_{j-1}^N(s)\|\, ds, \qquad t \geqq 0, \quad 1 \leqq j \leqq N. \tag{7.54}$$

From (7.48), (7.54) and Lemma (7.3), we have

$$\int_0^\infty \|\hat{w}_j(t)\|^2\, dt \leqq \mu^{-1} \|\Pi_N\| \cdot \|z\|^2, \qquad 0 \leqq j \leqq N. \tag{7.55}$$

Since

$$\left\| \sum_{j=1}^N \hat{w}_j^N(t) \chi_j \right\|^2 = \sum_{j=1}^N \frac{r}{N} \|\hat{w}_j^N(t)\|^2, \tag{7.56}$$

from (7.50) and (7.55), we have

$$\int_0^\infty \left\| \int_0^t V_N(t-s) \frac{N}{r} w_0^N(s) \chi_1\, ds \right\|^2 dt \leqq r\mu^{-1} \|\Pi_N\| \cdot \|z\|^2. \tag{7.57}$$

Then, since $\|\phi_N(0)\| \leqq \|z\|$, (7.46), (7.49) and (7.57) yield

$$\int_0^\infty \|\phi_N(t)\|^2\, dt \leqq 2\left(\frac{M'^2}{(2\beta')} + r\mu^{-1} \|\Pi_N\| \right) \|z\|^2, \tag{7.58}$$

where M' and β' are independent of N for N sufficiently large. From (7.37), (7.48) and (7.58), we have, for N sufficiently large,

$$\int_0^\infty \|z_N(t)\|^2\, dt = \int_0^\infty \|S_N(t) z\|^2\, dt \leqq M_2 \|z\|^2, \qquad z \in Z_n, \tag{7.59}$$

where M_2 does not depend on N. Also, because we have subtracted $BR^{-1}B^*\Pi_N$ from the generator of $T_N(\cdot)$ to obtain the generator of $S_N(\cdot)$, Lemma 7.2 and (7.37) imply the existence of constants M_1 and α_1 such that

$$\|S_N(t)\| \leqq M_1 e^{\alpha_1 t}, \qquad t \geqq 0, \quad N \geqq 1. \tag{7.60}$$

The theorem now follows from Lemma 7.4. □

COROLLARY 7.1. *If the hereditary system* (7.2) *is stabilizable and* Q_0 *is positive definite, then, for* N *sufficiently large, there exists a bounded nonnegative, self-adjoint solution* Π_N *of the Riccati algebraic equation for the* Nth *approximate optimal control problem and* Π_N *converges in trace norm to* Π, *the bounded nonnegative, self-adjoint solution of the Riccati algebraic equation for the original hereditary optimal control problem.*

Proof. Replacing $\mathcal{A}_N$ with $\mathcal{A}_N - \beta(I - P_N)$ (see the paragraph following Remark 7.2), we obtain $\|S_N(t)|_{Z_N^\perp}\| \leqq e^{-\beta t}$, so that (7.47) implies (6.29). Thus Theorems 7.4 and 7.5 yield the hypotheses of Theorem 6.9. □

Remark 7.3. From Theorem 4.1 and Remark 6.1, we see that positive definiteness of Q_0 implies that $\Pi_N|_{Z_N}$ is positive definite and is the unique nonnegative, self-adjoint

solution of (7.33) in $\mathscr{L}(Z_N, Z_N)$. It seems most likely that Theorem 7.5 remains true if the requirement that Q_0 be positive definite is replaced by Hypothesis 4.1; however, the positive definiteness of Q_0 is essential to the proof here.

Of course, we have the steady-state versions of (7.17)–(7.32) ($G_0 = 0$); i.e., the same equations except that none of the operators depend on t. Especially important for engineering applications are the implications of Theorem 6.8 for the sequence of feedback control laws given by the time-invariant version of (7.30).

Remark 7.4. We need Conjecture 7.1 for Theorems 7.4 and 7.5 and Corollary 7.1. Without this conjecture, all is not lost. According to Theorem 6.7, we still can say that Π_N converges weakly to Π if $\|\Pi_N\|$ is bounded in N. Since Π_N is self-adjoint and nonnegative, $\|\Pi_N\|$ is equal to the maximum eigenvalue of $\tilde{P}^N$, the matrix representation of Π_N. Also (see § 5), we have $\|\Pi_N\| \leqq \operatorname{tr} \Pi_N = Tr\, \tilde{P}^N$. Thus we have an easy numerical test for weak convergence of Π_N. However, for stabilizable hereditary systems, the conjecture implies the much stronger convergence and stability statements in Theorem 7.5, Corollary 7.1, and Theorem 6.8.

The following theorem is a generalization of [38, Proposition 3.1]. The further generalization to include (2.1) should be obvious. As in the proof of Theorem 7.2, we define $\hat{D} \in \mathscr{L}(Z, Z)$ by

$$\hat{D}(x, \phi) = \left(\int_{-r}^{0} D(\theta)\phi(\theta)\, d\theta, 0 \right), \qquad (x, \phi) \in Z. \tag{7.61}$$

THEOREM 7.6. *If the pair (A_0, B_0) is stabilizable and the range of B_0 contains the range of A_1 and the range of $\hat{D}$, then* (7.2) *is stabilizable and the pair (A^N, B^N) is stabilizable for $N \geqq 1$.*

Proof. We may assume that B_0 is one-to-one, so that $B_0^T B_0$ is positive definite. By hypothesis, there exists an $m \times n$ matrix K_0 such that all the eigenvalues of $A_0 - B_0 K_0$ have negative real parts, so that the feedback control

$$u(t) = -K_0 x(t) - (B_0^T B_0)^{-1} B_0^T \left[A_1 x(t-r) + \int_{-r}^{0} D(\theta) x(t+\theta)\, d\theta \right] \tag{7.62}$$

makes (7.2) asymptotically stable. Thus, for each $(x(0), x_0) \in Z$, the $u(\cdot)$ in (7.62) is an admissible control according to Definition 4.1, and, by Corollary 4.1, (7.2) is stabilizable in the sense of Definition 2.2.

Similarly, we can define a stabilizing feedback control for the pair (A^N, B^N) for $N \geqq 1$. Recall that the terms D_j^N in A^N represent $\hat{D}\mathscr{P}_N$ and range $\hat{D}\mathscr{P}_N \subset$ range $\hat{D} \subset$ range B_0. □

For Theorem 7.6 itself, we do not need Conjecture 7.1. However, even for systems which satisfy the hypothesis of Theorem 7.6, we still need Conjecture 7.1 to make the statements in Theorems 7.4 and 7.5 and Corollary 7.1 because Theorem 7.6 does not yield the uniform bound in (7.47), which Theorem 6.9 requires.

7.4. Computational aspects of the finite dimensional Riccati algebraic equations. From here on, we will assume that (7.2) and each of the approximating systems is stabilizable, and that Q_0 is positive definite. As in the case where $t_f < \infty$, we denote the matrix representing $\Pi_N|_{Z_N}$ by $\tilde{P}^N$ and define $P^N = W^N \tilde{P}^N$. It follows from (7.33) that $\tilde{P}^N$ and P^N are constant solutions of (7.18) and (7.22), respectively. In particular

$$A^{N^T} P^N + P^N A^N - P^N B^N R^{-1} B^{N^T} P^N + Q^N = 0. \tag{7.63}$$

From Remark 7.1, we know that the $n(N+1) \times n(N+1)$ matrix P^N is positive definite and is the unique nonnegative, symmetric solution of the Riccati matrix equation (7.63).

Ross and Flügge-Lotz [38] derived (7.63) using an approximation scheme for the hereditary system that is essentially the same as the scheme used here, but without using infinite dimensional Riccati equations. While we have not bothered with a formula for the kernels of the operators Π_N^{11} (see (7.23)) because the operators Π_N^{11} do not appear in the control laws, the interested reader can find such formulas in [37] and [38]. Ross and Flügge-Lotz did not indicate the sense in which the solutions to the finite dimensional Riccati equations and corresponding control laws can be expected to converge. Neither did they discuss computational aspects of the problem, although [37] concludes that "the primary limitation of the approximate solution method is therefore the computational burden of solving very high dimension steady-state Riccati equations." In the remainder of this section, we give some results which we hope will illuminate and diminish the computational burden.

Probably the most efficient method for solving (7.63) is the method given in [34] by Potter (see also [2, p. 354], [25, p. 250]), which involves an eigenvector decomposition of the matrix

$$\hat{A}^N = \begin{bmatrix} A^N & -C^N \\ -Q^N & -A^{NT} \end{bmatrix}, \tag{7.64}$$

where $C^N = B^N R^{-1} B^{NT}$. Compare $\hat{A}^N$ to the $\hat{\mathscr{A}}$ of (4.20). When the discussion in the last part of § 4 is applied to finite dimensional control problems (i.e., the case $r = 0$) the next theorem results (see [2], [25]). Recall that we are assuming that (A^N, B^N) is stabilizable and Q_0 is positive definite.

THEOREM 7.7. *The matrix $\hat{A}^N$ has no purely imaginary eigenvalues, and the spectrum of $\hat{A}^N$ is symmetric about the imaginary axis, the symmetry including algebraic and geometric multiplicities of eigenvalues. A complex number λ is an eigenvalue of $A^N - B^N R^{-1} B^{NT} P^N$ if and only if* $\operatorname{Re}\lambda < 0$ *and λ is an eigenvalue of $\hat{A}^N$.*

COROLLARY 7.2. *Since $A^N - B^N R^{-1} B^{NT} P^N = A^N - B^N R^{-1} B^{NT} \tilde{P}^N$ is the matrix representation of $(\mathscr{A}_N - B_N R^{-1} B_N^* \Pi_N)|_{Z_N}$, λ is an eigenvalue of the approximating closed-loop system on Z_N if and only if* $\operatorname{Re}\lambda < 0$ *and λ is an eigenvalue of $\hat{A}^N$.*

We will denote a generalized eigenvector of $\hat{A}^N$ by $\binom{x^N}{y^N}$, where x^N and y^N have the form

$$x^N = \begin{bmatrix} x_0^N \\ x_1^N \\ \vdots \\ x_N^N \end{bmatrix}, \qquad y^N = \begin{bmatrix} y_0^N \\ y_1^N \\ \vdots \\ y_N^N \end{bmatrix}, \tag{7.65}$$

where x_j^N and y_j^N are complex n-vectors for $0 \leqq j \leqq N$. Since $\hat{A}^N$ is a $2n(N+1) \times 2n(N+1)$ matrix, Theorem 7.7 implies that the total number of linearly independent generalized eigenvectors corresponding to the eigenvalues of $\hat{A}^N$ with negative real parts is exactly $n(N+1)$. The following theorem, which is implicit in the finite dimensional version of Theorem 4.9 and its proof, particularly (4.39), gives the Potter method for solving (7.63).

THEOREM 7.8. *Let*

$$\binom{x^{N,i}}{y^{N,i}}, \qquad 1 \leqq i \leqq n(N+1),$$

be an enumeration of linearly independent generalized eigenvectors corresponding to the eigenvalues of $\hat{A}^N$ with negative real parts, and define the $n(N+1) \times n(N+1)$ matrices

$$X^N = [x^{N,1} x^{N,2} \cdots x^{N,n(N+1)}] \quad \text{and} \quad Y^N = [y^{N,1} y^{N,2} \cdots y^{N,n(N+1)}]. \tag{7.66}$$

Then the unique nonnegative, symmetric solution of (7.63) *is*

$$P^N = Y^N X^{N-1}. \tag{7.67}$$

Note that, even though X^N and Y^N may be complex, P^N is real.

COROLLARY 7.3 (see Corollary 4.3). *Let* λ *be an eigenvalue of* $A^N - B^N R^{-1} B^{NT} P^N$ *with negative real part,* x^N *a complex* $n(N+1)$*-vector, and* k *a positive integer. Then* $(\lambda - (A^N - B^N R^{-1} B^{NT} P^N))^k x^N = 0$ *if and only if*

$$(\lambda - \hat{A}^N)^k \begin{pmatrix} x^N \\ P^N x^N \end{pmatrix} = 0.$$

Next we will derive a characteristic equation for $\hat{A}^N$, similar to (4.24). For $N \geqq 1$, we define an $n \times n$ matrix function of a complex number λ by

$$L_0^N(\lambda) = A_0 + \sum_{j=1}^{N} \left(\frac{N}{N+\lambda r}\right)^j \frac{r}{N} D_j^N + \left(\frac{N}{N+\lambda r}\right)^N A_1. \tag{7.68}$$

Now λ is an eigenvalue of $\hat{A}^N$ if and only if there is a nonzero vector $\binom{x^N}{y^N}$ such that

$$\hat{A}^N \begin{pmatrix} x^N \\ y^N \end{pmatrix} = \lambda \begin{pmatrix} x^N \\ y^N \end{pmatrix}, \tag{7.69}$$

which is equivalent to the following set of equations (recall (7.65)):

$$A_0 x_0^N + \sum_{j=1}^{N} \frac{r}{N} D_j^N x_j^N + A_1 x_N^N - C_0 y_0^N = \lambda x_0^N, \tag{7.70a}$$

$$\frac{N}{r} x_{j-1}^N - \frac{N}{r} x_j^N = \lambda x_j^N, \qquad 1 \leqq j \leqq N, \tag{7.70b}$$

$$-Q_0 x_0^N - A_0^T y_0^N - \frac{N}{r} y_1^N = \lambda y_0^N, \tag{7.70c}$$

$$-\frac{r}{N} D_j^{NT} y_0^N + \frac{N}{r} y_j^N - \frac{N}{r} y_{j+1}^N = \lambda y_j^N, \qquad 1 \leqq j \leqq N-1, \tag{7.70d}$$

$$-\left(\frac{r}{N} D_N^N + A_1\right)^T y_0^N + \frac{N}{r} y_N^N = \lambda y_N^N, \tag{7.70e}$$

where, as in (4.23), $C_0 = B_0 R^{-1} B_0^T$.

Suppose $\lambda \neq \pm N/r$ then, from (7.70b), we have

$$x_j^N = \left(\frac{N}{N+\lambda r}\right)^j x_0^N, \qquad 1 \leqq j \leqq N; \tag{7.71}$$

from (7.70d) and (7.70e), we have

$$y_j^N = \frac{r}{N}\left[\sum_{k=j}^{N} \left(\frac{N}{N-\lambda r}\right)^{k-j+1} \frac{r}{N} D_k^N + \left(\frac{N}{N-\lambda r}\right)^{N-j+1} A_1\right]^T y_0^N, \qquad 1 \leqq j \leqq N. \tag{7.72}$$

With (7.71), (7.70a) becomes

$$L_0^N(\lambda) x_0^N - C_0 y_0^N = \lambda x_0^N \tag{7.73}$$

and, with (7.72), (7.70c) becomes

$$-Q_0 x_0^N - L_0^{NT}(-\lambda) y_0^N = \lambda y_0^N. \tag{7.74}$$

We can write (7.73) and (7.74) as

$$\hat{\Delta}^N(\lambda)\begin{pmatrix} x_0^N \\ y_0^N \end{pmatrix} = 0, \tag{7.75}$$

where the $2n \times 2n$ matrix $\hat{\Delta}^N(\lambda)$ is given by

$$\hat{\Delta}^N(\lambda) = \lambda I - \begin{bmatrix} L_0^N(\lambda) & -C_0 \\ -Q_0 & -L_0^{N^T}(-\lambda) \end{bmatrix}. \tag{7.76}$$

We have then the following theorem.

THEOREM 7.9. *Suppose $\lambda \neq \pm N/r$. Then λ is an eigenvalue of $\hat{A}^N$ if and only if*

$$\det \hat{\Delta}^N(\lambda) = 0. \tag{7.77}$$

If λ satisfies (7.77), *then the corresponding eigenvectors of $\hat{A}^N$ are described by* (7.71)–(7.74).

Some comments are in order here. First, Theorem 7.9 holds regardless of whether (A^N, B^N) is stabilizable or Q_0 is positive definite. For the eigenvalues of A^N that are not equal to $\pm N/r$, we could use (7.70a) with $C_0 = 0$, (7.70b), and (7.71) to obtain

$$\det \Delta^N(\lambda) = 0, \tag{7.78}$$

where

$$\Delta^N(\lambda) = \lambda I - L_0^N(\lambda). \tag{7.79}$$

Banks and Burns derived (7.78) and (7.79) in [4].

It is quite instructive to compare $L_0^N(\lambda)$, $\Delta^N(\lambda)$ and $\hat{\Delta}^N(\lambda)$ to their counterparts for the hereditary system ((2.6), (4.23)) in light of the well-known result that

$$\left(\frac{N}{N+\lambda r}\right)^N \to e^{-\lambda r} \quad \text{as } N \to \infty. \tag{7.80}$$

In particular, as $N \to \infty$, $L_0^N(\lambda) \to L_0(\lambda)$ uniformly in compact subsets of the complex plane. Although we will not pursue the details here, standard methods of complex analysis can be used to show that, for each λ_0 satisfying $\det \hat{\Delta}(\lambda_0) = 0$, there is a sequence λ_N such that $\det \hat{\Delta}^N(\lambda_N) = 0$ and $\lambda_N \to \lambda_0$, and conversely, if $\det \hat{\Delta}(\lambda_0) \neq 0$, then there is a neighborhood $\mathcal{N}$ of λ_0 such that, for N sufficiently large, $\mathcal{N}$ contains no roots of $\det \hat{\Delta}^N(\lambda) = 0$. This comment parallels a corresponding comment by Banks and Burns concerning $\Delta^N(\lambda)$.

Now consider the equation

$$\left(\frac{N}{r}+\lambda\right)^{nN}\left(\frac{N}{r}-\lambda\right)^{nN} \det \hat{\Delta}(\lambda) = 0. \tag{7.81}$$

It follows from (7.68) and (7.76) that the left side of (7.81) is a polynomial of degree $2n(N+1)$ in λ. We expect that the left side of (7.81) is $\det(\lambda - \hat{A}^N)$. It appears that this result would follow from an argument quite similar to that used by Banks and Burns to prove $\det(\lambda - A^N) = (N/r + \lambda)^{nN} \det \Delta(\lambda)$. Banks and Burns based their derivation on manipulation of determinants instead of construction of eigenvectors, which we use here because the eigenvectors are needed for the solution of the Riccati equation. With Theorem 7.9 we certainly can say that any $\lambda \neq \pm N/r$ which satisfies (7.81) is an eigenvalue of $\hat{A}^N$, but there remain questions about multiplicities and possible roots of (7.81) equal to $\pm N/r$.

Anyone using the method of Theorem 7.8 to solve (7.63) should be aware of the conditions under which $-N/r$ is an eigenvalue of $\hat{A}^N$ and the form of the corresponding generalized eigenvectors. Recall that Theorem 7.7 says that the spectral characteristics of N/r as an eigenvalue of $\hat{A}^N$ are identical to those of $-N/r$ as an eigenvalue of $\hat{A}^N$. It is a straightforward exercise to use (7.70a–e), (7.72), (7.74) and their appropriate generalizations for generalized eigenvectors of $\hat{A}^N$ to prove the following.

THEOREM 7.10. *Suppose $(N/rI - L_0^N(N/r))$ is nonsingular. Then $-N/r$ is an eigenvalue of $\hat{A}^N$ if and only if $(A_1+(r/N)D_N^N)$ is singular. For each nonzero n-vector w such that $(A_1+(r/N)D_N^N)w=0$, there is a chain of N generalized eigenvectors $\binom{x^{N,k}}{y^{N,k}}$ satisfying*

$$\left(-\frac{N}{r}-\hat{A}^N\right)^k\binom{x^{N,k}}{y^{N,k}}=0, \qquad 1\leqq k\leqq N, \tag{7.82}$$

where

$$x_j^{N,k}=\begin{cases} w, & j=N-k+1,\\ 0, & \text{otherwise,}\end{cases} \qquad y^{N,k}=0, \qquad 1\leqq k\leqq N. \tag{7.83}$$

Furthermore, the only eigenvectors corresponding to $-N/r$ are given by (7.83) with $k=1$.

Note that $(r/N)D_N^N\to 0$ as $N\to\infty$, and that $\det(N/rI-L_0^N(N/r))\neq 0$ for N sufficiently large.

In general, it is possible for $\hat{A}^N$ to have generalized eigenvectors corresponding to $-N/r$ of rank greater than k; however, we can say the following about a class of problems that often arise from physical systems.

THEOREM 7.11. *Suppose that n is even, $D=0$, and A_0, A_1 and B_0 have the forms*

$$A_0=\begin{bmatrix}0 & I\\ A_{01} & A_{02}\end{bmatrix}, \quad A_1=\begin{bmatrix}0 & 0\\ A_{11} & A_{12}\end{bmatrix}, \quad B_0=\begin{bmatrix}0\\ B_{01}\end{bmatrix}, \tag{7.84}$$

where I (the identity), A_{01}, A_{02}, A_{11} and A_{12} are $n/2\times n/2$ matrices and B_{01} is $n/2\times n$. If $\det(A_{11}-(N/r)A_{12})\neq 0$, then $\hat{A}^N$ has no generalized eigenvector corresponding to $-N/r$ of rank greater than N.

Proof. Suppose there is a generalized eigenvector of rank $N+1$. Then we have

$$\binom{x^{N,N+1}}{y^{N,N+1}}\ni\left(-\frac{N}{r}-\hat{A}^N\right)\binom{x^{N,N+1}}{y^{N,N+1}}=\binom{x^{N,N}}{y^{N,N}}, \tag{7.85}$$

where $\binom{x^{N,N}}{y^{N,N}}$ is given by (7.83) with $k=N$ and w a nonzero n-vector such that $A_1w=0$. The set of n-vector equations represented by (7.85) then yield, in the same way that (7.70a–e) yielded (7.71) and (7.73)

$$-\left(\frac{N}{r}+A_0\right)x_0^{N,N+1}=-\frac{r}{N}w, \tag{7.86}$$

$$-\left(\frac{N}{r}+A_0\right)x_0^{N,N+1}-A_1x_N^{N,N+1}+C_0y_0^{N,N+1}=0, \tag{7.87}$$

where C_0 is given by

$$C_0=\begin{bmatrix}0 & 0\\ 0 & B_{01}R^{-1}B_{01}^T\end{bmatrix}. \tag{7.88}$$

Now we write $w=\binom{w_1}{w_2}$ where w_1 and w_2 are $n/2$-vectors. Since the first $n/2$ rows of the matrices A_1 and C_0 are all zero, (7.86) and (7.87) yield

$$\frac{N}{r}w_1+w_2=0. \tag{7.89}$$

Since $A_1 w = 0$,

$$A_{11} w_1 + A_{12} w_2 = 0. \tag{7.90}$$

Thus, if $\hat{A}^N$ has a generalized eigenvector corresponding to $-N/r$ of rank greater than N, we must have

$$\left(A_{11} - \frac{N}{r} A_{12}\right) w_1 = 0 \tag{7.91}$$

for some nonzero w_1, contradicting $\det (A_{11} - (N/r)A_{12}) \neq 0$. □

Let us summarize the computational implications of our results pertaining to the solution of the Riccati matrix equation (7.63) via Theorem 7.8. There are now two possible routes for computing the required generalized eigenvectors of the matrix $\hat{A}^N$ of (7.64). The first is to give $\hat{A}^N$ to a software package which computes eigenvalues and eigenvectors of matrices. Because some such packages handle only diagonalizable matrices, while others handle general matrices, Theorems 7.10 and 7.11 should guide the software selection. The second approach to computing the eigenvectors is to solve either (7.77) or (7.81) numerically for the eigenvalues of $\hat{A}^N$, solve (7.73) and (7.74) for the n-vectors x_0^N and y_0^N, and then construct the required generalized eigenvectors according to (7.71), (7.72) and (7.83). We are assuming here that the eigenvalues not equal to $\pm N/r$ are simple, and that there are no generalized eigenvectors corresponding to $-N/r$ of rank greater than N. If either assumption does not hold, the additional generalized eigenvectors can be obtained from equations similar to (7.70a–e), but the formulas are considerably more complicated than (7.71), (7.72) and (7.83).

For most of us, the choice between these two approaches amounts to the choice between using one canned program to compute the eigenvalues and eigenvectors of $\hat{A}^N$ or using another canned program to compute the roots to (7.77) or (7.81). We tried both approaches on the examples in the next section. In Example 1, $\hat{A}^N$ has no generalized eigenvectors (i.e., $\hat{A}^N$ is diagonalizable), and the first method—computing the eigenvalues and eigenvectors of $\hat{A}^N$—worked quite well. In Examples 2 and 3, $\hat{A}^N$ has generalized eigenvectors given by Theorem 7.10, and the programs available at UCLA did not compute these eigenvectors reliably, although they did compute the eigenvalues accurately. We tried several programs, including the Muller method [32], for computing the roots of (7.77) directly, but had little success. So, for Examples 2 and 3, we actually combined the two approaches, using a canned matrix program to compute the eigenvalues of $\hat{A}^N$ and then constructing the generalized eigenvectors according to Theorem 7.10.

8. Examples. In this section, we present numerical results for three examples of optimal control on the infinite interval. For each example, we solved the Riccati matrix equation (7.63) for the matrix P^N for several values of N. The Nth feedback control is given by the time-invariant version of (7.30):

$$u_N(t) = -R^{-1} B_0^T \left[\Pi_N^{00} x(t) + \int_{-r}^{0} \Pi_N^{10}(\theta)^T x(t+\theta)\, d\theta \right], \tag{8.1}$$

where Π_N^{00} and $\Pi_N^{10}(\theta)^T$ are given by the time-invariant versions of (7.27) and (7.29), respectively. For the examples here, Corollary 7.1 (with Conjecture 7.1) implies $\Pi_N^{00} \to \Pi^{00}$, $\Pi_N^{10} \to \Pi^{10}$ in $L_2(-r, 0; R^{n^2})$, and $\operatorname{tr} \Pi_N \to \operatorname{tr} \Pi$, where, as in (7.32),

$$\operatorname{tr} \Pi_N = \operatorname{tr} (W^{N-1} P^N). \tag{8.2}$$

Also, recall Remark 7.4.

The results here do not depend on the initial conditions $(x(0), x_0)$ because we are computing the feedback control law and the average minimum performance index discussed at the end of § 5. Here, we take $K = I$ in (5.9).

Example 8.1. We take $n = m = r = A_0 = A_1 = B_0 = Q_0 = R = 1$, so that (2.1) is the scalar differential equation

$$\dot{x}(t) = x(t) + x(t-1) + u(t) \tag{8.3}$$

and the performance index of (4.1) is

$$J((x(0), x_0), u) = \int_0^\infty (x^2(t) + u^2(t))\, dt. \tag{8.4}$$

For each N, Π_N^{00} is a scalar, $\Pi_N^{10^T} = \Pi_N^{10} \in L_2(-1, 0; R)$, and the Nth feedback control law is

$$u_N(t) = -\Pi_N^{00} x(t) - \int_{-1}^0 \Pi_N^{10}(\theta) x(t+\theta)\, d\theta. \tag{8.5}$$

We have the numbers in Table 8.1.

TABLE 8.1

$\Pi_{17}^{00} = 2.8260$, $\Pi_{29}^{00} = 2.8190$, $\Pi_{50}^{00} = 2.8148$, $\Pi_{74}^{00} = 2.8130$

θ	$\Pi_{17}^{10}(\theta)$	$\Pi_{29}^{10}(\theta)$	$\Pi_{50}^{10}(\theta)$	$\Pi_{74}^{10}(\theta)$
0.0	0.6684	0.6547	0.6469	0.6435
−0.1	0.7726	0.7169	0.7179	0.7273
−0.2	0.8467	0.8239	0.8209	0.8258
−0.3	0.9961	0.9508	0.9434	0.9607
−0.4	1.0822	1.1020	1.0895	1.1023
−0.5	1.2811	1.2822	1.2633	1.2694
−0.6	1.5220	1.4963	1.4693	1.4965
−0.7	1.6606	1.7501	1.7125	1.7315
−0.8	1.9802	2.0499	1.9987	2.0480
−0.9	2.3648	2.4033	2.3347	2.3748
−1.0	2.5852	2.6730	2.7284	2.7541

$\operatorname{tr} \Pi_{17} = 3.7742$, $\operatorname{tr} \Pi_{29} = 3.7978$, $\operatorname{tr} \Pi_{50} = 3.8161$, $\operatorname{tr} \Pi_{74} = 3.8265$

It is instructive to compare Π_N^{00} and $\Pi_N^{10}(-1)$ to see how closely we have approximated the boundary condition (4.14).

Example 8.2. We take $n = 2$, $m = r = R = 1$,

$$A_0 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

If we write $x(t) = (x_1(t), x_2(t))$, (2.1) is equivalent to

$$\ddot{x}_1(t) + x_1(t-1) = u(t), \tag{8.6}$$

and the performance index is

$$J((x(0), x_0); u) = \int_0^\infty (x_1^2(t) + \dot{x}_1^2(t) + u^2(t))\, dt. \tag{8.7}$$

The optimal control in feedback form is

$$(8.8)\quad u(t) = -\Pi^{00}_{,21}x_1(t) - \Pi^{00}_{,22}x_2(t) - \int_{-1}^{0} (\Pi^{10}_{,12}(\theta)x_1(t+\theta) + \Pi^{10}_{,22}(\theta)x_2(t+\theta))\, d\theta,$$

where $\Pi^{00}_{,ij}$ and $\Pi^{10}_{,ij}(\theta)$ are the ij elements of the matrices Π^{00} and $\Pi^{10}(\theta)$, respectively.

The Nth feedback control law is

$$(8.9)\quad u_N(t) = -\Pi^{00}_{N,21}x_1(t) - \Pi^{00}_{N,22}x_2(t) - \int_{-1}^{0} (\Pi^{10}_{N,12}(\theta)x_1(t+\theta) + \Pi^{10}_{N,22}(\theta)x_2(t+\theta))\, d\theta.$$

Note that the initial conditions

$$(8.10)\qquad x_1(0) = x_2(0) = 0 \quad \text{and} \quad x_1(\theta) = 0, \qquad -1 \leqq \theta \leqq 0,$$

yield $x_1(t) = x_2(t) = 0$, $t \geqq 0$, regardless of the initial function $x_2(\theta)$, $-1 \leqq \theta \leqq 0$. Hence for the initial conditions in (8.10) and any initial history $x_2(\cdot)$, the optimal control is $u(t) = 0$, $t \geqq 0$. Therefore, we must have

$$(8.11)\qquad \Pi^{10}_{,22} = 0.$$

Similarly, from the A^N for this example and (7.28)–(7.30), it can be seen that

$$(8.12)\qquad \Pi^{10}_{N,22} = 0, \qquad n \geqq 1.$$

Numerically, we have the results in Table 8.2.

TABLE 8.2

$$\Pi^{00}_{10} = \begin{bmatrix} 2.9525 & 1.3594 \\ 1.3594 & 1.9284 \end{bmatrix}, \quad \Pi^{00}_{14} = \begin{bmatrix} 2.9823 & 1.3761 \\ 1.3761 & 1.9370 \end{bmatrix}, \quad \Pi^{00}_{22} = \begin{bmatrix} 3.0109 & 1.3917 \\ 1.3917 & 1.9451 \end{bmatrix}.$$

θ	$\Pi^{10}_{10,12}(\theta)$	$\Pi^{10}_{14,12}(\theta)$	$\Pi^{10}_{22,12}(\theta)$
−0.0	−0.3309	−0.3168	−0.3038
−0.1	−0.3309	−0.3926	−0.4022
−0.2	−0.4371	−0.4736	−0.5094
−0.3	−0.5537	−0.6524	−0.6258
−0.4	−0.6813	−0.7505	−0.7517
−0.5	−0.8203	−0.8548	−0.8877
−0.6	−0.9713	−1.0825	−1.1119
−0.7	−1.1350	−1.2065	−1.2754
−0.8	−1.3119	−1.3375	−1.4510
−0.9	−1.5027	−1.6218	−1.6390
−1.0	−1.7081	−1.7754	−1.8399

$\operatorname{tr}\Pi_{10} = 5.7606$, $\operatorname{tr}\Pi_{14} = 5.8300$, $\operatorname{tr}\Pi_{22} = 5.8988$

For checking (4.14), we give

$$\Pi^{10}_{10}(-1) = \begin{bmatrix} -1.0900 & -1.7081 \\ 0 & 0 \end{bmatrix}, \quad \Pi^{10}_{14}(-1) = \begin{bmatrix} -1.1845 & -1.7754 \\ 0 & 0 \end{bmatrix},$$

$$\Pi^{10}_{22}(-1) = \begin{bmatrix} -1.2606 & -1.8399 \\ 0 & 0 \end{bmatrix}.$$

Example 8.3. We take $n = 2$, $m = r = R = 1$,

$$A_0 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A_1 = \begin{bmatrix} 0 & 0 \\ -1 & -1 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then (2.1) is equivalent to the second order scalar equation

$$\ddot{x}_1(t)+\dot{x}_1(t-1)+x_1(t-1)=u(t), \tag{8.13}$$

and the performance index is the same as (8.7). Also, the controls $u(\cdot)$ and $u_N(\cdot)$ have the forms (8.8) and (8.9), respectively.

Reasoning similar to that which in the previous example led to (8.11) and (8.12) now shows that

$$\Pi^{10}_{\cdot,12}=\Pi^{10}_{\cdot,22} \tag{8.14}$$

and

$$\Pi^{10}_{N,12}=\Pi^{10}_{N,22}, \qquad N\geqq 1. \tag{8.15}$$

The numbers then are as in Table 8.3.

TABLE 8.3

$$\Pi^{00}_{10}=\begin{bmatrix}2.0665 & 1.2164\\ 1.2164 & 1.7637\end{bmatrix},\quad \Pi^{00}_{14}=\begin{bmatrix}2.0853 & 1.2373\\ 1.2373 & 1.7884\end{bmatrix},\quad \Pi^{00}_{22}=\begin{bmatrix}2.1034 & 1.2574\\ 1.2574 & 1.8123\end{bmatrix}$$

θ	$\Pi^{10}_{10,12}(\theta)$	$\Pi^{10}_{14,12}(\theta)$	$\Pi^{10}_{22,12}(\theta)$,
−0.0	−0.1610	−0.1373	−0.1152
−0.1	−0.1610	−0.2201	−0.2247
−0.2	−0.2738	−0.3087	−0.3449
−0.3	−0.3986	−0.5043	−0.4750
−0.4	−0.5351	−0.6110	−0.6147
−0.5	−0.6837	−0.7231	−0.7631
−0.6	−0.8408	−0.9631	−1.0013
−0.7	−1.1087	−1.0903	−1.1698
−0.8	−1.0857	−1.3579	−1.3455
−0.9	−1.3711	−1.4979	−1.5278
−1.0	−1.5640	−1.6415	−1.7160

$\operatorname{tr}\Pi_{10}=5.5549$, $\operatorname{tr}\Pi_{14}=5.6771$, $\operatorname{tr}\Pi_{22}=5.8000$

$$\Pi^{10}_{10}(-1)=\begin{bmatrix}-1.0021 & -1.5640\\ -1.0021 & -1.5640\end{bmatrix},\quad \Pi^{10}_{14}(-1)=\begin{bmatrix}-1.0779 & -1.6415\\ -1.0779 & -1.6415\end{bmatrix},\quad \Pi^{10}_{22}(-1)=\begin{bmatrix}-1.1521 & -1.7160\\ -1.1521 & -1.7160\end{bmatrix}$$

9. Conclusions. The approximation theory of § 6 gives conditions under which a sequence of finite dimensional Riccati equations can be solved for sequences of operators which converge in trace norm to the solutions of the infinite dimensional Riccati equations which yield the optimal control laws for the hereditary control problems of §§ 3 and 4. We have shown that the approximation scheme of § 7 satisfies these conditions for problems on finite time intervals and, if Conjecture 7.1 holds, on the infinite interval. This conjecture seems certain to be true, and it is unfortunate that we have had to state it here as a conjecture rather than a theorem. Nevertheless, probably the most significant results of this paper concern control on the infinite interval.

From the mathematical point of view, the importance of trace-norm convergence lies in the fact that the trace-norm is the strongest of the operator norms in (5.4) and (5.5). For engineering design, trace-norm convergence of the approximate solutions of the infinite dimensional Riccati equations has two important implications. First, it ensures that we can compute approximately the trace of the solution of the appropriate

infinite dimensional Riccati equation, so that we can use this trace as a performance measure which accounts for an infinite number of initial conditions, as discussed in § 5. Also, trace-norm convergence of the solutions of the finite dimensional Riccati equations implies strong convergence, so that Theorems 6.2 and 6.8 ensure that we can choose the approximation order sufficiently large that, when the approximate optimal feedback control law is applied to the hereditary system, the response of the closed-loop system will be arbitrarily close to optimal and, for the infinite-time problem, asymptotically stable.

Note that Theorems 6.2 and 6.8 require only strong convergence of the solutions to the finite dimensional Riccati equations and, like the other results of § 6 (see Remark 6.1), do not depend on the infinite dimensional control system being a hereditary system and could be applied to control systems governed by partial differential equations. The results on trace-norm convergence in Theorems 6.3 and 6.9 depend only on the appropriate operators in the performance indices having finite rank and all the approximating operators converging strongly. As we have said earlier, a primary objective here is to point out the importance of having strong convergence of the sequences of solutions to the finite dimensional Riccati equations so that Theorems 6.2 and 6.8 hold. For this convergence, as we have seen, strong convergence of the approximating adjoint semigroups is essential.

Also among our most important results, are the characterizations, including characteristic equations, of the closed-loop eigenvalues and eigenvectors for both the hereditary system (Theorem 4.6–Corollary 4.4) and the approximating systems (Theorems 7.9–7.11). As we have said, the characterizations of § 7 should facilitate numerical solution of the finite dimensional Riccati equations of that section. In a subsequent paper, we will pursue approximate solution of the Riccati algebraic equation for the hereditary system by decomposition of the state space Z with the eigenvectors of the closed-loop system, which are given by Corollary 4.4.

Acknowledgments. The author is indebted to David Ng for the numerical work on the examples in § 8. Also, the referees made many valuable suggestions for the revision.

REFERENCES

[1] Y. ALEKAL, P. BRUNOVSKY, D. H. CHYUNG AND E. B. LEE, *The quadratic problem for systems with time delay*, IEEE Trans. Automat. Contr., AC-16 (1971), pp. 702–711.

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

[3] A. V. BALAKRISHNAN, *Applied Functional Mathematics*, Springer-Verlag, New York, 1976.

[4] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.

[5] H. T. BANKS AND A. MANITIUS, *Projection series for retarded functional differential equations with applications to optimal control problems*, J. Differential Equations, 8 (1975), pp. 296–332.

[6] R. F. CURTAIN, *The infinite-dimensional Riccati equations with application to affine hereditary differential systems*, this Journal, 13 (1975), pp. 1130–1143.

[7] 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.

[8] R. DATKO, *Uniform asymptotic stability of evolutionary processes in a Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428–554.

[9] M. C. DELFOUR, *State theory of linear hereditary differential systems*, J. Math. Anal. Appl., 60 (1977), pp. 8–35.

[10] ———, *The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution*, Appl. Math. and Optim., 3 (1977), pp. 101–162.

[11] 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.
[12] M. C. DELFOUR, C. MCCALLA AND S. K. MITTER, *Stability and the infinite-time quadratic cost problem for linear hereditary differential systems*, this Journal, 13 (1975), pp. 48–88.
[13] M. C. DELFOUR AND S. K. MITTER, *Controllability, observability and optimal feedback control of hereditary differential systems*, this Journal, 10 (1972), pp. 298–328.
[14] ———, *Hereditary differential systems with constant delays, I: general case*, J. Differential Equations, 12 (1972), pp. 213–235.
[15] ———, *Hereditary differential systems with constant delays, II: a class of affine systems and the adjoint problem*, J. Differential Equations, 18 (1975), pp. 18–28.
[16] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators*, Part I, Wiley-Interscience, New York, 1957.
[17] ———, *Linear Operators*, Part II, Wiley-Interscience, New York, 1963.
[18] J. S. GIBSON, *An analysis of optimal modal regulation: convergence and stability*, this Journal, 19 (1981), pp. 686–707.
[19] ———, *The Riccati integral equations for optimal control problems on Hilbert spaces*, this Journal, 17 (1979), pp. 537–565.
[20] J. K. HALE, *Functional Differential Equations*, Springer-Verlag, New York, 1977.
[21] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, AMS Colloquium Publications, 31, American Mathematical Society, Providence, RI, 1957.
[22] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[23] N. N. KRAVSOVSKII, *The approximation of a problem of analytic design of controls in a system with time-lag*, J. Appl. Math. Mech., 28 (1964), pp. 876–885.
[24] K. KUNISCH, *Approximation schemes for the linear-quadratic optimal control problem associated with delay-equations*, this Journal, 20 (1982), pp. 506–540.
[25] NUIBERT KWAKERNAAK AND RAPHAEL SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.
[26] R. H. KWONG AND A. S. WILLSKY, *Estimation and filter stability of stochastic delay systems*, this Journal, 16 (1978), pp. 660–681.
[27] E. B. LEE AND A. MANITIUS, *Computational approaches to synthesis of feedback controllers for multivariable systems with delays*, Proc. 1974 IEEE Conference on Decision and Control, 1974, pp. 791–792.
[28] WILLIAM S. LEVINE AND M. ATHANS, *On the determination of the optimal constant output feedback gains for linear multivariable systems*, IEEE Trans. Automat. Contr., AC-15 (1970), pp. 44–48.
[29] B. W. LEVINGER, *A folk theorem in functional differential equations*, J. Differential Equations, 4 (1968), pp. 612–619.
[30] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.
[31] 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.
[32] D. E. MULLER, *A method for solving algebraic equations using an automatic computer*, Mathematical Tables and Other Aids to Computation, 10 (1956), pp. 208–215.
[33] A. PAZY, *Semigroups of linear operators and applications to partial differential equations*, Math. Dept. Lecture Notes, Vol. 10, Univ. of Maryland, College Park, 1974.
[34] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14, 3 (1966), pp. 496–501.
[35] D. C. REBER, *A finite difference technique for solving optimization problems governed by linear functional differential equations*, J. Differential Equations, 32 (1979), pp. 193–232.
[36] IU M. REPIN, *On the approximate replacement of systems with lag by ordinary differential equations*, J. Appl. Math. Mech., 29 (1966), pp. 254–264.
[37] D. W. ROSS, *Controller design for time lag systems via a quadratic criterion*, IEEE Trans. Automat. Contr., AC-16 (1971), pp. 644–672.
[38] D. W. ROSS AND I. FLÜGGE-LOTZ, *An optimal control problem for systems with differential-difference equation dynamics*, this Journal, 7 (1969), pp. 609–623.
[39] M. E. SALUKVADZE, *Concerning the synthesis of an optimal controller in linear delay systems subjected to constantly acting perturbations*, Auto. Remote Contr., 23 (1962), pp. 1495–1501.
[40] H. F. TROTTER, *Approximations of semigroups of operators*, Pacific J. Math., 8 (1958), pp. 887–919.
[41] R. B. VINTER, *On the evolution of the state of linear differential delay equations in M^2: Properties of the generator*, J. Inst. Maths. Appls., 21 (1978), pp. 13–23.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 1, January 1983

0363-0129/83/2101-0007 $01.25/0

UNSTRUCTURED MEAN ITERATIVE PROCESSES IN REFLEXIVE BANACH SPACES*

A. J. DAVID† AND G. G. L. MEYER‡

Abstract. The unstructured mean iterative process is an algorithm schema that may be used to model numerical methods for solving fixed point problems. This schema relates the sequences $\{y_i\}$ and $\{z_i\}$ in a linear space X and a scalar sequence $\{\lambda_i\}$ via the relation $z_{i+1} = z_i + \lambda_i(y_i - z_i)$. The "unstructured" use of y_i rather than the explicit use of an algorithmic map involving the iterates $z_i, z_{i-1}, \cdots, z_0$, allows the properties of the schema to be separated from the properties of any particular class of applications. The results presented concern the boundedness and convergence properties of $\{y_i\}$ and $\{z_i\}$ in reflexive Banach spaces, as controlled by the parameter sequence $\{\lambda_i\}$. These properties provide guidelines that are useful in a systematic approach to the synthesis of the iterative algorithms for solving fixed point problems.

Key words. algorithm, mean iterative process, strong convergence, weak convergence, iterate, cluster point, boundedness, point-to-set map

1. Introduction. The results obtained in this paper provide guidelines to algorithm synthesis and convergence verification for a large class of fixed point algorithms. This class of algorithms is modeled by the one-dimensional mean iterative process defined on a linear space X.

ALGORITHM 1. Let z_0 and $\{y_i\}$ be in X and $\{\lambda_i\}$ be in E.
Step 0. Set $i = 0$.
Step 1. Let $z_{i+1} = z_i + \lambda_i(y_i - z_i)$.
Step 2. Set $i = i + 1$ and go to Step 1.

Algorithm 1 does not directly involve a fixed point map. Thus, a wide range of applications may be studied through examination of the algorithm schema in a suitably general space. To illustrate how specific algorithms may be recovered as special cases of Algorithm 1, we discuss some well-known applications.

When $y_i = a(z_i)$ for all i, where $a(\cdot)$ is a nonlinear map from X into X, different choices for the sequence $\{\lambda_i\}$ result in different known algorithms. If $\lambda_i = 1$ for all i, then Algorithm 1 reduces to the Picard iteration [20]. If $\lambda_i = \frac{1}{2}$ for all i, the algorithm is the Krasnoselskii bisection method [8], [11]. For $\lambda_i = 1/(i+1)$ and the choice $z_0 = y_0$, the centroidal method of Mann is recovered [5], [17]. If for all i, Y_i is a random variable with conditional mean $E(Y_i \mid Z_i) = a(Z_i)$ and $\lambda_i = 1/(i+1)$, then for Y_i and Z_i scalar random variables Algorithm 1 reduces to the stochastic approximation method of Robbins and Monro [24]. For Y_i and Z_i n-vector-valued random variables, this is the multidimensional stochastic approximation algorithm [1]. The learning and pattern recognition algorithms of Tsypkin, Braverman and Rozonoer are of the form of the schema [2], [3], [26]. A class of nonlinear programming algorithms discussed by Nurminskii [18] is also of this form. The convergence of stochastic versions of these and similar algorithms is covered in [13], [14], [15], [19]. So far, the algorithms that have been mentioned use a sequence of step lengths $\{\lambda_i\}$ that is given in advance. But the sequence $\{\lambda_i\}$ may also be computed adaptively, as in the steepest descent algorithm where $y_i = z_i - \nabla f(z_i)$ and λ_i is the "best step" with respect to minimizing the value of $f(z_{i+1})$. Other nonlinear programming algorithms of this form include Newton's method and conjugate gradient methods [16], [21].

* Received by the editors October 28, 1980, and in final revised form March 8, 1982.

† Bell Laboratories, Holmdel, New Jersey 07733. The work of this author was done while he was at The Johns Hopkins University, Baltimore, Maryland 21218.

‡ Electrical Engineering Department, The Johns Hopkins University, Baltimore, Maryland 21218.

In § 2, the algorithm schema is treated from the system theoretic point of view as a means of relating an output sequence to an input sequence via the parameter sequence $\{\lambda_i\}$. First $\{z_i\}$ is viewed as the output and $\{y_i\}$ as the input, and then the roles are reversed. In § 3, the input–output analysis is used to obtain results on algorithms that involves point-to-set maps. Conclusions and relations with some previous results are presented in § 4. For ease in reading, all proofs are given in the Appendix.

In this paper the following notation is used.

$\bar{T}$	The closure of the set T.
$\overline{\text{co}}\,(T)$	The convex closure of the set T.
q_x	The weak sequential cluster point set of the sequence $\{x_i\}$.
E	The real line.
(a, b)	The ordered pair denoting an element of E^2.
$[a; b)$	The set of points x in E such that $a \leqq x < b$.
$\{a, b\}$	The set containing the elements a and b.
$B(x, \delta)$	The open sphere with center x and radius δ.
$B(T, \delta)$	The union of open sets $\cup_{x \in T} B(x, \delta)$.
$\sum \lambda_i$	The infinite summation $\sum_{i=0}^{\infty} \lambda_i$.
$\limsup \{x_i\}$	$\lim_{i \to \infty} \sup_{j \geqq i} \|x_j\|$.

2. Input-output study. Our study of the unstructured mean iterative process centers on the influence of the parameter sequence $\{\lambda_i\}$ on the relationships between such basic properties of $\{y_i\}$ and $\{z_i\}$ as boundedness, strong and weak convergence, and the location of their respective weak cluster point sets. The presentation is in two parts. First, it is assumed that $\{y_i\}$ is bounded. As increasingly strong conditions are placed on $\{\lambda_i\}$, increasingly precise results are obtained linking the properties of $\{y_i\}$ and $\{z_i\}$. Later, it is assumed that $\{z_i\}$ is bounded, and the class of sequences $\{\lambda_i\}$ is determined so that the properties of $\{y_i\}$ and $\{z_i\}$ are usefully related.

In this section, X denotes a reflexive Banach space over the reals, and the sequences $\{y_i\}$ and $\{z_i\}$ in X and the sequence $\{\lambda_i\}$ in E are assumed to be related as in Algorithm 1.

Examples 1 and 2 illustrate that if $\{\lambda_i\}$ is outside the interval $[0; 2)$ infinitely many times then the boundedness of the sequence $\{y_i\}$ does not assure that $\{z_i\}$ is bounded.

Example 1. Let $X = E$. Suppose that $z_0 \geqq 0$, $\{\lambda_i\}$ is in $(-\infty; 0)$, and $\sum \lambda_i$ diverges. Let $y_i = -1$ for all i. Then

$$z_{i+1} > z_0 - \sum_{j=0}^{i} \lambda_j > 0,$$

for all i, and since $\sum \lambda_i$ diverges, the sequence $\{z_i\}$ is unbounded.

Example 2. Let $X = E$. Suppose that $\{\lambda_i\}$ is in $[2; \infty)$, $z_0 \geqq 0$, and

$$y_i = (-1)^{i+1} \frac{1}{i+1}$$

for all i. The process may be written in terms of alternate iterates as

$$z_{i+2} = (1-\lambda_{i+1})(1-\lambda_i) z_i + (1-\lambda_{i+1}) \lambda_i y_i + \lambda_{i+1} y_{i+1}.$$

For all even indices i,

$$z_{i+2} \geqq z_i + \frac{2}{i+1} + \frac{2}{i+2} > 0,$$

and therefore, for $i \geqq 0$, i even,

$$z_{i+2} \geqq z_0 + 2\left[1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{i+2}\right].$$

The subsequence of even iterates is unbounded and so $\{z_i\}$ is unbounded.

Now we show that even if $\{\lambda_i\}$ belongs to $[0; 2)$ and $\{y_i\}$ is bounded, $\{z_i\}$ need not be bounded.

Example 3. Let $X = E$. Let $z_0 = 1$, and for $i = 0, 1, 2, \cdots$, let $\lambda_i = 2 - 1/(i+1)$ and $y_i = (-1)^i$. Then for all $i \geqq 1$, $z_i = (-1)^{i-1} i$.

Example 3 illustrates that when $\{\lambda_i\}$ in $[0; 2)$ is allowed to become arbitrarily close to 2, then the boundedness of the sequence $\{y_i\}$ may not assure the boundedness of the sequence $\{z_i\}$. Thus, a relationship between the boundedness of $\{y_i\}$ and $\{z_i\}$ exists only if we rule out all sequences $\{\lambda_i\}$ which leave the closed interval $[0; \lambda_M]$, $\lambda_M < 2$, infinitely many times.

The following basic result allows the boundedness of the sequences $\{y_i\}$ and $\{z_i\}$ to be related. The role of the interval $[0; 2)$ is seen to be crucial.

LEMMA 1. *Let w, x, y, z in X, λ in $[0; 2)$, $\delta > 0$, and β be such that $x = z + \lambda(y - z)$, $\|y - w\| \leqq \delta$ and $\beta = \max(1; \lambda/(2-\lambda))$.*

Then:

(i) $\|x - w\| \leqq \delta\beta$ *whenever* $\|z - w\| \leqq \delta\beta$, *and*

(ii) *for every* $\varepsilon \geqq 0$, $\|x - w\| \leqq \|z - w\| - \lambda\varepsilon$, *whenever* $\|z - w\| \geqq (\delta + \varepsilon)\beta$.

The boundedness results we have been seeking are now immediate. Using the mean iterative process to generate the sequence $\{z_i\}$, the points z_{i+1}, y_i and z_i can be identified with the variables x, y and z, respectively, of Lemma 1. If w is the center of a sphere containing the sequence $\{y_i\}$, then a consequence of Lemma 1 is that regardless of the choice of the initial point z_0, the distance of subsequent iterates z_i from w will be bounded.

THEOREM 1. *If* (i) $\{y_i\}$ *is bounded, and* (ii) $\{\lambda_i\}$ *is in* $[0; \lambda_M]$ *for some* $\lambda_M < 2$, *then* $\{z_i\}$ *is bounded.*

This boundedness relationship between $\{y_i\}$ and $\{z_i\}$ may be strengthened when further requirements are imposed on the sequence $\{\lambda_i\}$. Essentially, when $\sum \lambda_i$ diverges, the smallest sphere containing all but finitely many members of the sequence $\{y_i\}$ serves as a region of attraction [25] for the sequence $\{z_i\}$.

LEMMA 2. *If* (i) $\{y_i\}$ *is bounded,* (ii) $\{\lambda_i\}$ *is in* $[0; \lambda_M]$, $\lambda_M < 2$, *and* (iii) $\sum \lambda_i$ *diverges, then for all* w,

$$\limsup \|z_i - w\| \leqq \beta \limsup \|y_i - w\|,$$

where $\beta = \max(1; \lambda_M/(2-\lambda_M))$.

In the following example we illustrate the result of Lemma 2 in the Banach space l_1.

Example 4. Let $\{\lambda_i\}$ be in $[0; 1.5]$, let $z_0 = (0, 0, 0, \cdots)$ and let $y_i = e_i$ for all i, where $e_0 = (0, 0, 0, \cdots)$, and e_i is the ith unit vector for all $i \geqq 1$, i.e., $e_1 = (1, 0, ,0 \cdots)$, $e_2 = (0, 1, 0, \cdots)$, etc. Then $\|y_i\|_1 = 1$ for all i, $\beta = 3$, and from Lemma 2, using $w = e_0$ we obtain $\limsup \|z_i\| \leqq 3$.

Lemma 2 may be used to relate the sequences $\{y_i\}$ and $\{z_i\}$ when $\{y_i\}$ converges either strongly or weakly.

COROLLARY 1. *If* (i) $\{y_i\}$ *converges strongly to some point* y_*, (ii) $\{\lambda_i\}$ *is in the* $[0; \lambda_M]$, $\lambda_M < 2$, *and* (iii) $\sum \lambda_i$ *diverges, then* $\{z_i\}$ *converges strongly to* y_*.

COROLLARY 2. *If* (i) $\{y_i\}$ *converges weakly to some point* y_*, (ii) $\{\lambda_i\}$ *is in* $[0; \lambda_M]$, $\lambda_M < 2$, *and* (iii) $\sum \lambda_i$ *diverges, then* $\{z_i\}$ *converges weakly to* y_*.

We know that bounded sets are weakly sequentially compact in reflexive Banach spaces [6]. Thus Lemma 2 can be specialized to provide a relationship between the weak sequential cluster point sets q_y and q_z of the sequences $\{y_i\}$ and $\{z_i\}$, respectively. We recall that a point x in a normed linear space is a weak sequential cluster point of a sequence $\{x_i\}$ if and only if there exists an infinite subset K of the integers such that $\{f(x_i)\}_K$ converges to $f(x)$ for every continuous linear functional $f(\cdot)$. Also, if $\{x_i\}$ is a bounded sequence in a reflexive Banach space and q_x is its weak cluster point set, then for any continuous linear functional $f(\cdot)$ on X, and for any $\varepsilon > 0$, there exists an index k, depending on $f(\cdot)$ and ε, such that for all $i \geqq k$

$$f(x_i) \in f(B(\overline{\text{co}}\,(q_x), \varepsilon)).$$

THEOREM 2. *If* (i) $\{y_i\}$ *is bounded,* (ii) $\{\lambda_i\}$ *is in* $[0; \lambda_M]$ *for some* $\lambda_M < 2$, *and* (iii) $\sum \lambda_i$ *diverges, then the weak sequential cluster point sets* q_y *and* q_z *of the sequences* $\{y_i\}$ *and* $\{z_i\}$, *respectively are nonempty and bounded, and*

$$q_z \subseteq \bar{B}(w, \delta\beta)$$

for every point w *in* X *and scalar* δ *such that*

$$q_y \subseteq \bar{B}(w, \delta),$$

where $\beta = \max(1; \lambda_M/(2-\lambda_M))$.

The following example in E^2 illustrates the relationship between cluster point sets of sequences given in Theorem 2.

Example 5. Let $\lambda_i = 1.5$ for all i, let $z_0 = (3, 3)$, and let $\{y_i\}$ be defined by $y_0 = (3, -1)$, $y_1 = (-1, -3)$, $y_2 = (-3, 1)$, $y_3 = (1, 3)$, $y_4 = y_0$, $y_5 = y_1$, etc. The corresponding sequence $\{z_i\}$ is given by $z_1 = (3, -3)$, $z_2 = (-3, -3)$, $z_3 = (-3, 3)$, $z_4 = (3, 3)$, $z_5 = z_1$, $z_6 = z_2$, etc. If we let w be the origin of E^2, i.e., $w = (0, 0)$, then q_y is contained in the closed ball $\bar{B}(w, \delta)$ with $\delta = 10^{0.5}$, and therefore, from Theorem 2, we conclude that the cluster point set of $\{z_i\}$ is contained in the closed ball centered at the origin with radius $3(10)^{0.5}$. Note that q_z is not contained in the convex closure of q_y. As we will see below, further restrictions on the sequence $\{\lambda_i\}$ are needed to insure that q_z is a subset of the convex closure of q_y.

Theorem 2 depends heavily on the reflexivity of the space X since reflexivity provides a convenient way of relating a bounded sequence with its weak sequential cluster point set. In a nonreflexive space, it is possible to have a bounded sequence which does not possess any weak cluster points, or to have a bounded sequence which possesses one and only one weak cluster point, but which does not converge weakly to it.

The relative sizes of the regions containing the weak sequential cluster point sets q_y and q_z have been related provided that $\{\lambda_i\}$ is in $[0; \lambda_M]$, $\lambda_M < 2$ and $\sum \lambda_i$ diverges. An improved bound for q_z is obtained if we restrict $\{\lambda_i\}$ to the interval $[0; 1]$. A first step towards clarifying the relationship between q_y and q_z under this restriction is to assume that at least one point z_* in q_z is known.

LEMMA 3. *If* (i) $\{y_i\}$ *is bounded,* (ii) $\{\lambda_i\}$ *is in* $[0; 1]$, *and* (iii) z_* *is in* q_z, *then* q_z *is a subset of the convex closure of* z_* *and* q_y.

The next lemma is a technical result that will be required subsequently.

LEMMA 4. *Let* C *be a convex subset of* X *and let* v *and* w *be points in* X *not contained in* C. *If* v *is an element of the convex closure of* w *and* C, *and* w *is an element of the convex closure of* v *and* C, *then* v *is equal to* w.

Lemmas 3 and 4 imply the following result.

LEMMA 5. *If* (i) $\{y_i\}$ *is bounded,* (ii) $\{\lambda_i\}$ *is in* $[0; 1]$, *and* (iii) $\{z_i\}$ *is not weakly convergent, then* q_z *is a subset of* $\overline{\text{co}}\,(q_y)$.

The assumptions of Lemma 5 insure that $\{z_i\}$ is bounded but not weakly convergent, and this implies that $\sum \lambda_i$ diverges. If, on the other hand, $\{z_i\}$ converges to some point z_*, then in order to relate z_* and q_y it must be explicitly assumed that $\sum \lambda_i$ diverges. We begin with the case in which the sequence $\{y_i\}$ is contained in a closed and convex set S_y.

LEMMA 6. *If* (i) $\{y_i\}$ *is contained in a closed and convex set* S_y, (ii) $\{\lambda_i\}$ *is in* $[0; \infty)$, (iii) $\sum \lambda_i$ *diverges, and* (iv) $\{z_i\}$ *converges weakly to* z_*, *then*

$$z_* \in S_y.$$

The desired result relating z_* and the weak sequential cluster point set q_y is now obtained easily provided that the sequence $\{y_i\}$ is bounded.

COROLLARY 3. *If* (i) $\{y_i\}$ *is bounded,* (ii) $\{\lambda_i\}$ *is in* $[0; \infty)$, (iii) $\sum \lambda_i$ *diverges, and* (iv) $\{z_i\}$ *converges weakly to* z_*, *then*

$$z_* \in \overline{\text{co}}\,(q_y).$$

The next main result of this section is a consequence of Lemma 5 and Corollary 3.

THEOREM 3. *If* (i) $\{y_i\}$ *is bounded,* (ii) $\{\lambda_i\}$ *is in* $[0; 1]$, *and* (iii) $\sum \lambda_i$ *diverges, then the weak cluster point set* q_z *of the sequence* $\{z_i\}$ *is contained in the convex closure of the weak cluster point set* q_y *of the sequence* $\{y_i\}$.

Note that if $\{y_i\}$ is bounded and $\sum y_i$ diverges, then different bounds on q_z are obtained depending on whether $\{\lambda_i\}$ is in $[0; 1]$ or $[0; \lambda_M]$, for some $\lambda_M < 2$.

The following example in E shows that q_z may be a strict subset of the convex closure of q_y. Equality of the two cluster point sets is not to be obtained in general.

Example 6. Let $z_0 = 0$, let $\lambda_i = \frac{1}{2}$ for all i, let $y_i = -1$ for all i odd, and let $y_i = 2$ for all i even. Then $\{y_i\}$ is bounded, $q_y = \{-1, 2\}$, $z_i = 0$ for i even, $z_i = 1$ for i odd, and the set q_z which consists of the two points 0 and 1 is a strict subset of the convex closure of q_y.

We have discussed the consequences of having $\{\lambda_i\}$ in $[0; \lambda_M]$, for some $\lambda_M < 2$, such that $\sum \lambda_i$ diverges. Now we examine the consequence of removing the requirement that $\sum \lambda_i$ diverges.

LEMMA 7. *If* (i) $\{y_i\}$ *converges strongly to* y_*, *and* (ii) $\{\lambda_i\}$ *is in* $[0; \lambda_M]$ *for some* $\lambda_M < 2$, *then the sequence* $\{z_i\}$ *converges strongly, but not necessarily to* y_*.

LEMMA 8. *If* (i) $\{y_i\}$ *converges weakly to* y_*, *and* (ii) $\{\lambda_i\}$ *is in* $[0; \lambda_M]$ *for some* $\lambda_M < 2$, *then the sequence* $\{z_i\}$ *converges weakly, but not necessarily to* y_*.

So far it has been assumed that some of the properties of the sequence $\{y_i\}$ are known; thus each result has as one of its assumptions that $\{y_i\}$ is either bounded or convergent. Now we examine the consequences of the assumption that $\{z_i\}$ is bounded or convergent. For algorithm synthesis, this is an important question: If $\{z_i\}$ converges to some point z_*, it is necessary to know how the sequence $\{\lambda_i\}$ influences the behavior of $\{y_i\}$. The next two results illustrate the consequences of choosing sequences $\{\lambda_i\}$ that are bounded away from 0.

THEOREM 4. *If* (i) $\{z_i\}$ *is bounded, and* (ii) *there is some* $\lambda_m > 0$ *such that* $\{\lambda_i\}$ *is in* $[\lambda_m; \infty)$, *then* $\{y_i\}$ *is bounded.*

When $\{z_i\}$ converges either strongly or weakly to some limit z_*, and $\{\lambda_i\}$ is in $[\lambda_m; \infty)$, then Theorem 4 implies that $\{y_i\}$ is bounded; thus the hypotheses of Corollary 3 are satisfied and so we know that z_* belongs the $\overline{\text{co}}\,(q_y)$. However, Corollary 3

requires only that $\{\lambda_i\}$ belong to $[0; \infty)$. With the additional hypothesis that $\{\lambda_i\}$ is bounded away from 0, an even stronger relationship is implied.

COROLLARY 4. *If* (i) $\{z_i\}$ *converges strongly to* z_*, *and* (ii) *there is some* $\lambda_m > 0$ *such that* $\{\lambda_i\}$ *is in* $[\lambda_m; \infty)$, *then* $\{y_i\}$ *converges strongly to* z_*.

COROLLARY 5. *If* (i) $\{z_i\}$ *converges weakly to* z_*, *and* (ii) *there is some* $\lambda_m > 0$ *such that* $\{\lambda_i\}$ *is in* $[\lambda_m; \infty)$, *then* $\{y_i\}$ *converges weakly to* z_*.

Theorem 4 and Corollaries 4 and 5 are sufficient conditions in the sense that for any bounded (or convergent) sequence $\{z_i\}$, as long as $\{\lambda_i\}$ is bounded away from 0, then $\{y_i\}$ is bounded or convergent. The next example illustrates that these results are also necessary in the following sense: We exhibit a convergent sequence $\{z_i\}$ and a sequence $\{\lambda_i\}$ in $[0; 1]$ where $\sum \lambda_i$ diverges such that $\{y_i\}$ is unbounded.

Example 7. Let $z_0 = 1$, for i even, let $y_i = 1$ and $\lambda_i = 1$, and for i odd, let $y_i = i$ and $\lambda_i = 0$. Then $z_i = 1$ for all i, $\sum \lambda_i$ diverges, and $\{y_i\}$ is unbounded.

This concludes our discussion of the influence of the parameter sequence $\{\lambda_i\}$ on the relationships that exist between the sequences $\{y_i\}$ and $\{z_i\}$ as related by the unstructured mean iterative process. Next we turn to an application of the preceding results.

3. Applications of the input–output theory. Many problems of interest in nonlinear programming and elsewhere can be reformulated as fixed point problems involving a point-to-set map $A(\cdot)$ defined on all of X or defined only on a closed and convex subset T of the space X. We now proceed to apply the results of the preceding section to mean iterative processes that use a point-to-set map $A(\cdot)$ from X into all the nonempty subsets of X as follows.

ALGORITHM 2. Let z_0 be a point in X.
Step 0. Set $i = 0$.
Step 1. Pick y_i in $A(z_i)$.
Step 2. Set $z_{i+1} = z_i + \lambda_i(y_i - z)$.
Step 3. Set $i = i + 1$ and go to Step 1.

In this section we assume that X is a reflexive Banach space over the reals, and that $\{y_i\}$ and $\{z_i\}$ are sequences in X and $\{\lambda_i\}$ is a sequence in E related as in Algorithm 2. We now find those properties of $\{\lambda_i\}$ that assure that if $\{z_i\}$ converges to z_*, then z_* is in $A(z_*)$. First we recall that a point-to-set map $A(\cdot)$ is strongly (weakly) sequentially closed at a point x in X if whenever $\{x_i\}$ in X converges strongly (weakly) to x, w_i is in $A(z_i)$ for all i, and $\{w_i\}$ converges strongly (weakly) to some point w in X, then w is in $A(x)$.

When $A(\cdot)$ is defined on all of X, the following results may be obtained using Corollaries 4 and 5.

THEOREM 5. *Let the map* $A(\cdot)$ *be strongly sequentially closed on* X. *If* (i) $\{\lambda_i\}$ *is in* $[\lambda_m; \infty)$ *for some* $\lambda_m > 0$, *and* (ii) $\{z_i\}$ *converges strongly to some point* z_*, *then* z_* *belongs to* $A(z_*)$.

THEOREM 6. *Let the map* $A(\cdot)$ *be weakly sequentially closed on* X. *If* (i) $\{\lambda_i\}$ *is in* $[\lambda_m; \infty)$ *for some* $\lambda_m > 0$, *and* (ii) $\{z_i\}$ *converges either weakly or strongly to some point* z_*, *then* z_* *belongs to* $A(z_*)$.

The preceding results are satisfactory if $A(\cdot)$ is closed in the appropriate sense and if a value of λ_m can be estimated. But in many applications it is necessary to let $\{\lambda_i\}$ converge to 0—for example if $\{y_i\}$ cannot converge (it may have a random component) it is easy to see that $\{z_i\}$ does not converge when $\{\lambda_i\}$ is bounded away from 0. Hence, if $A(\cdot)$ possesses an additional useful property, more generality in the choice of $\{\lambda_i\}$ results.

THEOREM 7. *Let the map $A(\cdot)$ be weakly sequentially closed on X and assume that the set $A(x)$ is convex for every x in X. If* (i) $\{y_i\}$ *is bounded,* (ii) $\{\lambda_i\}$ *is in* $[0; \infty)$, (iii) $\sum \lambda_i$ *diverges, and* (iv) $\{z_i\}$ *converges either weakly or strongly to* z_*, *then*

$$z_* \in A(z_*).$$

Often, problems in nonlinear programming are defined on a subset T of X. In this case, Algorithm 2 is well-defined provided that (i) T is closed and convex, (ii) $A(\cdot)$ maps T into all the nonempty subsets of T, (iii) $\{\lambda_i\}$ is in $[0; 1]$, and (iv) z_0 is in T. Under these restrictions, Theorems 5, 6 and 7 remain valid.

The reader may note that when the point-to-set map $A(\cdot)$ is actually a weakly continuous point-to-point map, then Theorem 7 is simplified. This is because $A(x)$ is convex for all x in X and the sequence $\{y_i\}$ is bounded whenever $\{z_i\}$ converges either weakly or strongly.

COROLLARY 6. *Let the map $A(\cdot)$ be weakly sequentially closed on X and assume that the set $A(x)$ contains one and only one point for every x in X. If* (i) $\{\lambda_i\}$ *is in* $[0; \infty)$; (ii) $\sum \lambda_i$ *diverges, and* (iii) $\{z_i\}$ *converges either weakly or strongly to* z_*, *then*

$$z_* \in A(z_*).$$

4. Conclusions. Central to the field of fixed point algorithms is Mann's 1953 paper, *Mean Value Methods in Iteration* [17] in which the Mann averaging process is introduced and some of its basic properties are discussed. Our results in reflexive Banach space on the relationships between the weak sequential cluster point sets q_y and q_z are generalizations, for the mean unstructured iterative process, of Mann's results on a compact, convex subset of Banach space. Other papers on mean iteration view it as a special case of the Mann process, thereby limiting the sequence $\{\lambda_i\}$ to $[0; 1]$, [5], [23]. Our work shows that although some results are obtained when $\{\lambda_i\}$ is in $[0; 1]$, others are obtained when $\{\lambda_i\}$ is in $[0; \lambda_M]$, $\lambda_M < 2$, or when $\{\lambda_i\}$ is in $[0; \infty)$.

Whereas the concern of this paper is with the influence of the sequence $\{\lambda_i\}$ on the relationships between the weak sequential cluster point sets q_y and q_z, most authors concentrate on verifying the convergence of algorithms which, for some particular class of maps $B(\cdot)$, use $y_i = B(z_i)$ for all i. For instance, using the properties of the Mann process, existence of fixed points and convergence, in the appropriate sense, to these fixed points, can be shown for nonexpansive mappings in Banach space and even in nonnormable space [10], [22]. Likewise, mean iteration may be used to find the fixed points of various classes of maps in Hilbert space possessing contractive properties [4], [9] or maps whose ranges are contained in compact subsets of Banach space [12], to obtain sequences whose weak cluster points coincide with the fixed points of the maps under consideration. Another topic that has been investigated is the selection of specific sequences $\{\lambda_i\}$ that improve the rate of convergence of the mean iterative process to fixed points of particular classes of maps in Hilbert space [7].

In closing, we wish to remark on several aspects of our approach. An effort has been made to avoid strong assumptions on the nature of the space in which the results are given. For example, local compactness is not required; thus we forego the opportunity of obtaining results involving strong sequential cluster point sets. Hence our effort centers on weak convergence and weak sequential cluster points. A common hypothesis in this paper is that $\{y_i\}$ be a bounded sequence. Often this requirement is not difficult to fulfill: The map generating $\{y_i\}$ may have a bounded range, and if not, a bounded neighborhood of the solution set may be known in advance so that if for some i, y_i is not in this neighborhood, then y_i can be projected onto it.

The thrust of our results in this paper is also different from much other work in fixed point and algorithm theory. We view the mean iterative process in the absence of a surrogate map or potential function [2], [3]. But when for all i, y_i is given by $B(z_i)$ for some map $B(\cdot)$ with contractive or nonexpansive properties, then an implicit potential function exists involving the distance of $B(z_i)$ from the fixed point. Thus, our main results can be used to guide the initial steps in synthesizing algorithms, e.g., showing that $\{z_i\}$ is a bounded sequence, and that if $\{z_i\}$ is weakly convergent to some point z_*, then z_* is a solution of the given problem.

Appendix: Proofs for §§2 and 3.

Lemma 1. By construction

$$x = (1-\lambda)z + \lambda y,$$

and therefore

$$x - w = (1-\lambda)(z-w) + \lambda(y-w).$$

The triangle inequality and the assumption that $\|y-w\| \leqq \delta$ imply directly that

$$\|x-w\| \leqq |1-\lambda|\,\|z-w\| + |\lambda|\delta. \tag{1}$$

Now we show that results (i) and (ii) hold.

(i) Assume that $\|z-w\| \leqq \delta\beta$. If λ is in $[0;1]$, then $\beta = 1$, and

$$\|z-w\| \leqq \delta. \tag{2}$$

Thus (1) becomes

$$\|x-w\| \leqq (1-\lambda)\|z-w\| + \lambda\delta,$$

and using (2) we obtain

$$\|x-w\| \leqq \delta\beta.$$

If λ is in $[1;2)$, then $\beta = \lambda/(2-\lambda)$, and

$$\|z-w\| \leqq \frac{\delta\lambda}{2-\lambda}. \tag{3}$$

Thus, (1) becomes

$$\|x-w\| \leqq (\lambda-1)\|z-w\| + \lambda\delta$$

and using (3) we obtain

$$\|x-w\| \leqq \frac{\delta\lambda(\lambda-1)}{2-\lambda} + \delta\lambda,$$

which reduces to

$$\|x-w\| \leqq \delta\beta.$$

Therefore result (i) holds.

(ii) Now assume that $\|z-w\| \geqq (\delta+\varepsilon)\beta$ for some $\varepsilon \geqq 0$. If λ is in $[0;1]$, then $\beta = 1$, and

$$\|z-w\| - \varepsilon \geqq \delta. \tag{4}$$

Thus, (1) becomes

$$\|x-w\| \leqq (1-\lambda)\|z-w\| + \lambda\delta,$$

and using the bound for δ from (4), we obtain

$$\|x-w\| \leqq \|z-w\| - \lambda\varepsilon.$$

If λ is in $[1; 2)$, then $\beta = \lambda/(2-\lambda)$, and

$$\left(\frac{2-\lambda}{\lambda}\right)\|z-w\| - \varepsilon \geqq \delta. \tag{5}$$

But, by (1),

$$\|x-w\| \leqq (\lambda-1)\|z-w\| + \lambda\delta,$$

and therefore, using (5) to bound δ, we obtain

$$\|x-w\| \leqq \|z-w\| - \lambda\varepsilon.$$

We conclude that result (ii) of the lemma holds.

Theorem 1. No proof needed.

Lemma 2. Let w be a point in X. Since $\{y_i\}$ is bounded there exists a scalar ρ such that

$$\rho = \limsup \|y_i - w\|.$$

Thus, given a scalar $\alpha > 0$, there exists an index k such that

$$\|y_i - w\| \leqq \rho + \alpha$$

for all $i \geqq k$.

Suppose that there exists a scalar $\varepsilon > 0$ and an index $j \geqq k$ such that

$$\|z_i - w\| \geqq (\rho + \alpha + \varepsilon)\beta$$

for all $i \geqq j$. Then, from Lemma 1,

$$\|z_{i+1} - w\| \leqq \|z_i - w\| - \lambda_i\varepsilon \tag{6}$$

for all $i \geqq j$, and since we assume that $\sum \lambda_i$ diverges, (6) implies that $\{\|z_i - w\|\}$ is unbounded from below, which is impossible. We conclude that given any $\varepsilon > 0$, there exists an index $j \geqq k$ such that

$$\|z_j - w\| < (\rho + \alpha + \varepsilon)\beta. \tag{7}$$

From Lemma 1 we know that for all $i \geqq k$, either

$$\|z_i - w\| \leqq (\rho + \alpha)\beta \quad \text{or} \quad \|z_{i+1} - w\| < \|z_i - w\|.$$

Thus, (7) implies that

$$\|z_i - w\| < (\rho + \alpha + \varepsilon)\beta$$

for all $i \geqq j$. Therefore

$$\limsup \|z_i - w\| \leqq (\rho + \alpha + \varepsilon)\beta$$

for all $\alpha > 0$ and for all $\varepsilon > 0$, and the result is proved.

Corollary 1. No proof needed.

Corollary 2. By hypothesis, the sequence $\{y_i\}$ converges weakly to the point y_*, and therefore the sequence $\{f(y_i)\}$ converges to $f(y_*)$ for every continuous linear

functional $f(\cdot)$ on X. By the linearity of $f(\cdot)$, the sequences $\{f(z_i)\}$ and $\{f(y_i)\}$ are related via the mean iterative process. Thus we may apply Lemma 2. By assumption, $\sum \lambda_i$ diverges. Hence, the sequence $\{f(z_i)\}$ converges to $f(y_*)$ for every continuous linear functional $f(\cdot)$ on X and so $\{z_i\}$ converges weakly to y_*.

Theorem 2. To each continuous linear functional $f(\cdot)$ on X, with $\|f(\cdot)\| = 1$, and to each scalar $\varepsilon_1 > 0$, there corresponds an index k_1 depending on $f(\cdot)$ and ε_1 such that for all $i \geqq k_1$,

$$|f(y_i) - f(w)| \leqq \delta + \varepsilon_1.$$

Using Lemma 2 we see that given $\varepsilon_2 > 0$, there exists an index k_2 (also depending on $f(\cdot)$ and ε_1) such that for all $i \geqq k_2$,

$$|f(z_i) - f(w)| \leqq \beta(\delta + \varepsilon_1) + \varepsilon_2. \tag{8}$$

The sequence $\{z_i\}$ in X is bounded and so has a bounded, nonempty weak cluster point set q_z. It follows from (8) that for all z_* in q_z,

$$|f(z_*) - f(w)| \leqq \beta\delta,$$

and thus

$$\|z_* - w\| \leqq \beta\delta.$$

Lemma 3. Let $f(\cdot)$ be a continuous linear functional on X and let ε be a positive scalar. The sequence $\{y_i\}$ is bounded, $\{\lambda_i\}$ is in $[0; 1]$, and therefore we know from Theorem 1 that $\{z_i\}$ is bounded. The definitions of q_y and z_* imply the existence of a scalar k such that

$$f(y_i) \in f(B(\overline{\text{co}}\,(q_y), \varepsilon))$$

for all $i \geqq k$, and

$$f(z_k) \in f(B(z_*, \varepsilon)).$$

It follows immediately that

$$f(y_i) \in f(B(\overline{\text{co}}\,(q_y, z_*), \varepsilon))$$

for all $i \geqq k$ and that

$$f(z_k) \in f(B(\overline{\text{co}}\,(q_y, z_*), \varepsilon)).$$

By assumption, the scalar λ_i is in $[0; 1]$ for all i and we conclude that

$$f(z_i) \in f(B(\overline{\text{co}}\,(q_y, z_*), \varepsilon))$$

for all $i \geqq k$. The sequence $\{z_i\}$ is bounded, and therefore every weak cluster point of $\{z_i\}$ is in $B(\overline{\text{co}}\,(q_y, z_*), \varepsilon)$ for all positive ε; hence the result.

Lemma 4. Assume that v is not equal to w. Since v is in the convex closure of w and C, we can write it as the convex combination of w and some point c_1 in C:

$$v = (1 - \alpha_1)w + \alpha_1 c_1, \qquad 0 \leqq \alpha_1 \leqq 1. \tag{9}$$

Likewise, for w and some c_2 in C,

$$w = (1 - \alpha_2)v + \alpha_2 c_2, \qquad 0 \leqq \alpha_2 \leqq 1. \tag{10}$$

By assumption v is not equal to w, and v and w are not in C, so α_1 and α_2 cannot equal 0 or 1. Hence they belong to the interval $(0; 1)$. Substituting (10) into (9) yields

$$v = (1 - \alpha_1)(1 - \alpha_2)v + (1 - \alpha_1)\alpha_2 c_2 + \alpha_1 c_1,$$

and therefore

$$(1-(1-\alpha_1)(1-\alpha_2))v = (1-\alpha_1)\alpha_2 c_2 + \alpha_1 c_1.$$

But, $0<\alpha_1<1$ and $0<\alpha_2<1$ imply that the coefficient $1-(1-\alpha_1)(1-\alpha_2)$ is strictly positive. Thus, v can be written as

$$v = \frac{(1-\alpha_1)\alpha_2}{1-(1-\alpha_1)(1-\alpha_2)} c_2 + \frac{\alpha_1}{1-(1-\alpha_1)(1-\alpha_2)} c_1.$$

Since the coefficients of c_1 and c_2 are positive and sum to unity, v is a convex combination of elements of C. But since C is convex, this implies that v must be an element of C. This contradicts the hypothesis that v is not in C; hence v equals w.

Lemma 5. The sequence $\{z_i\}$ is bounded and therefore if $\{z_i\}$ is not weakly convergent, then it possesses at least two distinct weak sequential cluster points. From Lemma 4, we know that at most one weak sequential cluster point of $\{z_i\}$ may be exterior to $\overline{\text{co}}\,(q_y)$, and therefore at least one weak sequential cluster point of $\{z_i\}$ is in $\overline{\text{co}}\,(q_y)$. It follows from Lemma 3 that all the weak sequential cluster points of $\{z_i\}$ must be in $\overline{\text{co}}\,(q_y)$.

Lemma 6. If z_* does not belong to S_y, then there exists a continuous linear functional $f(\cdot)$ on X and constants c and $\varepsilon>0$ such that

$$f(z_*) \leqq c-\varepsilon < c \leqq f(S_y).$$

The sequence $\{z_i\}$ converges weakly to z_*, so therefore there exists an integer k such that for all $i \geqq k$,

$$f(z_i) \leqq c - \frac{\varepsilon}{2} < c \leqq f(y_i).$$

Now,

$$f(z_{i+1}) = f(z_i) + \lambda_i (f(y_i) - f(z_i)),$$

so for all $i \geqq k$,

$$f(z_{i+1}) \geqq f(z_i) + \lambda_i \frac{\varepsilon}{2}.$$

The assumption that $\sum \lambda_i$ diverges then implies that $\{f(z_i)\}$ is unbounded from above. We know that $\{f(z_i)\}$ is a convergent sequence and so must be bounded, and therefore the assumption that z_* does not belong to S_y leads to a contradiction; the lemma now follows.

Corollary 3. Since X is reflexive and $\{y_i\}$ is bounded, the weak sequential cluster point set q_y exists and is nonempty. Thus, asymptotically, $\overline{\text{co}}\,(q_y)$ can be used as S_y in Lemma 6.

Theorem 3. No proof needed.

Lemma 7. If $\sum \lambda_i$ diverges, then Corollary 1 implies that $\{z_i\}$ converges strongly. If $\sum \lambda_i$ is bounded from above then $\sum_{i=t}^{s-1} \lambda_i$ converges to 0 as t goes to infinity. Let s and t be two indices, with $s>t$. Then

$$\|z_s - z_t\| \leqq \|z_s - z_{s-1}\| + \|z_{s-1} - z_{s-2}\| + \cdots + \|z_{t+1} - z_t\|,$$

and

$$\|z_s - z_t\| \leqq \lambda_{s-1}\|y_{s-1} - z_{s-1}\| + \lambda_{s-2}\|y_{s-2} - z_{s-2}\| + \cdots + \lambda_t \|y_t - z_t\|.$$

The sequences $\{y_i\}$ and $\{z_i\}$ are bounded and there exists a scalar α such that

$$\|y_i - z_i\| \leqq \alpha$$

for all $i = 0, 1, 2, \cdots$. It follows that

$$\|z_s - z_t\| \leqq \alpha \sum_{i=t}^{s-1} \lambda_i$$

and the fact that $\sum_{i=t}^{s-1} \lambda_i$ converges to 0 as t increases implies that $\{z_i\}$ is a Cauchy sequence which converges strongly to some point z_*.

Lemma 8. If $\sum \lambda_i$ diverges, then Corollary 2 implies that $\{z_i\}$ converges weakly. If $\sum \lambda_i$ is bounded from above, then there exists an index k such that λ_i is in $[0; 1]$ for all $i \geqq k$, and using Lemma 5 we see that $\{z_i\}$ must be weakly convergent.

Theorem 4. We observe that

$$\|z_{i+1} - z_i\| = \lambda_i \|y_i - z_i\|.$$

Since $\lambda_i \geqq \lambda_m$ for all i,

$$\|y_i - z_i\| \leqq \frac{1}{\lambda_m} \|z_{i+1} - z_i\|. \tag{11}$$

But $\{z_i\}$ is bounded so that the right-hand side of (11) is bounded for all i. Thus the distance between y_i and z_i is bounded for all i, and so $\{y_i\}$ is bounded.

Corollary 4. By the triangle inequality we may write

$$\|y_i - z_*\| \leqq \|y_i - z_i\| + \|z_i - z_*\|,$$

and using (11), we obtain

$$\|y_i - z_*\| \leqq \frac{1}{\lambda_m} \|z_{i+1} - z_i\| + \|z_i - z_*\|. \tag{12}$$

Since $\{z_i\}$ converges strongly to z_*, the right-hand side of (12) converges to 0, and we conclude that $\{y_i\}$ converges strongly to z_*.

Corollary 5. Let $f(\cdot)$ be a continuous linear functional on X. Then

$$|f(y_i - z_*)| \leqq \frac{1}{\lambda_m} |f(z_{i+1} - z_i)| + |f(z_i - z_*)|. \tag{13}$$

By assumption, the sequence $\{z_i\}$ converges weakly to z_* and therefore the right-hand side of (13) converges to 0. Thus $\{f(y_i)\}$ converges to $f(z_*)$ for all continuous linear functionals $f(\cdot)$ on X and so $\{y_i\}$ converges weakly to z_*.

Theorem 5. If $\{z_i\}$ converges strongly to z_*, then by Corollary 4, $\{y_i\}$ converges strongly to z_*. Since $A(\cdot)$ is strongly sequentially closed, z_* is in $A(z_*)$.

Theorem 6. *If* $\{z_i\}$ converges weakly to z_*, then by Corollary 5, $\{y_i\}$ converges weakly to z_*. Since $A(\cdot)$ is weakly sequentially closed, z_* is in $A(z_*)$.

Theorem 7. It is sufficient to prove this result for the case where $\{z_i\}$ converges weakly to z_*. We now show that for all $\varepsilon > 0$, there exists an index k such that y_i is in $\bar{B}(A(z_*), \varepsilon)$ for all $i \geqq k$. In other words, for all continuous linear functionals $f(\cdot)$ on X and for all $\varepsilon > 0$, there exists an index k such that $f(y_i)$ is in $f(\bar{B}(A(z_*), \varepsilon))$ for all $i \geqq k$. Suppose this is not true: That for some linear continuous $f(\cdot)$ and for all k there is some $\varepsilon > 0$ such that $f(y_i)$ is not in $f(\bar{B}(A(z_*), \varepsilon))$ for some $i \geqq k$. But $\{y_i\}$ is a bounded sequence in reflexive Banach space and so possesses a weakly convergent subsequence $\{y_i\}_K$ such that $\{f(y_i)\}_K$ converges to $f(y_*)$, where $f(y_*)$ is not in $f(A(z_*))$,

i.e., y_* is not in $A(z_*)$. But since $\{z_i\}$ converges weakly to z_*, the subsequence $\{z_i\}_K$ has a weak limit z_* and y_* is not in $A(z_*)$. This contradicts the fact that $A(\cdot)$ is weakly sequentially closed. Hence for all $\varepsilon > 0$, there is some index k such that y_i is in $\bar{B}(A(z_*), \varepsilon)$ for all $i \geqq k$. Moreover, $\bar{B}(A(z_*), \varepsilon)$ is a closed and convex set. The result of the theorem is thus an immediate consequence of Lemma 6.

Corollary 6. No proof needed.

REFERENCES

[1] J. R. BLUM, *Multidimensional stochastic approximation methods*, Ann. Math. Statist., 25 (1954), pp. 737–744.

[2] E. M. BRAVERMAN AND L. I. ROZONOER, *Convergence of random processes in learning machine theory. Part I*, Automat. Remote Control, 1 (January 1969), pp. 44–64.

[3] ———, *Convergence of random processes in learning machine theory. Part II*, Automat. Remote Control, 3 (March 1969), pp. 386–402.

[4] F. E. BROWDER AND W. V. PETRYSHYN, *Construction of fixed points of nonlinear mappings in Hilbert space*, J. Math. Anal. Appl., 20 (1967), pp. 197–228.

[5] W. G. DOTSON, *On the Mann iterative process*, Trans. Amer. Math. Soc., 149 (1970), pp. 65–73.

[6] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* I: *General Theory*, Interscience, New York, 1967.

[7] J. C. DUNN, *Iterative construction of fixed points for multivalued operators of the monotone type*, J. Funct. Anal., 27 (1978), pp. 38–50.

[8] M. EDELSTEIN, *A remark on a theorem of M. A. Krasnoselskii*, Amer. Math. Monthly, 73 (1966), pp. 509–510.

[9] T. L. HICKS AND J. D. KUBICEK, *On the Mann iteration process in a Hilbert space*, J. Math. Anal. Appl., 59 (1977), pp. 498–504.

[10] T. L. HICKS, *Fixed point theorems in locally convex spaces*, Pacific J. Math., 79 (1978), pp. 111–115.

[11] M. A. KRASNOSELSKII, *Two remarks on the method of successive approximations*, Uspekhi Mat. Nauk, X, 1(63) (1955), pp. 123–127. (In Russian.)

[12] P. K. KUHFITTIG, *The mean-value iteration for set-valued mappings*, Proc. Amer. Math. Soc., 80 (1980), pp. 401–405.

[13] H. J. KUSHNER, *Stochastic approximation algorithms for the local optimization of functions with nonunique stationary points*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 646–654.

[14] H. J. KUSHNER AND T. GAVIN, *Stochastic approximation type methods for constrained systems: algorithms and numerical results*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 349–357.

[15] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551–575.

[16] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[17] W. R. MANN, *Mean value methods in iteration*, Proc. Amer. Math. Soc., 4 (1953), pp. 506–510.

[18] E. A. NURMINSKII, *Convergence conditions for nonlinear programming algorithms*, Cybernetics, 8 (1972), pp. 959–962.

[19] ———, *Convergence conditions of stochastic programming algorithms*, Cybernetics, 9 (1973), pp. 464–468.

[20] E. M. PICARD, *Sur l'application des methodes d'approximations successives a l'etude de certaines equations differentielles ordinaires*, J. Mathematiques, 4 Serie, Tome IX, Fasc. III (1893), pp. 28–271.

[21] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[22] S. REICH, *Fixed point iteration of nonexpansive mappings*, Pacific J. Math., 60 (1975), pp. 195–198.

[23] B. E. RHOADES, *Fixed point iteration using infinite matrices*, Trans. Amer. Math. Soc., 196 (1974), pp. 161–176.

[24] H. ROBBINS AND S. MUNRO, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.

[25] A. N. SARKOVSKII, *Attracting and attracted sets*, Soviet Math. Dokl., 6 (1965), pp. 268–270.

[26] YA. Z. TSYPKIN, *Adaptation, training, and self organization*, Automat. Remote Control, 27 (1966), pp. 16–51.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 1, January 1983

0363-0129/83/2101-0008 $01.25/0

AN EFFICIENT METHOD OF FEASIBLE DIRECTIONS*

GERARD G. L. MEYER†

Abstract. This paper presents a new method of feasible directions which uses an efficient antizig-zagging scheme. At every iteration, the gradient of the cost function and the gradients of the active constraints (usually one) are computed, and the previously computed gradients of the almost active constraints are used to prevent zig-zagging.

Key words. nonlinear programming, feasible directions, antizig-zagging, active constraints, iterative

1. Introduction. The past ten years have seen the development of new methods for solving the general nonlinear programming problem, namely, augmented Lagrangian and multiplier methods and more recently extended Newton's methods [1], [2], [3], [6]–[11], [19]–[21]. These methods are efficient, but are not applicable in all cases: the iterates generated by the new methods are usually not feasible, the constraints must satisfy stringent continuity requirements, and the new methods require the evaluation of the gradients of all the constraints at every iteration.

Methods of feasible directions, which have been overshadowed by these new methods, may still be useful and competitive provided that: the constraint set has an interior or a relative interior [18]; the iterates are required to be feasible; the constraints do not satisfy strong continuity hypotheses; and the nature of the constraints is such that the computation of their gradients is very time consuming. Clearly, the feasible directions approach would also be more appealing if its efficiency could be increased.

This paper presents a new feasible directions algorithm which uses the evaluated gradients more efficiently than existing methods. The main difference between the new method and the more classical ones lies in the use of an efficient antizig-zagging scheme. The "almost active" constraints are selected according to a procedure which possesses several attractive properties: at each iteration, only the gradient of the cost function and the gradients of the active constraints need be evaluated; the redundant constraints are automatically ignored; and the sequence of iterates generated depends on the shape of the constraint set and not its description.

2. Preliminaries. Consider the following nonlinear programming problem:

Problem 1. Given $m+1$ maps $f^0(\cdot), f^1(\cdot), \cdots, f^m(\cdot)$ from E^n into E, let T be the subset of E^n defined by

$$T = \{z \mid f^j(z) \leqq 0, j = 1, 2, \cdots, m\}. \tag{1}$$

Find a point z in T such that

$$f^0(z) \leqq f^0(y) \tag{2}$$

for all y in T.

The solution set $D(P)$ of Problem 1 may be empty; when it is not, the characterization of the points in $D(P)$ is usually not easy. To remove these difficulties, it will be assumed that the maps involved in the description of Problem 1 satisfy a set of simplifying assumptions.

* Received by the editors March 24, 1981, and in revised form January 25, 1982.
† Electrical Engineering Department, The Johns Hopkins University, Baltimore, Maryland 21218.

Hypothesis 1.

(i) The maps $f^0(\cdot), f^1(\cdot), \cdots, f^m(\cdot)$ are continuously differentiable and convex.

(ii) The set T is nonempty and compact.

(iii) For every z in T, the set

$$\{\nabla f^j(z) | f^j(z) = 0, j = 1, 2, \cdots, m\} \tag{3}$$

is linearly independent.

When Hypothesis 1 is satisfied, the properties of Problem 1 are well known [5], [18], [24] and the results are given without proof.

LEMMA 1. *If Hypothesis* 1 *is satisfied, then*:

(i) *Problem* 1 *possesses at least one solution.*

(ii) *A point z in T is a solution to Problem* 1 *if and only if*

$$\min \{\max \{\langle \nabla f^j(z), h\rangle | j \text{ in } J(z, \infty)\} | h \text{ in } S\} = 0, \tag{4}$$

where S is a compact neighborhood of the origin in E^n, *and* $J(z, \infty)$ *is the set that contains the indices of all the active constraints at z and the index of the cost function, namely*

$$J(z, \infty) = \{j \text{ in } \{1, 2, \cdots, m\} | f^j(z) = 0\} \cup \{0\}. \tag{5}$$

Essentially, Lemma 1 says that a point z in T is a solution to Problem 1 if and only if the origin of E^n belongs to the convex hull of the set

$$G(z) = \{\nabla f^j(z) | j \text{ in } J(z, \infty)\}. \tag{6}$$

Let the point $NG(z)$ be the projection of the origin of E^n onto the convex hull of $G(z)$. Then a point z in T is a solution to Problem 1 if and only if the origin of E^n belongs to $NG(z)$ [23].

The class of methods of feasible directions examined in this paper is based on a specific approximation of the necessary conditions of optimality. Given a point z in T, and a scalar $\alpha > 0$, let $J(z, \alpha)$ be the index set defined by

$$J(z, \alpha) = \left\{ j \text{ in } \{1, 2, \cdots, m\} | f^j(z) + \frac{1}{\alpha} \geqq 0 \right\} \cup \{0\}. \tag{7}$$

Instead of trying to find a point z in T that satisfies (4), one can generate a sequence $\{z_i\}$ of points in T that satisfies

$$\min \{\max \{\langle \nabla f^j(z_i), h\rangle | j \text{ in } J(z_i, \alpha_i)\} | h \text{ in } S\} \geqq -\frac{1}{\beta_i} \tag{8}$$

for two given sequences of positive scalars, $\{\alpha_i\}$ and $\{\beta_i\}$. For each index i, the parameter α_i controls the precision with which the conditions of optimality for Problem 1 are approximated, and the parameter β_i assures the necessary degree of negativity of the inner product between the gradient of the appropriate maps $f^j(\cdot)$ and the direction of descent. The various schemes used for choosing the sequences $\{\alpha_i\}$ and $\{\beta_i\}$ are called antizig-zagging rules. Several such schemes have been presented by Zoutendijk [25], Zukhovitskii, Polyak and Primak [26], Polak [18], and Meyer [14], [17].

One should note that the class of methods under consideration does not involve all the constraints. At each iteration, a point z_i is found which satisfies (8), and thus the only constraints involved through their gradients are the active and "almost" active constraints.

Methods of feasible directions exist that involve all the constraints: Huard's method of centers [4], [12], [13], and Topkis and Veinott's method of feasible directions [22] provide two such examples. In view of the introductory comments, it is clear that these methods are not competitive with the latest ones proposed, and therefore the thrust of this paper is directed at methods based on (4) and (8).

The theoretical justification for the use of iterative methods based on (8) in solving Problem 1 was presented in [14], [16]:

THEOREM 1. *If Hypotheses* 1 *is satisfied and* Q *is a closed subset of* T *which contains no solution to Problem* 1, *then there exist scalars* $\varepsilon > 0$ *and* $\bar{\alpha} > 0$, *and a neighborhood* $N(Q)$ *of* Q *such that*

$$\min \{\max \{\langle \nabla f^j(z), h\rangle | j \text{ in } J(z, \alpha)\} | h \text{ in } S\} < -\varepsilon \tag{9}$$

for all z *in* $N(Q) \cap T$, *and for all* $\alpha > \bar{\alpha}$.

Theorem 1 implies that when Hypothesis 1 is satisfied, a point that satisfies (8) can be computed in a finite number of iterations, thus justifying the use of iterative approaches based on (8) for solving Problem 1.

3. A class of feasible directions methods. The existing feasible directions methods require that at every iteration i, a feasible direction h_i be computed. To insure that the direction h_i is feasible, and that zig-zagging does not occur, one must take into account the gradients of both the active and almost active constraints at iteration i. The direction h_i must be a descent direction, and therefore one must insure that the gradient of the cost function and h_i make the proper angle. It follows that at every iteration, one must compute the quantities $\nabla f^j(z_i)$ for all indices j in the set $J(z_i, \alpha_i)$, where z_i is the ith iterate and α_i is the current value of the controlling parameter α.

If one wishes to use a direction h_i that is feasible and that makes an obtuse angle with the gradient of the cost function, it is necessary to take into account the gradients of the active constraints at iteration i and the gradient of the cost function. Thus, when zig-zagging does not occur, the gradients of the almost active constraints need not be computed. The algorithm presented below computes only the gradients that are absolutely necessary, and uses the previously computed gradients of the almost active constraints to approximate the tangent cone of the constraint set at the current point z_i. Thus, the gradient of the cost function and the gradients of the active constraints are used to obtain a direction of descent that is feasible, and the gradients of the previously active constraints are used to prevent the zig-zagging effect. One should note that usually only one constraint is active at any iteration. Thus, if such an algorithm is shown to solve Problem 1, just two gradients need be computed at each iteration: the gradient of the cost function and the gradient of the active constraint.

The approach just described requires a modification of the way in which the almost active constraints are accounted for. At iteration i, one uses an index set $K(i, p(r_i))$, which consists of the cost function index, the indices of all the active constraints, and the indices of all the constraints that have been active "not too long ago" at points that are "not too far" from the current iterate. Thus, in order to compute $K(i, p(r_i))$, one must keep track of the "age" of the previously computed gradients with respect to the present index of iteration: to each constraint index j is associated a parameter a_i^j which takes only nonnegative integer values and which is computed so that

$$f^j(z_{i-a_i^j}) = 0. \tag{10}$$

Since an algorithm with an unbounded memory is not desirable, the constraint gradients that have not been evaluated prior to iteration i are ignored, as are those that have been evaluated at iterations prior to iteration $i-\bar{t}$ (where $\bar{t}$ is a properly chosen nonnegative integer).

The proposed algorithm is parametrized by two maps $p(\cdot)$ and $q(\cdot)$ from E into E, a scalar ρ, a subset S of E^n, and a pre-specified sequence of integers $\{t_i\}$ that must satisfy the assumption given below.

Hypothesis 2.

(i) The maps $p(\cdot)$ and $q(\cdot)$ are positively radially unbounded, i.e., to every scalar $y>0$ corresponds an integer $k>0$ such that $q(x)>y$ and $q(x)>y$ for all $x>k$.

(ii) The scalar ρ is strictly positive.

(iii) The set S is a compact neighborhood of the origin in E^n.

(iv) $0 \leqq t_i \leqq \bar{t}$.

The algorithm consists of four steps, which have been subdivided into sub-steps for easier presentation.

Step 1 initializes the variables used in the algorithm. An initial feasible point z_1 in T is computed, and the gradient of the cost function and the gradients of the active constraints (usually one) are evaluated at the point z_1.

In Step 2, the feasible direction h_i is found, and the antizig-zagging parameter r_i is adjusted.

Step 3 contains a procedure for determining the step length so that the next iterate is feasible and the cost function is decreased.

In Step 4, the value of the gradient of the cost function and the gradients of the active constraints (usually one) are determined, and the ages of the various gradients are updated.

The maps $p(\cdot)$ and $q(\cdot)$ are used in conjunction with an auxiliary parameter r_i adjusted by the algorithm itself; the roles played by $p(r_i)$ and $q(r_i)$ are similar to those played by the parameters α and β discussed in the preliminaries. Thus, $p(r_i)$ controls the way in which the conditions of optimality for Problem 1 are approximated and $q(r_i)$ controls the precision with which the approximate optimality conditions are satisfied. The scalar ρ is introduced only to insure that Step 3a is well defined, even when $h_i = 0$. The set S is an arbitrary compact neighborhood of the origin in E^n, which is usually chosen to be a polyhedron [18] or a sphere [23]. In the first case, the computation of the feasible direction h_i in Step 2b reduces to solving a linear programming problem, and in the second case, it reduces to solving a quadratic programming problem.

ALGORITHM 1. The sequence $\{t_i\}$ in E is given.

Step 1a. Compute a point z_1 in T.

Step 1b. Set $b_1^j = \nabla f^j(z_1)$ and set $a_1^j = 0$ for all j in $J(z_1, \infty)$.

Step 1c. Set $b_1^j = 0$ and set $a_1^j = \bar{t}+1$ for all j not in $J(z_1, \infty)$.

Step 1d. Set $r_1 = 1$ and set $i = 1$.

Step 2a. Set $K(i, p(r_i)) = \{j | a_i^j \leqq t_i \text{ and } \|z_i - z_{i-a_i^j}\| \leqq 1/p(r_i)\}$.

Step 2b. Compute h_i in S such that for all h in S

$$\max \{\langle b_i^j, h_i\rangle | j \text{ in } K(i, p(r_i))\} \leqq \max \{\langle b_i^j, h\rangle | j \text{ in } K(i, p(r_i))\}.$$

Step 2c. If $\max \{\langle b_i^j, h_i\rangle | j \text{ in } K(i, p(r_i))\} > -1/q(r_i)$, set $r_i = r_i + 1$ and go to Step 2a; otherwise, go to Step 3a.

Step 3a. Compute $\lambda_i^j = \max \{\lambda \text{ in } [0; \rho] | f^j(z_i + \lambda h_i) \leqq 0\}$ for $j = 1, 2, \cdots, m$, and let $\mu_i = \min \{\lambda_i^1, \lambda_i^2, \cdots, \lambda_i^m\}$.

Step 3b. Compute λ_i^0 in $[0; \mu_i]$ such that

$$f^0(z_i + \lambda_i^0 h_i) \leqq f^0(z_i + \lambda h_i)$$

for all λ in $[0; \mu_i]$, and set $z_{i+1} = z_i + \lambda_i^0 h_i$.
Step 4a. Set $b_{i+1}^j = \nabla f^j(z_{i+1})$ and set $a_{i+1}^j = 0$ for all j in $J(z_{i+1}, \infty)$.
Step 4b. Set $b_{i+1}^j = b_i^j$ and set $a_{i+1}^j = \min(a_i^j + 1, \bar{t} + 1)$ for all j not in $J(z_{i+1}, \infty)$.
Step 4c. Set $r_{i+1} = r_i$, set $i = i + 1$, and go to Step 2a.

The sequence $\{t_i\}$ plays a major role in the antizig-zagging scheme. It is demonstrated below that if $t_i = 0$ for all i, and if the sequence $\{z_i\}$ generated by Algorithm 1 does not converge, then every cluster point of $\{z_i\}$ is a solution to Problem 1; if $t_i = 0$ for all i, and the sequence $\{z_i\}$ converges to some z_*, then z_* may or may not be a solution to Problem 1. On the other hand, if t_i is large enough (i.e., $t_i \geqq m - 1$ for all i, where m is the number of constraints), then every cluster point of every sequence generated by Algorithm 1 is a solution to Problem 1. The sequence $\{t_i\}$ controls the "memory" of the algorithm and is thus an important part of the antizig-zagging scheme.

4. Analysis of the algorithm. Algorithm 1, as presented in the preceding section, does not possess a stop rule. Nevertheless, the algorithm may generate finite sequences since it may "jam" in Step 2. Thus, before examining the asymptotic properties of the algorithm, one must analyze the properties of the finite sequences that may be generated by the algorithm.

LEMMA 2. *Suppose that Hypotheses* 1 *and* 2 *are satisfied and that* $\{z_i\}$ *is a sequence generated by Algorithm* 1. *If* z_k *is a solution to Problem* 1 *for some index* k, *the sequence* $\{z_i\}$ *is finite and* z_k *is the last point of the sequence. If* z_k *is not a solution to Problem* 1, *Algorithm* 1 *generates the point* z_{k+1} *after a finite number of adjustments of the parameter* r *in Step* 2.

Proof. (i) Suppose that z_k is a solution to Problem 1. Then

$$\min\{\max\{\langle \nabla f^j(z_k), h\rangle | j \text{ in } J(z_k, \infty)\} | h \text{ in } S\} = 0. \tag{11}$$

By construction,

$$K(k, p(r)) \supseteq J(z_k, \infty) \tag{12}$$

for all integers r, and

$$\nabla f^j(z_k) = b_k^j \tag{13}$$

for all j in $J(z_k, \infty)$. Thus,

$$\min\{\max\{\langle b_k^j, h\rangle | j \text{ in } K(k, p(r))\} | h \text{ in } S\} = 0, \tag{14}$$

and

$$\max\{\langle b_k^j, h_k\rangle | j \text{ in } K(k, p(r))\} > -\frac{1}{q(r)} \tag{15}$$

for all integers r. It follows that Algorithm 1 continues to adjust the parameter r in Step 2, and the point z_{k+1} is not generated. Thus, if z_k is a solution to Problem 1, the point z_k is the last element of the sequence $\{z_i\}$.

(ii) Suppose that z_k is not a solution to Problem 1: then,

$$\min\{\max\{\langle \nabla f^j(z_k), h\rangle | j \text{ in } J(z_k, \infty)\} | h \text{ in } S\} < 0. \tag{16}$$

Equation (16) and part (i) of Hypothesis 2 imply that a scalar $\bar{r}_1$ exists such that

$$\min\{\max\{\langle \nabla f^j(z_k), h\rangle | j \text{ in } J(z_k, \infty)\} | h \text{ in } S\} \leqq -\frac{1}{q(r)} \tag{17}$$

for all scalars $r \geqq \bar{r}_1$. The subsequence $\{z_1, z_2, \cdots, z_k\}$ contains a finite number of points, and therefore a scalar $\bar{r}_2 \geqq \bar{r}_1$ exists such that $z_i = z_k$ whenever the index i belongs to the set

$$\left\{i = 1, 2, \cdots, k \,|\, \|z_k - z_i\| \leqq \frac{1}{p(r)}\right\} \tag{18}$$

and $r \geqq \bar{r}_2$. The definition of the vectors b_k^j implies that whenever $r \geqq \bar{r}_2$,

$$b_k^j = \nabla f^j(z_k) \tag{19}$$

for all j in $K(z_k, p(r))$. It follows that

$$\min\{\max\{\langle b_k^j, h\rangle | j \text{ in } K(z_k, p(r))\} | h \text{ in } S\} \leqq \frac{-1}{q(r)} \tag{20}$$

for all $r \geqq \bar{r}_2$, and therefore Algorithm 1 generates the appropriate value for the parameter r after a finite number of adjustments in Step 2. Thus, if z_k is not a solution to Problem 1, the point z_{k+1} is generated in a finite time.

One should note that Lemma 2 implies that the characteristic set of Algorithm 1 [15] is equal to the solution set of Problem 1 whenever Hypotheses 1 and 2 are satisfied.

The construction of the point z_{i+1} from the point z_i in Step 4 of Algorithm 1 and the continuity of the cost function $f^0(\cdot)$ directly imply the following result:

LEMMA 3. *Suppose that Hypotheses* 1 *and* 2 *are satisfied, and let* $\{z_i\}$ *be an infinite sequence generated by Algorithm* 1. *Then*:

(i) $f^0(z_{i+1}) \leqq f^0(z_i)$ *for all* i; *and*

(ii) *if a cluster point of* $\{z_i\}$ *is a solution to Problem* 1, *then all the cluster points of* $\{z_i\}$ *are solutions to Problem* 1.

The asymptotic properties of Algorithm 1 are now analyzed. First, it is proved that if the sequence $\{r_i\}$ is unbounded from above, then every cluster point of every sequence generated by Algorithm 1 is a solution to Problem 1. Then it is shown that the sequence $\{r_i\}$ is unbounded from above, whether or not the infinite sequence $\{z_i\}$ generated by Algorithm 1 is asymptotically regular. One should recall that a sequence $\{z_i\}$ is asymptotically regular if and only if the corresponding scalar sequence $\{\|z_i - z_{i-1}\|\}$ converges to 0.

LEMMA 4. *If* (i) *Hypotheses* 1 *and* 2 *are satisfied*; (ii) $\{z_i\}$ *is an infinite sequence generated by Algorithm* 1; *and* (iii) *the sequence* $\{r_i\}$ *is unbounded from above, then every cluster point* z_* *of* $\{z_i\}$ *is a solution to Problem* 1.

Proof. The parameter r_i can only be increased in Step 2 of Algorithm 1. Thus, the assumption that the sequence $\{r_i\}$ is unbounded from above implies that

$$\max\{\langle b_i^j, h_i\rangle | j \text{ in } K(i, p(r_i))\} > \frac{-1}{q(r_i)} \tag{21}$$

infinitely many times. The map $q(\cdot)$ is positively radially unbounded, and therefore an infinite subset L of the integers exists such that

$$\{\max\{\langle b_i^j, h_i\rangle | j \text{ in } K(i, p(r_i))\}\}_L \quad \text{converges to 0.} \tag{22}$$

The set T is compact, $K(i, p(r_i))$ is a subset of the finite set $\{0, 1, \cdots, m\}$ for all i, and it follows that z_*, K_*, and an infinite subset M of L exist such that

$$\{z_i\}_M \quad \text{converges to } z_*, \tag{23}$$

and

$$K(i, p(r_i)) = K_* \quad \text{for all } i \text{ in } M. \tag{24}$$

Let j be an index in K_*: then for all i in M

$$b_i^j = \nabla f^j(z_{i-a_i^j}), \tag{25}$$

$$\|z_i - z_{i-a_i^j}\| \leqq \frac{1}{p(r_i)}, \tag{26}$$

and

$$0 \leqq a_i^j \leqq t_i \leqq \bar{t}. \tag{27}$$

The map $p(\cdot)$ is positively radially unbounded; $\{r_i\}$ is unbounded from above; the map $f^j(\cdot)$ is continuously differentiable; therefore,

$$\{b_i^j\}_M \quad \text{converges to } \nabla f^j(z_*); \tag{28}$$

and hence,

$$\min \{\max \{\langle \nabla f^j(z_*), h\rangle | j \text{ in } K_*\} | h \text{ in } S\} = 0. \tag{29}$$

To show that z_* is a solution to Problem 1, it remains to be proven that K_* is a subset of $J(z_*, \infty)$.

Let j be in K_*; then j is in $K(i, p(r_i))$ for all i in M, i.e.,

$$\|z_i - z_{i-a_i^j}\| \leqq \frac{1}{p(r_i)} \tag{30}$$

and

$$a_i^j \leqq t_i \leqq \bar{t} \tag{31}$$

for all i in M. The subsequence $\{z_i\}_M$ converges to z_*; the map $p(\cdot)$ is positively radially unbounded; the sequence $\{r_i\}$ is unbounded from above; and therefore,

$$\{z_{i-a_i^j}\}_M \quad \text{converges to } z_*. \tag{32}$$

When $a_i^j \leqq \bar{t}$, then

$$f^j(z_{i-a_i^j}) = 0. \tag{33}$$

It follows that $f^j(z_*) = 0$ and hence, j is in $J(z_*, \infty)$.

The point z_* is a solution to Problem 1 (i.e., at least one cluster point of $\{z_i\}$ is a solution to Problem 1), and using Lemma 3 one concludes that every cluster point of $\{z_i\}$ is a solution to Problem 1.

LEMMA 5. *If* (i) *Hypotheses* 1 *and* 2 *are satisfied, and* (ii) $\{z_i\}$ *is an infinite sequence generated by Algorithm* 1 *that is not asymptotically regular, then the sequence* $\{r_i\}$ *is unbounded from above.*

Proof. The sequence $\{z_i\}$ is not asymptotically regular and thus an infinite subset L of the integers and a scalar $\delta > 0$ exist such that

$$\|z_{i+1} - z_i\| \geqq \delta \tag{34}$$

for all i in L.

It follows that for all the indices i in L

$$\max\{\langle b_i^j, h_i\rangle | j \text{ in } K(i, p(r_i))\} \leqq \frac{-1}{q(r_i)}, \tag{35}$$

$$0 \text{ is in } K(i, p(r_i)), \tag{36}$$

$$b_i^0 = \nabla f^0(z_i), \tag{37}$$

and

$$\mu_i \geqq \bar{\delta}, \tag{38}$$

where

$$\bar{\delta} = \frac{\delta}{\text{diameter of } S}. \tag{39}$$

Suppose that the sequence $\{r_i\}$ is bounded from above: then a scalar r and an integer k exist such that $r_i = r$ for all $i \geqq k$.

It follows that

$$\langle \nabla f^0(z_i), h_i\rangle \leqq \frac{-1}{q(r)} \tag{40}$$

and

$$\mu_i \geqq \bar{\delta} \tag{41}$$

for all $i \geqq k, i$ in L. Equations (40) and (41) imply that the sequence $\{f^0(z_i)\}$ is unbounded from below, which is impossible because $f^0(\cdot)$ is continuous on the compact set T. It follows that the sequence $\{r_i\}$ is unbounded from above whenever the infinite sequence $\{z_i\}$ is not asymptotically regular.

LEMMA 6. *If* (i) *Hypotheses* 1 *and* 2 *are satisfied*; (ii) $m-1 \leqq t_i \leqq \bar{t}$ *for all* i; *and* (iii) $\{z_i\}$ *is an infinite sequence generated by Algorithm* 1 *that is asymptotically regular, then the sequence* $\{r_i\}$ *is unbounded from above.*

Proof. The sequence $\{z_i\}$ is asymptotically regular and contained in the compact set T, every set in the sequence $\{K(i, p(r_i))\}$ is a subset of $\{0, 1, 2, \cdots, m\}$, and every element a_i^j is in $\{0, 1, 2, \cdots, \bar{t}+1\}$. It follows that an infinite subset L of the integers, a point z_* in T, sets $K_*, K_{*+1}, \cdots, K_{*+m}$, and elements $a_*^0, a_*^1, \cdots, a_*^m, a_{*+1}^0, a_{*+1}^1, \cdots, a_{*+1}^m, \cdots, a_{*+m}^0, a_{*+m}^1, \cdots, a_{*+m}^m$ exist such that:

(F1) The subsequences $\{z_i\}_L, \{z_{i+1}\}_L, \cdots, \{z_{i+m}\}_L$ converge to z_*.

(F2) $K(i, p(r_i)) = K_*, \quad K(i+1, p(r_{i+1})) = K_{*+1}, \cdots, K(i+m, p(r_{i+m})) = K_{*+m}$ for all i in L.

(F3) $a_i^j = a_*^j, a_{i+1}^j = a_{*+1}^j, \cdots, a_{i+m}^j = a_{*+m}^j$ for all $j = 0, 1, \cdots, m$ and for all i in L.

Assume that $\{r_i\}$ is bounded from above; then

(F4) A scalar r and an integer k_1 exist such that

$$r_i = r \tag{42}$$

for all $i \geqq k_1$.

(F5) The sequence $\{z_i\}$ is asymptotically regular and therefore an integer $k_2 > k_1$ exists such that

$$\|z_i - z_{i-a_i^j}\| \leqq \frac{1}{p(r_i)} \tag{43}$$

for all $i \geqq k_2$ and for all a_i^j in $\{0, 1, \cdots \bar{t}+1\}$. Hence, the index j is in $K(i, p(r_i))$ whenever $i \geqq k_2$ and $a_i^j \leqq m-1$.

(F6) By construction,

$$\langle b_i^j, h_i \rangle \leqq \frac{-1}{q(r)} \tag{44}$$

for all j in $K(i, p(r_i))$ and for all $i \geqq k_2$.

(F7) The sequence $\{\mu_i\}$ converges to 0. The index 0 is in $K(i, p(r_i))$ for all i, thus F6 implies that

$$\langle \nabla f^0(z_i), h_i \rangle \leqq \frac{1}{q(r)}.$$

for all $i \geqq k_2$, and if a subsequence of μ_i is bounded away from 0, then the sequence $f^0(z_i)$ must be unbounded from below, which is impossible.

The index 0 is in $J(z_i, \infty)$ for all i, and therefore, the index 0 is in K_*, $K_{*+1}, \cdots, K_{*+m}$. Facts F6 and F7 imply the existence of an index $j(1)$ in K_{*+1} which is not in K_*. Thus,

$$j(1) \neq 0, \tag{45}$$

and

$$a_{i+1}^{j(1)} = 0. \tag{46}$$

The existence of $j(1)$ simply means that in view of F6, there must be a constraint not taken into account through K_*, which comes into play to insure that F7 is satisfied.

By construction,

$$a_{i+s}^{j(1)} \leqq a_{i+1}^{j(1)} + s - 1 \tag{47}$$

and therefore, from F5 we know that $j(1)$ will be in $K_{*+2}, K_{*+3}, \cdots, K_{*+m}$.

Similarly, F6 and F7 imply the existence of an index $j(2)$ in K_{*+2} which is not in K_{*+1}. Thus,

$$j(2) \neq 0, \tag{48}$$

$$j(2) \neq j(1) \tag{49}$$

and

$$a_{i+2}^{j(2)} = 0. \tag{50}$$

One concludes, again using F5, that $j(2)$ will be in $K_{*+3}, K_{*+4}, \cdots, K_{*+m}$.
Pursuing this reasoning, one shows that the index set K_{*+m} contains the indices $0, j(1)$, $j(2), \cdots, j(m)$, which are all distinct, i.e.,

$$K_{*+m} = \{0, 1, \cdots, m\}. \tag{51}$$

From F6 and the fact that $\{z_i\}$ is asymptotically regular, it is clear that the inner product of the direction h_i with the gradient of the cost function and the gradients of every constraint is bounded from above by $-1/q(r)$ infinitely many times. This in turn

implies that the sequence $\{f(z_i\}$ is unbounded from below. Thus, the assumption that the sequence $(r_i\}$ is bounded from above leads to a contradiction, and one concludes that $\{r_i\}$ must be unbounded from above.

Using Lemmas 2–6, one obtains the following theorem, which shows that Algorithm 1 may be used to solve Problem 1.

THEOREM 2. *Suppose that* (i) *Hypotheses* 1 *and* 2 *are satisfied*; (ii) $m-1 \leqq t_i \leqq \bar{t}$ *for all* i; *and* (iii) $\{z_i\}$ *is a sequence generated by Algorithm* 1. *Then*:

(i) $\{r_i\}$ *is unbounded from above*;

(ii) *if* $\{z_i\}$ *is finite, then the last point of* $\{z_i\}$ *is a solution to Problem* 1;

(iii) *if* $\{z_i\}$ *is infinite, then* $\{z_i\}$ *possesses at least one cluster point, and every cluster point of* $\{z_i\}$ *is a solution to Problem* 1.

REFERENCES

[1] D. P. BERTSEKAS, *Combined primal-dual and penalty methods for constrained minimization*, this Journal, 13 (1975), pp. 521, 544.

[2] ———, *Multiplier methods: a survey*, Automatica, 12 (1976), pp. 133–145.

[3] ———, *On penalty and multiplier methods for constrained minimization*, this Journal, 14 (1976), pp. 216–235.

[4] BUI-TRONG-LIEU AND P. HUARD, *La methode des centres dans un espace topologique*, Numer. Math. 8 (1966), pp. 65–67.

[5] V. F. DEMYANOV AND A. M. RUBINOV, *The minimization of a smooth convex functional on a convex set*, this Journal, 5 (1967), pp. 280–294.

[6] J. E. DENNIS AND J. J. MORE, *Characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comput., 28 (1974), pp. 549–560.

[7] ———, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.

[8] U. M. GARCIA-PALOMARES AND O. L. MANGASARIAN, *Superlinearly convergent quasi-Newton algorithms for nonlinearly constrained optimization problems*, Math. Programming, Vol. 11 (1976), pp. 1–13.

[9] S. P. HAN, *Superlinearly convergent variable metric algorithms for general nonlinear programming problems*, Math. Programming, 11 (1976), pp. 263–282.

[10] ———, *Dual variable metric algorithms for constrained optimization*, this Journal, 15 (1977), pp. 546–565.

[11] M. R. HESTENES, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.

[12] P. HUARD, *Programmation mathematique convexe*, RAIRO 7 (1968), pp. 43–59.

[13] ———, *Methode des centres et methodes des centres par majorations*, E.D.F., Bull. Direction Études Rech. Sér. C Math. Inform., 2 (1970), pp. 33–52.

[14] G. G. L. MEYER, *A drivable method of feasible directions*, this Journal, 11 (1973), pp. 113–118.

[15] ———, *A canonical structure for iterative procedures*, Math. Anal. Appl., 52 (1975), pp. 120–128.

[16] ———, *A finitely solvable class of approximating problems*, this Journal, 15 (1977), pp. 400–406.

[17] ———, *Methods of feasible directions with increased gradient memory*, Lecture Notes in Control and Information Science, 7, Part 2, J. Stoer, ed., Springer-Verlag, Berlin, 1978, pp. 87–93.

[18] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[19] M. J. D. POWELL, *Algorithms for nonlinear constraints that use Lagrangian functions*, Math. Programming, 14 (1978), pp. 224–248.

[20] R. T. ROCKAFELLAR, *The multiplier method of Hestenes and Powell applied to convex programming*, J. Optim. Theory Appl., 12 (1973), pp. 555–562.

[21] R. A. TAPIA, *Diagonalized multiplier methods and quasi-Newton methods for constrained optimization*, J. Optim. Theory Appl., 22 (1977), pp. 135–194.

[22] D. M. TOPKIS AND A. VEINOTT, JR., *On the convergence of some feasible directions algorithms for nonlinear programming*, this Journal, 5 (1967) pp. 268–279.

[23] P. WOLFE, *A method of conjugate subgradients for minimizing nondifferentiable functions*, Math. Programming Study, 3 (1975), pp. 145–173.

[24] W. I. ZANGWILL, *Nonlinear Programming: A Unified Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[25] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

[26] S. I. ZUKHOVITSKII, R. A. POLYAK AND M. E. PRIMAK, *An algorithm for the solution of convex programming problems*, Doklady Akad Nauk SSSR USSR 153, 5 (1963), pp. 991–1000. (In Russian.)

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0001 $01.25/0

OPTIMAL ADAPTIVE CONTROL OF LINEAR-QUADRATIC-GAUSSIAN SYSTEMS*

P. R. KUMAR†

Abstract. We consider the problem of adaptively controlling an unknown linear-Gaussian system with a standard quadratic cost criterion, *including a control cost.* By means of a counterexample, it is shown that a commonly mentioned adapative control scheme can lead to severe problems. To overcome this, a *new* adaptive control law, based on biasing the usual least-squares parameter estimation criterion with a term favoring parameters associated with lower optimal costs, is introduced. A salient feature of this adaptive control scheme is its imperviousness to the closed-loop identification problem. Properties such as closed-loop system identification, convergence of the adaptive control law to an optimal control law, overall stability of the controlled system and optimality with respect to the long-term average cost of the adaptive controller are proved.

Key words. adaptive control, stochastic systems, linear-quadratic Gaussian systems, linear systems, optimal control

1. A preliminary counterexample. To see what can go wrong in a commonly mentioned adaptive control scheme consider the following counterexample. We have a system

$$x_{t+1} = ax_t + bu_t + e_{t+1},$$

where we know the value of (a, b) to be *either* $(0, -1)$ *or* $(1, 1)$, but we *do not know* which of these is the correct value. Let $\{e_t\}$ be a noise sequence of independent identically distributed $N(0, 1)$ random variables. Our goal is the minimization of the cost criterion

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} x_s^2 + 2u_s^2$$

for which we know that the *optimal* control law is

$$u_t = \begin{cases} 0 & \text{if } (a, b) = (0, -1), \\ (-\tfrac{1}{2})x_t & \text{if } (a, b) = (1, 1). \end{cases}$$

(See any book on the algebraic Riccati equation and the LQG problem, e.g., [14, p. 355].)

Since we do not know the value of (a, b) (except, of course that it is either $(0, -1)$ or $(1, 1)$), at each time instant t we make a *least-squares* (or equivalently, in this case, a maximum-likelihood) *estimate* based on the data $(x_0, u_0, x_1, u_1, \cdots, x_t)$ up to time t, and then we use the control law corresponding to that estimate to choose u_t. Thus,

if $\sum_{s=0}^{t-1} (x_{s+1} + u_s)^2 \leqq \sum_{s=0}^{t-1} (x_{s+1} - x_s - u_s)^2$, then choose $u_t = 0$,
and if $\sum_{s=0}^{t-1} (x_{s+1} + u_s)^2 > \sum_{s=0}^{t-1} (x_{s+1} - x_s - u_s)^2$, then choose $u_t = -\tfrac{1}{2}x_t$.

* Received by the editors September 25, 1981, and in revised form March 23, 1982. This research was supported by the U.S. Army Research Office under contract DAAG-29-80-K0038.

† Department of Mathematics, University of Maryland Baltimore County, Baltimore, Maryland 21228.

This adaptive control scheme is in many ways a very "natural" one. To see what can go wrong, note the following chain of implications:

$$\sum_{s=0}^{t-1}(x_{s+1}+u_s)^2 > \sum_{s=0}^{t-1}(x_{s+1}-x_s-u_s)^2$$

$$\Rightarrow u_t = -\tfrac{1}{2}x_t$$

$$\Rightarrow (x_{t+1}+u_t)^2 = (x_{t+1}-x_t-u_t)^2$$

$$\Rightarrow \sum_{s=0}^{t}(x_{s+1}+u_s)^2 > \sum_{s=0}^{t}(x_{s+1}-x_s-u_s)^2$$

$$\Rightarrow u_{t+1} = -\tfrac{1}{2}x_{t+1}$$

$$\vdots$$

$$\Rightarrow \sum_{s=0}^{k-1}(x_{s+1}+u_s)^2 > \sum_{s=0}^{k-1}(x_{s+1}-x_s-u_s)^2 \quad \text{for all } k \geqq t$$

$$\Rightarrow u_k = -\tfrac{1}{2}x_k \quad \text{for all } k \geqq t.$$

Thus, if at *any* time t the parameters (a, b) are estimated to be $(1, 1)$, then the parameter estimates will thereafter remain *unchanged* and the adaptive control law will "stick" at $u_k = -\frac{1}{2}x_k$ for all $k \geqq t$. This is clearly undesirable if the true value of (a, b) is $(0, -1)$.

To see that this can indeed happen with positive probability, suppose that $(a, b) = (0, -1)$ is indeed the true system and we start initially with $x_0 = 1$ and $u_0 = 0$. Then

$$u_1 = -\tfrac{1}{2}x_1 \Leftrightarrow (x_1+u_0)^2 > (x_1-x_0-u_0)^2 \Leftrightarrow e_1^2 > (e_1-1)^2 \Leftrightarrow e_1 > \tfrac{1}{2}.$$

Since $e_1 > \frac{1}{2}$ occurs with probability 0.31 (recall $e_1 \sim N(0, 1)$), the adaptive control law will "stick" with probability *at least* 0.31 at the *nonoptimal* (cost = 2 versus cost = 1 for the optimal control law) *control law* $u_t = -\frac{1}{2}x_t$.

The object of this paper is to prove rigorously and completely the optimality and stability of a new adaptive control law which will never run into difficulties such as these.

2. Problem statement, main results and discussion. The true system being controlled is described by

$$x_{t+1} = A(\theta^0)x_t + B(\theta^0)u_t + w_{t+1}, \tag{1}$$

where x_t is the state vector, u_t is the control vector and $\{w_t\}$ is a Gaussian stochastic process with

$$E[w_t] = 0 \quad \text{and} \quad E[w_t w_s'] = \delta_{ts} I \quad \text{for all } s, t. \tag{2}$$

We however do *not* know the value of θ^0 and only have knowledge of a *finite set* Θ one of whose elements is θ^0. Our goal is to obtain an adaptive control scheme for choosing $\{u_t\}$ such that the cost criterion

$$\lim_{t\to\infty} \frac{1}{t}\sum_{s=0}^{t-1} x_s'Qx_s + u_s'Ru_s \tag{3}$$

is a minimum. Here $R = R' > 0$ and $Q = C'C$. The *only* assumption we make is

$$\{A(\theta), B(\theta), C\} \text{ is controllable and observable for each } \theta \in \Theta. \tag{4}$$

2.1. Main results. Our adaptive control scheme will be based on making an estimate $\hat{\theta}_t = \hat{\theta}_t(x_0, u_0, x_1, u_1, \cdots, x_t)$ of the unknown parameter. The criterion for making the estimates *differs* from the usual least-squares criterion (and is described in § 3). After the generation of an estimate $\hat{\theta}_t$, a control input

$$u_t = K(\hat{\theta}_t)x_t$$

is applied where $K(\theta)$ is the (unique) optimal feedback gain matrix for the parameter θ. Our main results are:

(5)

(i) *Closed-loop identification takes place* (Theorem 7). For a.e. ω and every $\theta \in \Theta$:

$$\limsup_{t\to\infty} \frac{1}{\ln t} \sum_{s=0}^{t-1} 1(\hat{\theta}_s(\omega) = \theta) > 0 \Rightarrow A(\theta) + B(\theta)K(\theta) = A(\theta^0) + B(\theta^0)K(\theta).$$

(ii) *Nonoptimal control inputs are applied very, very rarely* (Theorem 8).

$$\lim_{t\to\infty} \frac{1}{\ln t} \sum_{s=0}^{t-1} 1(u_s \neq K(\theta^0)x_s) = 0 \quad \text{a.s.}$$

(iii) *The closed-loop system is stable* (Theorem 12). For every $p \in [1, \infty)$ there is an $M(p) < \infty$ such that

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \|x_s\|^p + \|u_s\|^p \leqq M(p) < \infty \quad \text{a.s.}$$

(iv) *Optimal cost performance is achieved* (Theorem 14).

$$\lim_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} x_s'Qx_s + u_s'Ru_s = J(\theta^0) \quad \text{a.s.},$$

where $J(\theta^0)$ is the optimal cost achievable for the true system if one knew the value of θ^0 at the start.

2.2. Discussion. As formulated here, the problem under consideration belongs to the realm of adaptive control of linear stochastic systems with complete state observations. In recent years considerable attention has been paid and much success obtained [1]–[9] with regard to certain problems in this general area. Notable results have been obtained for the so-called minimum-variance adaptive control problem where the control weighting matrix R in (3) is 0. Thus, the goal of such adaptive control problems is only to minimize the output variance.

Regarding the present case $R > 0$, the available results are, to the author's knowledge, inconclusive. The results, e.g., [2], [6], are contingent on assumptions regarding the stability of the closed-loop system whose validity is unknown. Specifically, for example, it is not known whether available adaptive control laws which are appropriate for the case $R > 0$ will even yield a closed-loop system for which $\{(1/t)\sum_{s=0}^{t-1} x_s'Qx_s + u_s'Ru_s\}$ is bounded. A notable exception is some work of Mandl, see [9, § 7.7], and [19], [20], but the limitation here is that the matrix $B(\theta^0)$ is known and only $A(\theta^0)$ is unknown.

In comparing our problem formulation and results obtained with those for the much studied case $R = 0$, [1]–[8], we see that our problem formulation is limited in at least two respects: (i) $\{w_t\}$ is not temporally correlated; (ii) Θ is finite. On the other hand the results obtained here are stronger. For example, our results (5.i), (5.ii) dealing with closed-loop identification and convergence of the adaptive control are not mentioned in [7]. Moreover, our stability result (5.iii) is also stronger than the corresponding result in [7] and even our optimality result (5.iv) appears a little more relevant.

The net conclusion that may be drawn is that for the limited problem examined there is now a fairly complete theory for the case $R>0$.

A notable feature of our treatment is the difference between our adaptive control law and those previously studied for linear systems. The specific feature which distinguishes it is the manner in which the parameter estimates are made. The usual least-squares criterion is modified by the addition of a term depending on the *optimal cost* associated with a parameter in such a way as to be biased in favor of parameter estimates with lower costs. This adaptive control law was first introduced in the context of adaptive control of Markov Chains in [10], [11] and general Markovian systems in [12] and was proved there to be optimal for such systems. In fact, it was designed precisely to circumvent the closed-loop identification problem central to the counter-example of § 1 which also occurs in the context of Markov chains; see [13].

Regarding the practicality of the present assumption of a *finite* set Θ of possible parameters, many factors need to be considered. Its utility in a given situation depends on whether one is willing to quantize the possible set of parameter values to obtain a finite set, the amount of computational capability available, whether the quantized set includes a model which closely approximates the true system at hand (which, in fact, may not even be linear) and can ultimately be decided only on the basis of the practical problem being tackled. In any case, future attempts should be made to obtain complete theoretical results, as here, for the case where Θ is a compact set.

3. Description of adaptive control scheme. Under the condition (4) it is well known [14] that there exists a unique solution $P(\theta)$ to the algebraic Riccati equation

$$P(\theta)=A'(\theta)[P(\theta)-P(\theta)B(\theta)(B'(\theta)P(\theta)B(\theta)+R)^{-1}B'(\theta)P(\theta)]A(\theta)+Q \tag{6}$$

within the class of symmetric nonnegative definite matrices and also $P(\theta)>0$. Let

$$K(\theta):=-[B'(\theta)P(\theta)B(\theta)+R]^{-1}B'(\theta)P(\theta)A(\theta). \tag{7}$$

It is also known that $u_t=K(\theta)x_t$ is optimal for the problem of minimizing the cost criterion (3) for the system with parameter θ, and the resulting optimal cost (3) (both almost surely and in expectation, see [9]) is

$$J(\theta):=\operatorname{tr} P(\theta). \tag{8}$$

For future reference we also note the following facts all of which are obtainable either readily, or by a slight extension, from [14]:

(9)

(i) $P(\theta)$ may also be written as

$$P(\theta)=[A(\theta)+B(\theta)K(\theta)]'P(\theta)[A(\theta)+B(\theta)K(\theta)]+K'(\theta)RK(\theta)+Q.$$

(ii) $A(\theta)+B(\theta)K(\theta)$ is stable, i.e., all its eigenvalues are in the open unit disk in the complex plane.

(iii) Within the class of matrices K for which $A(\theta)+B(\theta)K$ is stable, $K(\theta)$ is the *unique* optimal feedback gain matrix.

The "identification" problem now consists of choosing an estimate $\hat{\theta}_t$ based on the observed past history $(x_0, u_0, x_1, u_1, \cdots, x_t)$ and the adaptive control problem consists of choosing an input u_t to be applied to the system (1) based on the same history.

We begin by *choosing* an arbitrary function $o(t)$ such that

$$o(t)>0, \quad \lim_{t\to\infty} o(t)=+\infty, \quad \lim_{t\to\infty}\left(\frac{o(t)}{\ln t}\right)=0. \tag{10}$$

We define $\hat{\theta}_t$ by

(11)

$$\hat{\theta}_t := \begin{cases} \arg\min_{\theta\in\Theta} \left[o(t) \ln J(\theta) + \sum_{s=0}^{t-1} (x_{s+1} - A(\theta)x_s - B(\theta)u_s)'(x_{s+1} - A(\theta)x_s - B(\theta)u_s) \right] \\ \qquad\qquad \text{if } t \text{ is an even number,} \\ \hat{\theta}_{t-1} \quad \text{if } t \text{ is an odd number.} \end{cases}$$

(If more than one value of θ maximizes the expression in (11), then we shall assume some specified priority ordering of elements of Θ which enables us to choose between competing maximizers.) After making an estimate $\hat{\theta}_t$, the control input is chosen as

$$u_t := K(\hat{\theta}_t)x_t. \tag{12}$$

An important observation to make is that were it not for the presence of the term $o(t) \ln J(\theta)$ in (11), $\hat{\theta}_t$ would merely be a least-squares estimate of the unknown parameter and we have already seen that such an estimate can cause severe problems. As it stands, $\hat{\theta}_t$ will be mildly biased ("mild" because of the choice of $o(t)$ in (10)) in favor of those θ's for which $J(\theta)$ is small. In a sense to be precisely exhibited later, this biasing term becomes asymptotically negligible as $t \to +\infty$, and so asymptotically the estimates $\{\hat{\theta}_t\}$ will retain some of the valuable and desirable properties of least-squares estimates.

One possible generalization of (11) consists in replacing the term $o(t) \ln J(\theta)$ by $o(t) \ln f(J(\theta))$ where f is an arbitrary positive, strictly monotone increasing function. The ensuing analysis and all the results will continue to hold for such a generalization. Another relaxation consists of assuming "stabilizability" in (4), but we shall not pursue that here.

For the remainder of this paper, we shall let $(\Omega, \mathcal{F}, \mathcal{P})$ be the underlying probability space on which the random variables x_t, u_t, w_t, $\hat{\theta}_t$, etc. are defined. Elements of Ω will be denoted by "ω". We define $\mathcal{F}_t := \sigma(x_0, u_0, x_1, u_1, \cdots, x_t) = \sigma(x_0, w_1, w_2, \cdots, w_t)$ to be the σ-algebra generated by the history up to time t.

4. An important effect of cost biasing on the parameter estimates. In this section we show that one important effect of the biasing term $o(t) \ln J(\theta)$ in (11) is to eliminate elements of $\{\theta : J(\theta) > J(\theta^0)\}$ from occurring as limit (cluster) points of $\{\hat{\theta}_t\}$. For convenience, define for each $\theta \in \Theta$ the random variables,

$$V_t(\theta) := \sum_{s=0}^{t-1} (x_{s+1} - A(\theta)x_s - B(\theta)u_s)'(x_{s+1} - A(\theta)x_s - B(\theta)u_s), \tag{13}$$

$$D_t(\theta) := o(t) \ln J(\theta) + V_t(\theta). \tag{14}$$

LEMMA 1[1]. *If θ^* is a limit point of $\{\theta_t(\omega)\}_{t=1}^{\infty}$, then $J(\theta^*) \leqq J(\theta^0)$.*

Proof. It is easily calculated that

$$E[\exp -\tfrac{1}{2}\{V_{t+1}(\theta) - V_{t+1}(\theta^0)\} | \mathcal{F}_t] = \exp -\tfrac{1}{2}\{V_t(\theta) - V_t(\theta^0)\},$$

[1] This is a sample path result which we require to hold almost surely. Thus the qualifier "There exists a set $N \subseteq \Omega$, $\mathcal{P}(N) = 0$ such that *if* $\omega \in N^c$ and", needs to precede the statement of the lemma. However, for brevity, in this and all future statements dealing with sample path results such a qualifier will not be explicitly mentioned.

and so for each $\theta \in \Theta$, $\{\exp -\frac{1}{2}\{V_t(\theta) - V_t(\theta^0)\}, \mathcal{F}_t\}$ is a positive martingale which converges almost surely. Fixing ω, we have, therefore, that

$$0 \leqq \lim_{t\to\infty} \exp -\frac{1}{2}\{V_t(\theta, \omega) - V_t(\theta^0, \omega)\} < \infty \quad \text{for every } \theta \in \Theta.$$

From (14) it follows that if $J(\theta^*) > J(\theta^0)$ then

$$\lim_{t\to\infty} \exp -\frac{1}{2}\{D_t(\theta^*, \omega) - D_t(\theta^0, \omega)\} = 0. \tag{15}$$

By (11) however, $\hat{\theta}_t(\omega) = \arg\min_{\theta\in\Theta} D_t(\theta, \omega)$ for even t and so, in particular, $D_t(\hat{\theta}_t(\omega), \omega) \leqq D_t(\theta^0, \omega)$ for even t. If θ^* is a limit of $\{\hat{\theta}_t(\omega)\}$, then since Θ is finite, $\hat{\theta}_t(\omega) = \theta^*(\omega)$ for infinitely many even t's. Thus, for θ^* to be a limit point of $\{\hat{\theta}_t(\omega)\}$, we will need $D_t(\theta^*, \omega) \leqq D_t(\theta^0, \omega)$ infinitely often, which however contradicts (15). □

5. Asymptotic properties of parameter estimates. In this section we show that the addition of the bias term $o(t) \ln J(\theta)$ to (11) does not adversely affect what would otherwise be a least-squares estimator with useful consistency properties. Many preliminary results are needed before we can establish Theorem 7, the main result of this section. Define for each $\theta \in \Theta$ the random variables

$$\begin{aligned} &\phi_t(\theta) := [A(\theta^0) - A(\theta)]x_t + [B(\theta^0) - B(\theta)]u_t, \\ &\mu_t(\theta) := 1 + \sum_{s=1}^{t} \phi_s'(\theta)\phi_s(\theta), \qquad \mu_0(\theta) := 1, \\ &\lambda_t(\theta) := \sum_{s=0}^{t-1} \mu_s^{-1}(\theta)\phi_s'(\theta)w_{s+1}, \qquad \lambda_0(\theta) := 0. \end{aligned} \tag{16}$$

LEMMA 2. *Let $\theta \in \Theta$ be arbitrary. Then $\lim_{t\to\infty} \mu_t^{-1}(\theta) \sum_{s=0}^{t-1} \phi_s'(\theta)w_{s+1} = 0$ almost surely on the set $\{\omega \in \Omega: \lim_{t\to\infty} \mu_t(\theta, \omega) = +\infty\}$.*

Proof. This lemma is similar to that of Ljung [15]. The proof is given here only for the sake of completeness. Fix $\theta \in \Theta$. Clearly $E[\lambda_{t+1}(\theta)|\mathcal{F}_t] = \lambda_t(\theta)$ and so $\{\lambda_t(\theta), \mathcal{F}_t\}$ is a martingale. Since

$$\begin{aligned} E[\lambda_t^2(\theta)] &= E\sum_{s=1}^{t} E[\lambda_s^2(\theta) - \lambda_{s-1}^2(\theta)|\mathcal{F}_{s-1}] \\ &= E\sum_{s=1}^{t} E[(\lambda_s(\theta) - \lambda_{s-1}(\theta))^2|\mathcal{F}_{s-1}] \quad \text{(martingale property)} \\ &= E\sum_{s=1}^{t} \mu_{s-1}^{-2}(\theta)\phi_{s-1}'(\theta)\phi_{s-1}(\theta) \\ &= E\left[\phi_0'(\theta)\phi_0(\theta) + \sum_{s=1}^{t-1} \mu_s^{-2}(\theta)[\mu_s(\theta) - \mu_{s-1}(\theta)]\right] \\ &\leqq E\left[\phi_0'(\theta)\phi_0(\theta) + \sum_{s=1}^{t} \mu_s^{-1}(\theta)\mu_{s-1}^{-1}(\theta)[\mu_s(\theta) - \mu_{s-1}(\theta)]\right] \\ &= E\left[\phi_0'(\theta)\phi_0(\theta) + \sum_{s=1}^{t} [\mu_{s-1}^{-1}(\theta) - \mu_s^{-1}(\theta)]\right] \\ &\leqq E[\phi_0'(\theta)\phi_0(\theta) + 1], \end{aligned}$$

it follows that $\{\lambda_t(\theta), \mathcal{F}_t\}$ is an L_2-bounded martingale and so converges almost surely.

Hence, $-\infty < \sum_{t=0}^{\infty} \mu_t^{-1}(\theta)\phi_t'(\theta)w_{t+1} < \infty$ almost surely. Kronecker's lemma [16] now gives the desired result. □

LEMMA 3. *If $\{t_k\}$ is a subsequence of the even integers with $\hat{\theta}_{t_k}(\omega) = \theta^*$, then*

$$\lim_{t\to\infty}\left(\frac{1}{\ln t_k}\right)\sum_{s=0}^{t_k-1}\phi_s'(\theta^*,\omega)\phi_s(\theta^*,\omega) = 0.$$

Proof. From (11), $D_{t_k}(\theta^*,\omega) \leqq D_{t_k}(\theta^0,\omega)$ and so

$$\left(\frac{1}{\ln t_k}\right)[D_{t_k}(\theta^*,\omega) - D_{t_k}(\theta^0,\omega)] \leqq 0 \quad \text{for all } k. \tag{17}$$

A simple calculation using (14), (13) and (16) shows that

$$\begin{aligned}
&\limsup_{k\to\infty}\left(\frac{1}{\ln t_k}\right)[D_{t_k}(\theta^*,\omega) - D_{t_k}(\theta^0,\omega)] \\
&\quad = \limsup_{k\to\infty}\left\{\left(\frac{o(t_k)}{\ln t_k}\right)\ln\left[\frac{J(\theta^*)}{J(\theta^0)}\right]\right. \\
&\qquad \left. + \left(\frac{1}{\ln t_k}\right)\sum_{s=0}^{t_k-1}[\phi_s'(\theta^*,\omega)\phi_s(\theta^*,\omega) + 2\phi_s'(\theta^*,\omega)w_{s+1}(\omega)]\right\} \qquad (18)\\
&\quad = \limsup_{k\to\infty}\left\{\left(\frac{o(t_k)}{\ln t_k}\right)\ln\left[\frac{J(\theta^*)}{J(\theta^0)}\right] + \left(\frac{\ln(t_k-1)}{\ln t_k}\right)\left(\frac{\mu_{t_k-1}(\theta^*,\omega)}{\ln(t_k-1)}\right)\right. \\
&\qquad \left. \cdot\left[1 - \left(\frac{1}{\mu_{t_k-1}(\theta^*,\omega)}\right) + \left(\frac{2}{\mu_{t_k-1}(\theta^*,\omega)}\right)\sum_{s=0}^{t_k-1}\phi_s'(\theta^*,\omega)w_{s+1}(\omega)\right]\right\}.
\end{aligned}$$

Note that the first term on the RHS of (18) has limit 0 from (10). If $\limsup_{k\to\infty}(1/\ln t_k)\mu_{t_k-1}(\theta^*,\omega) > 0$, then $\lim_{k\to\infty}\mu_{t_k-1}(\theta^*,\omega) = +\infty$, and it follows from Lemma 2 that the second term on the RHS of (18) would have strictly positive limit superior. Thus the RHS would be >0 which contradicts (17). □

Define for each $\theta \in \Theta$ the matrices

$$F(\theta) := A(\theta^0) - A(\theta) + B(\theta^0)K(\theta) - B(\theta)K(\theta),$$
$$\Gamma(\theta) := F'(\theta)F(\theta)$$

and the random variables

$$\xi_t(\theta) := 1 + \sum_{s=1}^{t} 1(\hat{\theta}_s = \theta, s \text{ is odd}), \qquad \xi_0(\theta) := 1,$$
$$\gamma_t(\theta) := (x_t'\Gamma(\theta)x_t \wedge 1)1(\hat{\theta}_t = \theta, t \text{ is odd}),$$

where " $\wedge$ " denotes the minimum operation.

LEMMA 4. *If $\lim_{t\to\infty}\xi_t(\theta,\omega) = +\infty$, then*

$$\lim_{t\to\infty}\xi_t^{-1}(\theta,\omega)\sum_{s=1}^{t}\gamma_s(\theta,\omega) - E[\gamma_s(\theta)|\mathcal{F}_{s-1}](\omega) = 0.$$

Proof. For each $\theta \in \Theta$ define the random variable

$$\delta_t(\theta) := \sum_{s=1}^{t}\xi_s^{-1}(\theta)[\gamma_s(\theta) - E[\gamma_s(\theta)|\mathcal{F}_{s-1}]], \qquad \delta_0(\theta) := 0.$$

Note that if $\hat{\theta} = \theta$ for s odd then $\hat{\theta}_{s-1} = \theta$ by (11). Hence, $1(\hat{\theta}_s = \theta, s \text{ is odd}) = 1(\hat{\theta}_{s-1} = \theta, s \text{ is odd})$ thereby making $\xi_t(\theta)$ measurable with respect to $\mathcal{F}_{t-1}$. Hence, $\{\delta_t(\theta), \mathcal{F}_t\}$ is a martingale. Also by calculations resembling those for $E[\lambda_t^2(\theta)]$ in Lemma 2,

$$\begin{aligned}
E[\delta_t^2(\theta)] &= E\sum_{s=1}^{t} E[\xi_s^{-2}(\theta)[\gamma_s(\theta) - E[\gamma_s(\theta)|\mathcal{F}_{s-1}]]^2|\mathcal{F}_{s-1}] \\
&= E\sum_{s=1}^{t} \xi_s^{-2}(\theta) 1(\hat{\theta}_{s-1} = \theta, s \text{ is odd}) E[\{(x_s'\Gamma(\theta)x_s \wedge 1) \\
&\qquad - E[x_s'\Gamma(\theta)x_s \wedge 1|\mathcal{F}_{s-1}]\}^2|\mathcal{F}_{s-1}] \\
&\leqq E\sum_{s=1}^{t} \xi_s^{-2}(\theta) 1(\hat{\theta}_{s-1} = \theta, s \text{ is odd}) = E\sum_{s=1}^{t} \xi_s^{-2}(\theta)[\xi_s(\theta) - \xi_{s-1}(\theta)] \\
&\leqq 1.
\end{aligned}$$

Again, as in Lemma 2, the martingale convergence theorem applies and Kronecker's lemma gives the desired result. □

LEMMA 5. *Let w be a vector of independent mean zero, variance one normal random variables. If $F \neq 0$, then there exists $\varepsilon > 0$ such that*

$$E[(w+a)'F'F(w+a)] \geqq \varepsilon \quad \textit{for every vector } a.$$

Proof. Let $F = (f_{ij})$ and denote by $(Fa)_i$ the ith element of Fa. Since $F \neq 0$ there is some row, say the lth, for which $\sum_{j=1}^{n} (f_{lj})^2 > 0$. Now

$$\begin{aligned}
E[(w+a)'F'F(w+a) \wedge 1] &\geqq \text{Prob}\,((Fw+Fa)'(Fw+Fa) \geqq 1) \\
&\geqq \text{Prob}\,([(Fw)_l + (Fa)_l]^2 \geqq 1) \\
&= \text{Prob}\left(\sum_{j=1}^{n} f_{lj}w_j \geqq 1 - \sum_{j=1}^{n} f_{lj}a_j\right) \\
&\quad + \text{Prob}\left(\sum_{j=1}^{n} f_{lj}w_j \leqq -1 - \sum_{j=1}^{n} f_{lj}a_j\right) \\
&\geqq 2\int_1^{\infty} \left(\sqrt{2\pi \sum_{j=1}^{n} f_{lj}^2}\right)^{-1} \exp -\frac{1}{2}\left(\sum_{j=1}^{n} f_{lj}^2\right)^{-1} x^2\, dx,
\end{aligned}$$

since the variance of $\sum_{j=1}^{n} f_{lj}w_j$ is $\sum_{j=1}^{n} f_{lj}^2$. Taking ε to be the last integral, the proof is concluded. □

LEMMA 6. *If $F(\theta) \neq 0$ and $\lim_{t\to\infty} \xi_t(\theta, \omega) = +\infty$, then*

$$\liminf_{t\to\infty} \xi_t^{-1}(\theta, \omega) \sum_{s=1}^{t} \gamma_s(\theta, \omega) \geqq \varepsilon > 0.$$

Proof. By Lemma 5,

$$\begin{aligned}
E[\gamma_s(\theta)|\mathcal{F}_{s-1}] &= E[(x_s'F'(\theta)F(\theta)x_s \wedge 1) 1(\hat{\theta}_s = \theta, s \text{ is odd})|\mathcal{F}_{s-1}] \\
&= 1(\hat{\theta}_{s-1} = \theta, s \text{ is odd}) E[(A(\theta^0)x_{s-1} + B(\theta^0)u_{s-1} + w_s)' \\
&\qquad \cdot F'(\theta)F(\theta)(A(\theta^0)x_{s-1} + B(\theta^0)u_{s-1} + w_s) \wedge 1|\mathcal{F}_{s-1}] \\
&\geqq \varepsilon\, 1(\hat{\theta}_s = \theta, s \text{ is odd}).
\end{aligned}$$

Hence

$$\xi_t^{-1}(\theta,\omega)\sum_{s=1}^{t} E[\gamma_s(\theta)|\mathcal{F}_{s-1}] \geqq \xi_t^{-1}(\theta,\omega)\sum_{s=1}^{t} \varepsilon\, 1(\hat{\theta}_s=\theta, s \text{ is odd})$$

$$= \varepsilon\xi_t^{-1}(\theta,\omega)[\xi_t(\theta,\omega)-1],$$

and the result follows from Lemma 4. □

THEOREM 7. *If*

$$\limsup_{t\to\infty} (\ln t)^{-1}\sum_{s=0}^{t-1} 1(\hat{\theta}_s(\omega)=\theta^*)>0, \tag{19}$$

then

$$A(\theta^*)+B(\theta^*)K(\theta^*)=A(\theta^0)+B(\theta^0)K(\theta^*). \tag{20}$$

Proof. By assumption, therefore, there is a subsequence $\{t_k\}$ such that

$$\lim_{k\to\infty} (\ln t_k)^{-1}\xi_{t_k}(\theta^*,\omega)>0.$$

We may without loss of generality also suppose that $\hat{\theta}_{t_k}(\omega)=\theta^*$ for every k. Noting that $\hat{\theta}_s(\omega)=\theta^* \Rightarrow \phi_s(\theta^*,\omega)=F(\theta^*)x_s(\omega)$, we have

$$\begin{aligned}
&\liminf_{k\to\infty} (\ln t_k)^{-1}\sum_{s=0}^{t_k-1} \phi_s'(\theta^*,\omega)\phi_s(\theta^*,\omega)\\
&\quad\geqq \liminf_{k\to\infty} (\ln t_k)^{-1}\sum_{s=0}^{t_k-1} 1(\hat{\theta}_s(\omega)=\theta^*, s \text{ is odd})\,\phi_s'(\theta^*,\omega)\phi_s(\theta^*,\omega)\\
&\quad= \liminf_{k\to\infty} (\ln t_k)^{-1}\sum_{s=0}^{t_k-1} 1(\hat{\theta}_s(\omega)=\theta^*, s \text{ is odd})\, x_s'(\omega)F'(\theta^*)F(\theta^*)x_s(\omega) \qquad (21)\\
&\quad\geqq \liminf_{k\to\infty} (\ln t_k)^{-1}\sum_{s=1}^{t_k} \gamma_s(\theta^*,\omega)\\
&\quad= \left[\lim_{k\to\infty} (\ln t_k)^{-1}\xi_{t_k}(\theta^*,\omega)\right]\liminf_k \xi_{t_k}^{-1}(\theta^*,\omega)\sum_{s=1}^{t_k}\gamma_s(\theta^*,\omega).
\end{aligned}$$

The first term in the RHS of (21) is positive by assumption. If $F(\theta^*)\neq 0$, then the second term is also positive from Lemma 6. This would contradict Lemma 3. □

This result shows that if a limit point θ^* occurs with sufficient frequency, then it is *indistinguishable* from θ^0 under the feedback gain $K(\theta^*)$. Note that if a θ does *not* satisfy the condition of Theorem 7, then either (i) it occurs only finitely often or (ii) its occurrences are very, very rare. In some sense therefore the closed-loop consistency properties of such limit points are unimportant and may be neglected. Thus Theorem 7 provides valuable information regarding the limit points which occur with sufficient frequency for us to take note of their occurrence, and in particular, the closed-loop gain is identified for such limit points.

6. Convergence of adaptive control law. Since the true system corresponds to θ^0, a control input $u_t=K(\theta^0)x_t$ would be optimal. Here we show that for the present adaptive control law (11), (12) which is computed *without* knowledge of θ^0,

$$\lim_{t\to\infty} (\ln t)^{-1}\sum_{s=0}^{t-1} 1(u_s\neq K(\theta^0)x_s)=0 \quad \text{a.s.} \tag{22}$$

Note that this is weaker than claiming (which we do not)

$$u_t - K(\theta^0)x_t \to 0.$$

However for practical purposes (22) is sufficient since the time instants at which the control input u_t is *not* optimal are very, very rare. Later, in § 8, it will be shown that (22) leads to convergence of the actual cost incurred to the true optimal cost achievable if one knew θ^0 at the start.

For K such that $A(\theta)+B(\theta)K$ is stable (i.e., all eigenvalues strictly inside the unit disc) denote by $P(K, \theta)$ the unique (within the class of symmetric nonnegative definite matrices) positive definite solution of

$$P(K, \theta) = Q + K'RK + [A(\theta)+B(\theta)K]'P(K, \theta)[A(\theta)+B(\theta)K] \tag{23}$$

and set $J(K, \theta) := \operatorname{tr} P(K, \theta)$. Note that $J(K, \theta)$ represents the cost of using a feedback control law $u = Kx$ on the system θ. Note that

$$J(\theta) = J(K(\theta), \theta) \leqq J(K, \theta) \quad \text{whenever } A(\theta)+B(\theta)K \text{ is stable,} \tag{24}$$

because $K(\theta)$ is the (unique within the class of stabilizing feedback gains) optimal feedback gain for system θ.

THEOREM 8.

$$\lim_{t\to\infty} (\ln t)^{-1} \sum_{s=0}^{t-1} 1(K(\hat{\theta}_s) \neq K(\theta^0)) = 0 \quad a.s.,$$

and (22) *also holds.*

Proof. Fix ω and let θ^* satisfy (19). Then $J(K(\theta^*), \theta^0) = \operatorname{tr} P(K(\theta^*), \theta^0)$ is well defined since the LHS of (20) and therefore the RHS of (20) is stable. By (20) and uniqueness of solutions to (23) $J(K(\theta^*), \theta^0) = J(K(\theta^*), \theta^*)$. By (24) it therefore follows that

$$J(\theta^0) \leqq J(K(\theta^*), \theta^0) = J(K(\theta^*), \theta^*) = J(\theta^*). \tag{25}$$

But (19) also implies that θ^* is a limit point of $\{\hat{\theta}_t(\omega)\}$, and by Lemma 2, it follows that equality holds throughout (25). Thus, it follows that $K(\theta^*)$ is the *optimal* feedback gain for θ^0. But by the uniqueness property (9.iii),

$$K(\theta^*) = K(\theta^0). \tag{26}$$

Since (19) $\Rightarrow$ (26) it follows that,

$$\begin{aligned}
&\limsup_{t\to\infty} (\ln t)^{-1} \sum_{s=0}^{t-1} 1(K(\hat{\theta}_s(\omega)) \neq K(\theta^0)) \\
&\quad = \limsup_{t\to\infty} (\ln t)^{-1} \sum_{\{\theta^*:\, K(\theta^*)\neq K(\theta^0)\}} \cdot \sum_{s=0}^{t-1} 1(\hat{\theta}_s(\omega) = \theta^*) \\
&\quad \leqq \sum_{\{\theta^*:\, K(\theta^*)\neq K(\theta^0)\}} \limsup_{t\to\infty} (\ln t)^{-1} \sum_{s=0}^{t-1} 1(\hat{\theta}_s(\omega) = \theta^*) = 0.
\end{aligned}$$

Equation (22) now follows trivially. □

7. Stability results. In this section we demonstrate the stability of the overall system by showing that for every integer p there is an M with

$$\limsup_{t\to\infty} t^{-1} \sum_{s=0}^{t-1} \|x_s\|^p \leqq M < \infty \quad \text{a.s.} \tag{27}$$

This will be used later in § 8 to prove optimality of the cost incurred.

Note that Theorem 8 does *not* rule out the occurrence of an infinite set of times $\{t_k\}$ at which $K(\hat{\theta}_{t_k}) = K(\theta^0)$. At such times t_k the control input $u_{t_k} = K(\hat{\theta}_{t_k})x_{t_k}$ can very well be destabilizing in the sense that $A(\theta^0)+B(\theta^0)K(\hat{\theta}_{t_k})$ may be unstable. The purpose of this section is to show that the rarity of occurrence of such a set of times as given by Theorem 8 guarantees (27).

We first study the behavior of the scalar difference equation

$$y_{t+1} = \gamma_t y_t + e_{t+1} \tag{28}$$

and examine conditions on $\{\gamma_t\}$ and $\{e_t\}$ which ensure that $\limsup_{t\to\infty} t^{-1}\sum_{s=0}^{t-1} y_s < \infty$. Later in Theorem 12 we shall relate y_t to $\|x_t(\omega)\|^p$ and thus succeed in showing (27). The following conditions are imposed on (28):

(29)

(i) $$0 \leqq y_0 < \infty.$$

(ii) $0 < a < 1 \leqq b < \infty$ are arbitrary and for every t *either*

$$\gamma_t = a \quad \text{or} \quad \gamma_t = b.$$

(iii) m_t defined by $m_t := \sum_{s=0}^{t-1} 1(\gamma_s = b)$ satisfies $(\ln t)^{-1} m_t = 0$.

(iv) $\{e_t\}$ is a nonnegative sequence satisfying

$$\limsup_{t\to\infty} t^{-1} \sum_{s=0}^{t-1} e_s = \delta < \infty.$$

(v) There is an increasing $\{\psi_t\}$ with $e_t \leqq \psi_t$ for every t and

$$\limsup_{t\to\infty} (\ln t)^{-1} \ln \psi_t < 1.$$

LEMMA 9.

$$\sum_{s=0}^{t-1} y_s \leqq \left(\frac{1+b-a}{1-a}\right)\left(\frac{1}{1-a}\right) m_t \psi_t \left(\frac{b}{a}\right)^{m_t} + \left(\frac{1}{1-a}\right)\sum_{s=0}^{t-1} e_{s+1} + y_0\left(\frac{b}{a}\right)^{m_t}\left(\frac{1}{1-a}\right).$$

Proof. First consider the case $y_0 = 0$. Let $t_0 := 0$ and recursively set $t_{i+1} := \inf\{t > t_i : \gamma_t = b\}$. Then

$$y_t = \left(\sum_{j=t_i}^{t-1} \gamma_j\right) y_{t_i} + \sum_{j=t_i}^{t-1} e_{j+1} \prod_{l=j+1}^{t-1} \gamma_l, \qquad 0 < t_i < t < t_{i+1},$$

where throughout, by convention, vacuous products are set to 1 and vacuous sums to 0. Since $\gamma_t = a$ for $t_i < t < t_{i+1}$,

$$y_t = b a^{t-t_i-1} y_{t_i} + \sum_{j=t_i}^{t-1} e_{j+1} a^{t-j-1}, \qquad 0 < t_i < t < t_{i+1}$$

and so

$$\begin{aligned}
\sum_{t=t_i}^{t_{i+1}-1} y_t &= y_{t_i}\left[1 + b\sum_{t=t_i+1}^{t_{i+1}-1} a^{t-t_i-1}\right] + \sum_{t=t_i+1}^{t_{i+1}-1}\sum_{j=t_i}^{t-1} e_{j+1} a^{t-j-1} \\
&\leqq \left(1 + \left(\frac{b}{1-a}\right)\right) y_{t_i} + \sum_{j=t_i}^{t_{i+1}-2} e_{j+1} \sum_{t=j+1}^{t_{i+1}-1} a^{t-j-1} \\
&\leqq \left(\frac{1+b-a}{1-a}\right) y_{t_i} + \left(\frac{1}{1-a}\right)\sum_{j=t_i}^{t_{i+1}-2} e_{j+1}.
\end{aligned}$$

Hence,

$$\sum_{t=0}^{t_{i+1}-1} y_t = \sum_{k=0}^{i} \sum_{t=t_k}^{t_{k+1}-1} y_t \leqq \left(\frac{1+b-a}{1-a}\right) \sum_{k=0}^{i} y_{t_k} + \left(\frac{1}{1-a}\right) \sum_{j=0}^{t_{i+1}-2} e_{j+1}$$

$$= \left(\frac{1+b-a}{1-a}\right) \sum_{j=0}^{t_i} 1(\gamma_j = b) y_j + \left(\frac{1}{1-a}\right) \sum_{j=0}^{t_{i+1}-2} e_{j+1}$$

and so

$$\sum_{t=0}^{n} y_t \leqq \left(\frac{1+b-a}{1-a}\right) \sum_{k=0}^{n} 1(\gamma_k = b) y_k + \left(\frac{1}{1-a}\right) \sum_{j=0}^{n-2} e_{j+1}. \tag{30}$$

Now we bound y_k as follows:

$$y_t = \sum_{j=0}^{t-1} e_{j+1} \prod_{l=j+1}^{t-1} \gamma_l \leqq \psi_t \sum_{j=0}^{t-1} \prod_{l=j+1}^{t-1} \gamma_l$$

$$= \psi_t \sum_{j=0}^{t-1} a^{(t-j-1+m_{j+1}-m_t)} b^{(m_t-m_{j+1})} \tag{31}$$

$$= \psi_t \left(\frac{b}{a}\right)^{m_t} \sum_{j=0}^{t-1} \left(\frac{a}{b}\right)^{m_{j+1}} a^{t-j-1} \leqq \psi_t \left(\frac{b}{a}\right)^{m_t} \left(\frac{1}{1-a}\right).$$

Substituting (31) in (30)

$$\sum_{t=0}^{n} y_t \leqq \left(\frac{1+b-a}{1-a}\right) \sum_{k=0}^{n} 1(\gamma_k = b) \psi_k \left(\frac{b}{a}\right)^{m_k} \left(\frac{1}{1-a}\right) + \left(\frac{1}{1-a}\right) \sum_{j=0}^{n-2} e_{j+1}$$

$$\leqq \left(\frac{1+b-a}{1-a}\right)\left(\frac{1}{1-a}\right) \psi_n \left(\frac{b}{a}\right)^{m_n} \sum_{k=0}^{n} 1(\gamma_k = b) + \left(\frac{1}{1-a}\right) \sum_{j=0}^{n-2} e_{j+1} \tag{32}$$

$$= \left(\frac{1+b-a}{1-a}\right)\left(\frac{1}{1-a}\right) m_{n+1} \psi_n \left(\frac{b}{a}\right)^{m_n} + \left(\frac{1}{1-a}\right) \sum_{j=0}^{n-2} e_{j+1}.$$

If the initial condition $y_0 > 0$, then by linearity of (28) its contribution to $\sum_{t=0}^{n} y_t$ is

$$\sum_{t=0}^{n} y_0 \prod_{k=0}^{t-1} \gamma_k = \sum_{t=0}^{n} a^{(t-m_t)} b^{m_t} y_0 \leqq y_0 \left(\frac{b}{a}\right)^{m_n} \left(\frac{1}{1-a}\right). \tag{33}$$

The present result now follows by adding (32) and (33). □

LEMMA 10. *If* (29.i)–(29.v) *hold,*

$$\limsup_{n\to\infty} n^{-1} \sum_{t=0}^{n-1} y_t \leqq \left(\frac{\delta}{1-a}\right) < \infty.$$

Proof.

$$\limsup_{n\to\infty} n^{-1} m_n \psi_n \left(\frac{b}{a}\right)^{m_n} = \exp\left\{\limsup_{n\to\infty} \left[\ln \psi_n + \ln m_n + m_n \ln\left(\frac{b}{a}\right) - \ln n\right]\right\}$$

$$= \exp\left\{\limsup_{n\to\infty} \left[\left\{\left(\frac{\ln \psi_n}{\ln n}\right) + \left(\frac{\ln m_n}{\ln n}\right) + \left(\frac{m_n}{\ln n}\right) \ln\left(\frac{b}{a}\right) - 1\right\} \ln n\right]\right\}$$

$$= 0.$$

Similarly $\limsup_{n\to\infty} n^{-1} y_0 (b/a)^{m_n} (1/(1-a)) = 0$, and the results follows from (28.iv) and Lemma 9. □

LEMMA 11. *There exists an increasing sequence* $\{\psi_t\}$, $\limsup_{t\to\infty}(\ln \psi_t/\ln t)<1$ *with* $\mathcal{P}(\{\omega: \|w_t(\omega)\|^p > \psi_t \text{ i.o.}\}) = 0$ *for every* $p \in [1, \infty)$.

Proof. Note first that the particular norm used is irrelevant (since all norms on $\mathbb{R}^n$ are equivalent). Since w_t^i, the ith component of w_t is normal, by Chebyshev's inequality,

$$\text{Prob}\,(|w_t^i| \geqq |\alpha|) \leqq \frac{c}{|\alpha|^{2p}} \quad \text{for some } c>0.$$

Hence for any $\psi_t > 0$, taking the L_1 norm on $\mathbb{R}^n$,

$$\begin{aligned}\text{Prob}\,(\|w_t\|^p \geqq \psi_t) &= \text{Prob}\,(\|w_t\| \geqq \psi_t^{1/p}) \\ &= \text{Prob}\left(\sum_{i=1}^n |w_t^i| \geqq \psi_t^{1/p}\right) \leqq \sum_{i=1}^n \text{Prob}\,(|w_t^i| \geqq \psi_t^{1/p} n^{-1}) \\ &\leqq \sum_{i=1}^n c n^{2p} \psi_t^{-2}.\end{aligned}$$

The present result now follows by taking $\psi_t = t^{3/4}$ and using the Borel–Cantelli lemma. □

THEOREM 12. *For every* $p \in [1, \infty)$, *there exists an* $M(p) > 0$ *such that*

$$\limsup_{t\to\infty} \frac{1}{t} \sum_{s=0}^{t-1} \|x_s\|^p \leqq M(p) < \infty \quad a.s.$$

Proof. Note that for every $\varepsilon > 0$, a, b,

$$2ab = 2(\varepsilon a)(\varepsilon^{-1} b) \leqq (\varepsilon a)^2 + \varepsilon^{-2} b^2$$

and so

$$(a+b)^2 \leqq (1+\varepsilon^2) a^2 + (1+\varepsilon^{-2}) b^2.$$

Repeating this process n times, we get

$$(a+b)^{2^n} \leqq (1+\varepsilon^2)^{2^n-1} a^{2^n} + (1+\varepsilon^{-2})^{2^n-1} b^{2^n}.$$

Thus, since $x_{t+1} = [A(\theta^0) + B(\theta^0) K(\hat{\theta}_t)] x_t + w_{t+1}$, we have

$$\|x_{t+1}\|^{2^n} \leqq (1+\varepsilon^2)^{2^n-1} \|A(\theta^0) + B(\theta^0) K(\hat{\theta}_t)\|^{2^n} \|x_t\|^{2^n} + (1+\varepsilon^{-2})^{2^n} \|w_{t+1}\|^{2^n}. \tag{34}$$

Note that (34) is true regardless of what norm we assume on the statespace, as long as the norm on matrices is the corresponding induced operator norm. Let us now choose these norms carefully. Since $A(\theta^0) + B(\theta^0) K(\theta^0)$ is stable by (9.ii), its spectral radius is less than 1. By [17, § 1.3.6] there exists a norm on the state space such that under the corresponding induced operator norm $\|A(\theta^0) + B(\theta^0) K(\theta^0)\| < 1$. Hence, for some $\varepsilon > 0$, we will have

$$a := (1+\varepsilon^2)^{2^n-1} \|A(\theta^0) + B(\theta^0) K(\theta^0)\|^{2^n} < 1.$$

Now define $b := 1 + \max\{(1+\varepsilon^2)^{2^n-1} \|A(\theta^0) + B(\theta^0) K(\theta)\|^{2^n} : \theta \in \Theta\}$. Fixing ω, we see from (34) that

$$\|x_{t+1}(\omega)\|^{2^n} \leqq \gamma_t \|x_t(\omega)\|^{2^n} + e_{t+1},$$

where $\gamma_t := 1(K(\hat{\theta}_t(\omega)) = K(\theta^0)) a + 1(K(\hat{\theta}_t(\omega)) \neq K(\theta^0)) b$ and $e_{t+1} := (1+\varepsilon^{-2})^{2^n-1} \|w_{t+1}(\omega)\|^{2^n}$. Clearly, therefore, the solution of (28) is an upper-bound for $\|x_t(\omega)\|^{2^n}$. γ_t satisfies (29.ii) by definition and satisfies (29.iii) from Theorem 8. (29.iv) is satisfied

because of the applicability of the ergodic theorem to $\{\|w_{t+1}\|^{2^n}\}$. (29.v) holds due to Lemma 11 and also note that δ in (29.iv) does *not* depend on ω. Hence, by Lemma 10 we can deduce that

$$\lim_{t\to\infty} t^{-1}\sum_{s=0}^{t-1}\|x_s\|^{2^n} \leqq M < \infty \quad \text{a.s.}$$

The corresponding result for every $p \in [1, \infty)$ follow since $\|x_s\|^p \leqq 1 + \|x_s\|^{2^n}$ for $1 \leqq p \leqq 2^n$. Note also that the present result once established for a particular norm becomes true for every norm because of the equivalence of norms on finite dimensional spaces. □

8. Optimality of incurred cost. The last result which we wish to establish is that the cost of using the given adaptive control law is optimal and cannot be improved *even* if the value of the unknown parameter is known at the start.

LEMMA 13.

(35)

$$\text{(i)} \qquad \sum_{t=1}^{\infty} t^{-2}\|x_t\|^2 < \infty \quad a.s.,$$

$$\text{(ii)} \qquad \lim_{t\to\infty} t^{-1}\|x_t\|^2 = 0 \quad a.s.$$

Proof. Let $z_t := \sum_{s=1}^{t}\|x_s\|^{2p}$, $z_0 := 0$. Then

$$\begin{aligned}\left[\sum_{t=1}^{\infty} t^{-2}\|x_t\|^p\right]^2 &= \left[\sum_{t=1}^{\infty} t^{-1}\{t^{-1}(z_t - z_{t-1})^{1/2}\}\right] \\ &\leqq \left(\sum_{t=1}^{\infty} t^{-2}\right)\sum_{t=1}^{\infty} t^{-2}(z_t - z_{t-1}) \qquad \text{(Schwarz inequality)} \\ &\leqq \left(\sum_{t=1}^{\infty} t^{-2}\right)\left(\sum_{t=1}^{\infty} z_t(t^{-2} - (t+1)^{-2})\right) \\ &\leqq \left(\sum_{t=1}^{\infty} t^{-2}\right)\left(\sum_{t=1}^{\infty} t^{-3} z_t\right) < \infty \quad \text{a.s.},\end{aligned}$$

where the convergence follows from Theorem 12. With $p = 2$, (35.i) is proved. With $p = 4$, we obtain $\sum_{t=1}^{\infty} t^{-2}\|x_t\|^4 < \infty$ almost surely from which we have $t^{-2}\|x_t\|^4 \to 0$ almost surely. □

THEOREM 14.

$$\lim_{t\to\infty} t^{-1}\sum_{s=0}^{t-1} x_s' Q x_s + u_s' R u_s = J(\theta^0) \quad a.s.$$

Proof. The proof that we provide is similar to [9, Thm 7.17] and [19], [20] of P. Mandl. Writing $K(\hat{\theta}_t) = K(\theta^0) + [K(\hat{\theta}_t) - K(\theta^0)]$ and using (7) and (9.i) gives the matrix identity

$$\begin{aligned}&Q + K'(\hat{\theta}_t)RK(\hat{\theta}_t) + [A(\theta^0) + B(\theta^0)K(\hat{\theta}_t)]'P(\theta^0)[A(\theta^0) + B(\theta^0)K(\hat{\theta}_t)] \\ &\quad = [K(\hat{\theta}_t) - K(\theta^0)]'[R + B'(\theta^0)P(\theta^0)B(\theta^0)][K(\hat{\theta}_t) - K(\theta^0)] + P(\theta^0).\end{aligned}$$

Define now the random variable

(36)

$$\begin{aligned}y_{t+1} := {} & x_t' Q x_t + u_t' R u_t - J(\theta^0) + x_{t+1}' P(\theta^0) x_{t+1} - x_t' P(\theta^0) x_t \\ & - x_t'[K(\hat{\theta}_t) - K(\theta^0)]'[R + B'(\theta^0)P(\theta^0)][K(\hat{\theta}_t) - K(\theta^0)]x_t.\end{aligned}$$

Using the above matrix identity and substituting $x_{t+1} = [A(\theta^0) + B(\theta^0)K(\hat{\theta}_t)]x_t + w_{t+1}$ and $u_t = K(\hat{\theta}_t)x_t$ gives

$$y_{t+1} = w'_{t+1}P(\theta^0)w_{t+1} - J(\theta^0) + 2w'_{t+1}P(\theta^0)[A(\theta^0) + B(\theta^0)K(\hat{\theta}_t)]x_t. \tag{37}$$

Equations (37) and (8) now yield

$$E[y_{t+1}|\mathcal{F}_t] = 0. \tag{38}$$

Additionally from (37) we also obtain

$$\begin{aligned} E[y_{t+1}^2|\mathcal{F}_t] &= a + x_t'[A(\theta^0) + B(\theta^0)K(\hat{\theta}_t)]P(\theta^0)b \\ &\quad + 4x_t'[A(\theta^0) + B(\theta^0)K(\hat{\theta}_t)]'P(\theta^0)P(\theta^0)[A(\theta^0) + B(\theta^0)K(\hat{\theta}_t)]x_t, \end{aligned}$$

where

$$\begin{aligned} a &:= E[\{w'_{t+1}P(\theta^0)w_{t+1} - J(\theta^0)\}^2], \\ b &:= 4E[w_{t+1}(w'_{t+1}P(\theta^0)w_{t+1} - J(\theta^0))]. \end{aligned}$$

Defining also

$$c := \max\{\|[A(\theta^0) + B(\theta^0)K(\theta)]P(\theta^0)\|: \theta \in \Theta\},$$

we have

$$E[y_{t+1}^2|\mathcal{F}_t] \leqq a + c\|b\|\|x_t\| + 4c^2\|x_t\|^2 \leqq (a + c\|b\|) + (4c^2 + c\|b\|)\|x_t\|^2.$$

From Lemma 13 it then follows that $\sum_{t=1}^{\infty} t^{-2}E[y_{t+1}^2|\mathcal{F}_t] < \infty$ a.s. This together with (38) yields from a version of the stability theorem [9, Thm 2.18],

$$\lim_{t\to\infty} t^{-1} \sum_{s=1}^{t} y_s = 0 \quad \text{a.s.} \tag{39}$$

We also have by the Schwarz inequality, Theorem 18 and Theorem 12 that

$$\begin{aligned} &\limsup_{t\to\infty} t^{-1} \sum_{s=0}^{t-1} x_s'[K(\hat{\theta}_s) - K(\theta^0)]'[R + B'(\theta^0)P(\theta^0)B(\theta^0)][K(\hat{\theta}_s) - K(\theta^0)]x_s \\ &\leqq \limsup_{t\to\infty} \|R + B'(\theta^0)P(\theta^0)B(\theta^0)\| t^{-1} \sum_{s=0}^{t-1} \|K(\hat{\theta}_s) - K(\theta^0)\|^2\|x_s\|^2 \\ &\leqq \limsup_{t\to\infty} \|R + B'(\theta^0)P(\theta^0)B(\theta^0)\| \\ &\quad \cdot \left[t^{-1} \sum_{s=0}^{t-1} \|K(\hat{\theta}_s) - K(\theta^0)\|^4\right]^{1/2} \left[t^{-1} \sum_{s=0}^{t-1} \|x_s\|^4\right]^{1/2} \\ &= 0 \quad \text{a.s.} \end{aligned} \tag{40}$$

From (39), (36). Lemma 13 and (40) we get

$$\begin{aligned} 0 &= \lim_{t\to\infty} t^{-1} \sum_{s=1}^{t} y_s \\ &= \lim_{t\to\infty} \left[t^{-1} \sum_{s=0}^{t-1} (x_s'Qx_s + u_s'Ru_s) - J(\theta^0) + t^{-1}(x_t'P(\theta^0)x_t - x_0'P(\theta^0)x_0)\right. \\ &\qquad \left. - t^{-1} \sum_{s=0}^{t-1} x_s'[K(\hat{\theta}_s) - K(\theta^0)]'[R + B'(\theta^0)P(\theta^0)B(\theta^0)][K(\hat{\theta}_s) - K(\hat{\theta}^0)]x_s\right] \\ &= \lim_{t\to\infty} t^{-1} \sum_{s=0}^{t-1} (x_s'Qx_s + u_s'Ru_s) - J(\theta^0). \end{aligned}$$

□

9. Concluding remarks. The situation we have studied in this paper is limited in its formulation. However, we have presented a *new* adaptive control law which is impervious to the closed-loop identification problem, which issue does not appear to have been fully resolved in the literature. Our results and analysis are fairly complete, and we prove important properties such as asymptotic closed-loop identification, convergence of the adaptive control law, closed-loop stability and overall system optimality with respect to the cost criterion involving a control cost.

Extensions of the problem considered here to overcome the limited formulation are desirable. Future attempts should certainly be directed at (i) allowing temporal noise correlation and (ii) allowing for the parameter set Θ to be compact.

Acknowledgments. The results of § 7 are from [18] and were obtained in collaboration with Tom Seidman, to whom the author is grateful.

REFERENCES

[1] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica, 9, (1973), pp. 185–199.
[2] L. LJUNG AND B. WITTENMARK, *Asymptotic properties of self-tuning regulators*, Report 7404, Div. of Automatic Control, Lund Inst. of Technology, 1974.
[3] L. LJUNG, *On positive real transfer functions and the convergence of some recursive schemes*, IEEE Trans. Autom. Control, AC-22, (1977), pp. 539–550.
[4] ———, *Analysis of recursive stochastic algorithms*, IEEE Trans. Autom. Control, AC-22, (1977), pp. 551–575.
[5] L. LJUNG AND B. WITTENMARK, *On a stabilizing property of adaptive regulators*, Report 7528 (C), Div. of Automatic Control, Lund Inst. of Technology, 1975.
[6] K. J. ÅSTRÖM, U. BORISSON, L. LJUNG AND B. WITTENMARK, *Theory and applications of self-tuning regulators*, Automatica, 13 (1976), pp. 457–476.
[7] G. C. GOODWIN, P. J. RAMADGE AND P. E. CAINES, *Discrete time stochastic adaptive control*, this Journal, 19 (1981), pp. 829–853.
[8] B. EGARDT, *Stability of adaptive controllers*, Lecture Notes in Control and Information Sciences, 20, Springer-Verlag, Berlin, 1979.
[9] P. HALL AND C. C. HEYDE, *Martingale Limit Theory and Its Applications*, Academic Press, New York, 1980.
[10] P. R. KUMAR AND A. BECKER, *A new family of optimal adaptive controllers for Markov chains*, IEEE Trans. Autom. Control, AC-27, (1982), pp. 137–146.
[11] P. R. KUMAR AND W. LIN, *Optimal adaptive controllers for unknown Markov chains*, IEEE Trans. Autom. Control, AC-27, (1982) to appear (August).
[12] P. R. KUMAR, *Simultaneous identification and adaptive control of unknown systems over finite parameter sets*, IEEE Trans. Autom. Control, AC-27 (1982) to appear (December).
[13] V. BORKAR AND P. VARAIYA, *Adaptive control of Markov chains*, I: *Finite parameter set*, IEEE Trans. Autom. Control, AC-24, (1979), pp. 953–958.
[14] D. BERTSEKAS, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.
[15] L. LJUNG, *Consistency of the least-squares identification method*, IEEE Trans. Autom. Control, AC-21, (1976), pp. 779–781.
[16] K. L. CHUNG, *A Course in Probability Theory*, Academic Press, New York, 1974.
[17] J. M. ORTEGA, *Numerical Analysis: A Second Course*, Academic Press, New York, 1972.
[18] P. R. KUMAR AND T. I. SEIDMAN, *Average stability of deterministic and random linear systems*, Univ. Maryland Baltimore County Mathematics Research Report, 1981.
[19] P. MANDL, *The use of optimal stationary policies in the adaptive control of linear systems*, Proc. Symposium to honour Jerzy Neyman, Warsaw, 1974, pp. 223–242.
[20] ———, *Some results in the adaptive control of linear systems*, Trans. Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European Meeting of Statisticians, Prague, 1977, pp. 399–410.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0002 $01.25/0

ON THE EXTENSION OF CONSTRAINED OPTIMIZATION ALGORITHMS FROM DIFFERENTIABLE TO NONDIFFERENTIABLE PROBLEMS*

E. POLAK†, D. Q. MAYNE‡ AND Y. WARDI†

Abstract. This paper presents three general schemes for extending differentiable optimization algorithms to nondifferentiable problems. It is shown that the Armijo gradient method, phase-I–phase-II methods of feasible directions and exact penalty function methods have conceptual analogs for problems with locally Lipschitz functions and implementable analogs for problems with semismooth functions. The exact penalty method has required the development of a new optimality condition.

Key words. Optimization algorithms, nondifferentiable optimization, locally Lipschitz functions, constrained optimization, phase I–phase II methods.

Introduction. Over the last several years, we have witnessed systematic efforts first to extend the concepts of the calculus to locally Lipschitz functions (see e.g. [C1], [L1], [L2]) and then to extend optimality conditions for differentiable optimization problems to optimization problems with locally Lipschitz functions (see e.g. [C2], [C3], [P11], [P15], [P13]). As a result, we now have an analogue of the extended F. John multiplier rule for nondifferentiable mathematical programming problems [C2], analogues of Lagrangians [C1] and an analogue of the maximum principle for nondifferentiable optimal control problems [C3].

The development of nondifferentiable optimization algorithms, for the nonconvex case, has been far less systematic. Two distinct approaches have emerged: that of the Kiev school, which constructs algorithms without a monotonic descent property [S1], [S2], [P12], and the one favored in the West, which always insists on monotonic descent of the cost or of a surrogate cost [B2], [G1], [P2], [P4], [P7]. In this paper we are concerned with algorithms of the second type. Although the literature on nondifferentiable optimization algorithms of the second type is still extremely small, two principles seem to have emerged. The first principle (see, e.g. [B2], [G1], [L2], [D2], [P7]) is that in extending a differentiable optimization algorithm to the nondifferentiable case, it is necessary to replace gradients, not with corresponding generalized gradients, but with bundles of generalized gradients in order to make up for the lack of continuity of the generalized gradients. The bundle-size parameter (ε) then has to be driven to zero as an optimal point is approached. The second principle was developed in [M1], [L2], [L5], [W1], [W2], [P2], [P7]. The gist of it is that when functions are semismooth, it is possible to get a good approximation to the nearest point from the origin to their generalized gradient bundles in a finite number of operations. The importance of this fact is that it defines an important class of nondifferentiable optimization problems for which one can obtain *implementable* algorithms, i.e., algorithms in which all the computations that are required to be performed in each iteration can be carried out in a finite number of simple operations.

* Received by the editors June 29, 1981, and in revised form February 10, 1982. This research was sponsored by the National Science Foundation under grants ECS-79-13148, ECS-79-13148/CEE-8105790 and the Joint Services Electronics Program under contract F49620-79-C-0178.

† Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720.

‡ Department of Computing and Control, Imperial College of Science and Technology, London SW7 2BT, England.

In this paper we develop three general schemes for the extension of differentiable optimization algorithms to nondifferentiable problems. The first one is for unconstrained optimization, while the remaining ones are for constrained optimization algorithms. To illustrate the applicability of these schemes, we use them to construct several *conceptual algorithms* for optimization problems with locally Lipschitz functions. These include an extension of the Armijo gradient method (previously presented in [P7]), extensions of two phase-I–phase-II methods of feasible directions of the type discussed in [P3] and extensions of exact penalty methods [C5], [P14]. The extension of exact penalty methods has required the development of a sharper optimality condition for constrained problems than the ones found in [C2]. Finally, for the semismooth case, we show that the conceptual algorithms give rise to *implementable algorithms* in a totally systematic manner. We hope that the results presented in this paper will contribute to the understanding and development of nondifferentiable optimization algorithms.

1. Preliminary results. Our analysis of algorithms for nonsmooth optimization will be based on a very small number of nonsmooth analysis results. For the sake of convenience, we begin by summarizing these; for details and proofs, the reader is referred to [C1], [C2], [L1].

DEFINITION 1.1 [C1]. Let $f:\mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. The *generalized gradient* of f at x is defined to be the set

$$\partial f(x) \triangleq \operatorname{co}\left\{\lim_{v_i \to 0} \nabla f(x+v_i)\right\} \tag{1.1}$$

where $\nabla f(x)$ denotes the gradient of $f(\cdot)$ at x, co denotes the convex hull of a set, and the v_i are such that $\nabla f(x+v_i)$ exists[1], and $\lim_{v_i \to 0} \nabla f(x+v_i)$ exists.

DEFINITION 1.2 [C1]. Let $f:\mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. The *generalized directional derivative* of f at x in the direction h is defined to be

$$d^0 f(x;h) \triangleq \overline{\lim_{\substack{y \to 0 \\ \lambda \searrow 0}}} \frac{f(x+y+\lambda h)-f(x+y)}{\lambda}. \tag{1.2}$$

The generalized directional derivative always exists and is well defined by (1.2) (see [C1]).

FACT 1.1 [C1]. *Let $f:\mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. Then*

a) *$\partial f(x)$ exists and is compact at all $x \in \mathbb{R}^n$.*

b) *$\partial f(x)$ is bounded on bounded sets.*

c) *$\partial f(\cdot)$ is upper semicontinuous in the sense that $\{x_i \to \hat{x},\ y_i \in \partial f(x_i)$ and $y_i \to \hat{y}\} \Rightarrow \{\hat{y} \in \partial f(\hat{x})\}$.*

d) *$d^0 f(x;v)$ exists for all $x, v \in \mathbb{R}^n$.*

e)

$$d^0 f(x;v) = \max_{\xi \in \partial f(x)} \langle \xi, v \rangle. \tag{1.3}$$

f) *Whenever the directional derivative $df(x;v)$ exists,*

$$df(x;v) \leqq d^0 f(x;v), \tag{1.4}$$

and furthermore, when f is C^1 at x, equality holds;

[1] Since $f(\cdot)$ is locally Lipschitz, $\nabla f(\cdot)$ exists almost everywhere.

g) *If x and h are such that $d^0 f(x+sh; h) \leqq -\rho < 0$ for all $s \in [0, 1]$, then*

$$f(x+sh)-f(x) \leqq -\alpha\rho s \quad \forall s \in [0, 1], \quad \forall \alpha \in (0, 1). \tag{1.5}$$

FACT 1.2 (mean value theorem) [L1]. *Let $f: \mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. Then, given $x, y \in \mathbb{R}^n$,*

$$f(y)-f(x) = \langle \xi, y-x \rangle \tag{1.6}$$

for some $\xi \in \partial f(x+s(y-x))$ and $s \in [0, 1]$.

FACT 1.3 [C2]. *Let $f: \mathbb{R}^n \to \mathbb{R}^1$, $g^i: \mathbb{R}^n \to \mathbb{R}^1$, $i \in \mathbf{m} \triangleq \{1, 2, \cdots, m\}$; $h_j: \mathbb{R}^n \to \mathbb{R}^1$, $j \in \mathbf{l} \triangleq \{1, 2, \cdots, l\}$ be locally Lipschitz continuous and let $\hat{x}$ be a solution of the problem*

$$\min \{f(x) | g^i(x) \leqq 0, i \in \mathbf{m}, h^j(x) = 0, j \in \mathbf{l}\}. \tag{1.7}$$

Then

$$0 \in \operatorname{co} \{\partial f(\hat{x}) \cup \{\partial g^i(\hat{x}) | i \in I(\hat{x})\} \cup \{t_j \partial h^j(\hat{x}) | j \in \mathbf{l}\}\}, \tag{1.8}$$

where $I(\hat{x}) \triangleq \{i \in \mathbf{m} | g^i(\hat{x}) = 0\}$ and $t_j \in \{+1, -1\}$.

The above result is not quite strong enough to be used in the context of exact penalty function methods, and hence we have had to propose the new optimality condition stated below.

THEOREM 1.1. *Let f, g^i, $i \in \mathbf{m}$; h^j, $j \in \mathbf{l}$, from $\mathbb{R}^n$ into $\mathbb{R}^1$ be locally Lipschitz continuous. Let $\hat{x}$ be a solution to (1.7) and let $F: \mathbb{R}^n \to \mathbb{R}^1$ be defined by*

$$F(x) \triangleq \max \{f(x)-f(\hat{x}); g^i(x)_+, i \in \mathbf{m}; |h^j(x)|, j \in \mathbf{l}\}, \tag{1.9}$$

where $g^i(x)_+ \triangleq \max \{g^i(x), 0\}$. Suppose that $\{x | F(x) = 0\}$ has measure zero. Then

$$0 \in \operatorname{co} \{\partial f(\hat{x}) \cup \{\partial g^i(\hat{x})_+ \cap \partial g^i(\hat{x}) | i \in I(\hat{x})\} \cup \{t_j \partial h^j(\hat{x}) | j \in \mathbf{l}\}\} \tag{1.10}$$

where $I(\hat{x}) = \{i \in \mathbf{m} | g^i(\hat{x}) = 0\}$ and $t_j \in \{+1, -1\}$.

Proof. Although F. Clarke has proved the above result for a somewhat more general case [C4], we shall only give a proof for the slightly restrictive case where $\hat{x}$ is also a local solution to $\min \{f(x) | g^i(x) \leqq 0, i \in \mathbf{m}; t_i h^j(x) \leqq 0, j \in \mathbf{l}\}$. (We note that (1.8) is also an optimality condition for this case.) Now, $\max \{g^i(x)_+, i \in \mathbf{m}; |h^j(x)|, j \in \mathbf{l}\} > 0$ for all x which are infeasible and $f(x)-f(\hat{x}) \geqq 0$ for all x which are feasible. Hence, $F(x) \geqq 0$ for all x. Hence $\hat{x} = \arg \min_{x \in \mathbb{R}^n} F(x)$, so that $0 \in \partial F(\hat{x})$. Now, by assumption, $\{x | F(x) = 0\}$ has measure zero, and hence (1.10) follows directly from the fact that $\partial F(\hat{x})$ involves the limit of gradients $\nabla g^i(x)$ evaluated only at points x where $g^i(x) > 0$. □

2. Unconstrained optimization. Let $f: \mathbb{R}^n \to \mathbb{R}^1$ be locally Lipschitz continuous. Consider the problem

$$\min_{x \in \mathbb{R}^n} f(x). \tag{2.1}$$

We shall consider algorithms for solving (2.1) of the form

$$x_{i+1} = x_i + \lambda_i h_i, \tag{2.2a}$$

$$\lambda_i = \arg \max_{k \in \mathbb{N}^+} \{\beta^k | f(x_i + \beta^k h_i) - f(x_i) \leqq \alpha \beta^k \delta_i\}, \tag{2.2b}$$

where $\alpha, \beta \in (0, 1)$, $\mathbb{N}^+ = \{1, 2, 3, \cdots\}$, and $\delta_i < 0$. We recognize these algorithms as a generalization of the class of descent algorithms, utilizing the Armijo step size rule

[A1], that were discussed by Polak, Sargent and Sebastian in [P9] for the differentiable case. Although most, if not all, differentiable unconstrained optimization algorithms, of the form considered by Polak, Sargent and Sebastian, can be analyzed in terms of the convergence theorem [P8, Thm. (1.3.10)], their structure permits the introduction of more readily verifiable assumptions than those found in [P8, Thm. (1.3.10)]. Consequently in [P9] we find (in slightly different form) the following result, which is intended to be used for algorithms of the form (2.2a, b) when $\delta_i = df(x_i; h_i)$.

THEOREM 2.1. *Suppose that* $f: \mathbb{R}^n \to \mathbb{R}^1$ *is* C^1 *and that there exist two continuous functions* $N_1, N_2: \mathbb{R}^n \to \mathbb{R}^+$, *which vanish only at points* x *for which* $\nabla f(x) = 0$, *such that for* h_i *in* (2.2a)

$$df(x_i; h_i) = \langle \nabla f(x_i), h_i \rangle \leqq -N_1(x_i), \tag{2.3}$$

$$\|h_i\| \leqq N_2(x_i) \tag{2.4}$$

hold.

Then, given an $\bar{x}$ *such that* $\nabla f(\bar{x}) \neq 0$, *there exist a* $\bar{\rho} > 0$ *and a* $\bar{k} \in \mathbb{N}_+$ *such that for all* $x_i \in B(\bar{x}, \bar{\rho}) \triangleq \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| \leqq \bar{\rho}\}$,

$$f(x_i + \lambda h_i) - f(x_i) \leqq \lambda \alpha df(x_i; h_i) \leqq \frac{-\lambda \alpha N_1(\bar{x})}{2} \quad \forall \lambda \in [0, \beta^{\bar{k}}]. \tag{2.5}$$

Relation (2.5) leads to two conclusions: for all $x_i \in B(\bar{x}, \bar{\rho})$

(i) $\lambda_i \geqq \beta^{\bar{k}}$, and

(ii) $f(x_{i+1}) - f(x_i) \leqq -\beta^{\bar{k}} \alpha N_1(\bar{x})/2$,

i.e., the algorithm map defined by (2.2a, b), with $\delta_i \triangleq df(x_i; h_i)$, is *locally uniformly monotonic* (see [T1]). As an immediate consequence, we see from [P8, Thm. (1.3.9)] that any accumulation point $\hat{x}$ of $\{x_i\}$ satisfies $\nabla f(\hat{x}) = 0$.

Assumption 2.1. From now on, we shall assume that the function $f: \mathbb{R}^n \to \mathbb{R}^1$ is locally Lipschitz continuous.

Any attempt to extend Theorem 2.1 to the case of $f(\cdot)$ locally Lipschitz only, by replacing df with $d^0 f$ in (2.3), is doomed to failure, as can be seen from the counter example in [W2]. This is due to the fact that although an h_i satisfying $d^0 f(x_i; h_i) \leqq -N_1(x_i)$ and (2.4) is obviously a descent direction, it is not possible to ensure that the step size λ_i is bounded from below in a ball about an $\bar{x}$ such that $0 \notin \partial f(\bar{x})$. To insure that a nonsmooth optimization algorithm is locally uniformly monotonic, it becomes necessary to "look ahead" for the "corners" of $f(\cdot)$ by "smearing" $\partial f(x)$, as follows.

DEFINITION 2.1. For any $\varepsilon > 0$, we define the *ε-smeared generalized gradient* by

$$\partial_\varepsilon f(x) \triangleq \operatorname{co}\left\{ \bigcup_{x' \in B(x, \varepsilon)} \partial f(x') \right\}. \tag{2.6}$$

FACT 2.1. *For any* $\varepsilon > 0$, $\partial_\varepsilon f(x)$ *is compact, bounded on bounded sets; furthermore* $\partial_\varepsilon f(\cdot)$ *is upper semicontinuous (u.s.c.) (see* [P7]).

DEFINITION 2.2. For any $\varepsilon > 0$, we define the *ε-smeared generalized directional derivative* of $f(\cdot)$ at x in the direction h by

$$d^0_\varepsilon f(x; h) \triangleq \max_{\xi \in \partial_\varepsilon f(x)} \langle \xi, h \rangle. \tag{2.7}$$

With the introduction of $d^0_\varepsilon f(\cdot; \cdot)$, and ignoring for the moment the problem of choosing $\varepsilon > 0$ as well as that of computing $\partial_\varepsilon f(x)$ and $d^0_\varepsilon(x; h)$, we are ready to extend

Theorem 2.1 to the nonsmooth case. We shall refer to algorithms which assume that $\partial_\varepsilon f(x)$ and $d^0_\varepsilon(x; h)$ can be computed exactly, as *conceptual.*

In anticipation of the application of the new theorem to conceptual optimization algorithms for nonsmooth problems, we find it necessary to relax the continuity of N_1, N_2 in Theorem 2.1 to a requirement which is somewhat weaker than semicontinuity, as we shall now see.

THEOREM 2.2 (conceptual algorithms). *Let $\varepsilon > 0$ be given. Suppose that there exist two functions $N_1, N_2 : \mathbb{R}^n \to \mathbb{R}^+$ such that:*

(i) *if $N_1(x)N_2(x) = 0$, then $0 \in \partial_\varepsilon f(x)$*

(ii) *for every $x \in \mathbb{R}^n$ such that $0 \notin \partial_\varepsilon f(x)$, there exist a $\rho(x) > 0$ and $b_i(x) > 0$, $i = 1, 2$, such that for all $x' \in B(x, \rho(x))$,*

$$N_1(x') \geqq b_1(x), \tag{2.8a}$$

$$N_2(x') \leqq b_2(x). \tag{2.8b}$$

Now consider the process (2.2a, b) *and suppose that for $i = 0, 1, 2, \cdots$,*

$$d^0_\varepsilon f(x_i; h_i) \leqq -N_1(x_i), \tag{2.8c}$$

$$\|h_i\| \leqq N_2(x_i). \tag{2.8d}$$

Then, given any $\bar{x}$ such that $0 \notin \partial_\varepsilon f(\bar{x})$, there exists a $\bar{k} \in \mathbb{N}^+$ such that for all $x_i \in B(\bar{x}, \rho(\bar{x}))$ for all $\lambda \in [0, \beta^{\bar{k}}]$

$$f(x_i + \lambda h_i) - f(x_i) \leqq \lambda\alpha d^0_\varepsilon f(x_i; h_i) \leqq -\lambda\alpha b_1(\bar{x}). \tag{2.9}$$

Proof. Let $\bar{x} \in \mathbb{R}^n$ be such that $0 \notin \partial_\varepsilon f(\bar{x})$. Let $\bar{k} \in \mathbb{N}^+$ be such that $\beta^{\bar{k}} b_2(\bar{x}) \leqq \varepsilon$. Then, for all $x_i \in B(\bar{x}, \rho(\bar{x}))$ and for all $\lambda \in [0, \beta^{\bar{k}}]$, $(x_i + \lambda h_i) \in B(x_i, \varepsilon)$, and hence for all such x_i and λ,

$$\begin{aligned} d^0 f(x_i + \lambda h_i; h_i) &= \max_{\xi \in \partial f(x_i + \lambda h_i)} \langle \xi, h_i \rangle \leqq \max_{\xi \in \partial_\varepsilon f(x_i)} \langle \xi, h_i \rangle \\ &= d^0_\varepsilon f(x_i; h_i) \leqq -N_1(x_i) \leqq -b_1(\bar{x}). \end{aligned} \tag{2.10}$$

The desired result now follows from Fact 1.1(g). □

COROLLARY 2.1 (conceptual algorithms). *Let $\varepsilon > 0$ be given and suppose that the assumptions in Theorem 2.2 hold. Then any accumulation point $\hat{x}$ of a sequence $\{x_i\}_{i=0}^\infty$ constructed by an algorithm of the form* (2.2a, b) with $d^0_\varepsilon f(x_i; h_i) \leqq \delta_i \leqq -N_1(x_i)$ *satisfies* $0 \in \partial_\varepsilon f(\hat{x})$.

Proof. Suppose that $x_i \to_K \hat{x}$, with $K \subset \{0, 1, 2, \cdots\}$ infinite, and that $0 \notin \partial f(\hat{x})$. Then, by Theorem 2.2, there exists an i_0 and a $\hat{k} \in \mathbb{N}^+$ such that for all $i \geqq i_0$ and $i \in K$, $\lambda_i \geqq \beta^{\hat{k}}$ and

$$f(x_{i+1}) - f(x_i) \leqq \lambda_i\alpha d^0_\varepsilon f(x_i; h_i) \leqq \lambda_i\alpha\delta_i \leqq -\beta^{\hat{k}}\alpha b_1(\hat{x}). \tag{2.11}$$

Now, $\{f(x_i)\}$ is monotonically decreasing and $x_i \to_K \hat{x}$; hence, since $f(\cdot)$ is continuous, $f(x_i) \to f(\hat{x})$. But this contradicts (2.11) and hence we are done. □

The simplest algorithm in the class considered in Theorem 2.2 can be viewed as an "ε-smeared" steepest descent method. It sets

$$h_i = h_\varepsilon(x_i) \triangleq -\mathrm{Nr}\,(\partial_\varepsilon f(x_i)) \triangleq -\arg\min\{\|h\| \mid h \in \partial_\varepsilon f(x_i)\} \tag{2.12a}$$

and

$$\delta_i = -\|h_i\|^2. \tag{2.12b}$$

Hence

$$d_\varepsilon^0 f(x_i; h_i) = -\|h_i\|^2. \tag{2.13}$$

Setting $N_1(x_i) = \|h_i\|^2$, we see that $N_1(\cdot)$ is lower semicontinuous (l.s.c.) because $\partial_\varepsilon f(\cdot)$ is u.s.c. (see proof in [P7]). Next, if we define $N_2(x)$ by

$$N_2(x) = \max \{\|h\| \,|\, h \in \partial_\varepsilon f(x)\}, \tag{2.14}$$

we see that $\|h_i\| \leqq N_2(x_i)$ and that $N_2(\cdot)$ is u.s.c. because $\partial_\varepsilon f(\cdot)$ is u.s.c. (see proof in [P7]). Hence we can set $b_1(x) = N_1(x)/2$ and $b_2(x) = 2N_2(x)$ to show that this algorithm satisfies the assumptions of Theorem 2.2.

Obviously, we would prefer to have algorithms which generate accumulation points $\hat{x}$ such that $0 \in \partial f(\hat{x})$ rather than $0 \in \partial_\varepsilon f(\hat{x})$, with $\varepsilon > 0$. Hence, it is necessary to propose at least one ε-reduction scheme. The most natural thing to do is to reduce ε as x_i approaches a stationary point. This fact is not postulated in the theorem below, but unless it holds it is not possible to find a function $N_1(\cdot)$.

THEOREM 2.3 (conceptual algorithms). *Suppose that there exist three functions* $N_1, N_2, N_3 : \mathbb{R}^n \to \mathbb{R}^+$ *such that*:

(i) *if* $N_1(x)N_2(x)N_3(x) = 0$, *then* $0 \in \partial f(x)$,

(ii) *for every* $x \in \mathbb{R}^n$ *such that* $0 \notin \partial f(x)$, *there exist a* $\rho(x) > 0$ *and* $b_i(x) > 0$, $i = 1, 2, 3$, *such that for all* $x' \in B(x, \rho(x))$

$$N_1(x') \leqq b_1(x), \tag{2.15a}$$

$$N_2(x') \leq b_2(x), \tag{2.15b}$$

$$N_3(x') \geqq b_3(x). \tag{2.15c}$$

Now consider the process (2.2a, b) *and suppose that for* $i = 0, 1, 2, \cdots$,

$$d^0_{N_3(x_i)} f(x_i; h_i) \leqq -N_1(x_i), \tag{2.15d}$$

$$\|h_i\| \leqq N_2(x_i). \tag{2.15e}$$

Then, given any $\bar{x}$ *such that* $0 \notin \partial f(\bar{x})$, *there exists a* $\bar{k} \in \mathbb{N}^+$ *such that for all* $x_i \in B(\bar{x}, \rho(\bar{x}))$, *for all* $\lambda \in [0, \beta^{\bar{k}}]$,

$$f(x_i + \lambda h_i) - f(x_i) \leqq \lambda\alpha d^0_{N_3(x_i)} f(x_i; h_i) \leqq -\lambda\alpha b_1(\bar{x}). \tag{2.16}$$

Furthermore, any accumulation point $\hat{x}$ *of a sequence* $\{x_i\}_{i=0}^\infty$ *constructed by an algorithm of the form* (2.2a, b) *with* $\delta_i = d^0_{N_3}(x_i) f(x_i; h_i)$ *satisfies* $0 \in \partial f(\hat{x})$.

We omit a proof of this theorem since it is obtained by a trivial modification of the proofs of Theorem 2.2 and Corollary 2.1.

We shall now exhibit a natural candidate for $N_3(x)$ in extending the "ε-smeared" steepest descent method to one with an adjustable ε.

Thus, let $\nu \in (0, 1)$, $\varepsilon_0 > 0$, $\delta > 0$ be given. Let

$$\mathscr{E} \triangleq \{\varepsilon \,|\, \varepsilon = \varepsilon_0 \nu^k, k \in \mathbb{N}^+\} \cup \{0\}. \tag{2.17}$$

Next, for any $\varepsilon \geqq 0$, let as in (2.12),

$$h_\varepsilon(x) \triangleq -\mathrm{Nr}\,(\partial_\varepsilon f(x)) \triangleq -\arg\min \{\|h\|^2 \,|\, \in \partial_\varepsilon f(x)\}. \tag{2.18}$$

Then we define $\varepsilon : \mathbb{R}^n \to \mathbb{R}^1$ by

$$\varepsilon(x) \triangleq \max \{\varepsilon \in \mathscr{E} \,|\, \|h_\varepsilon(x)\|^2 \geqq \delta\varepsilon\}. \tag{2.19}$$

PROPOSITION 2.1. a) *For every* $\bar{x} \in \mathbb{R}^n$ *such that* $0 \notin \partial f(\bar{x})$, *there exists a* $\rho_3(\bar{x})$ *such that*

$$\varepsilon(x_i) \geqq \nu\varepsilon(\bar{x}) > 0 \quad \forall x_i \in B(\bar{x}, \rho_3(\bar{x})). \tag{2.20}$$

b) *If* $x_j \to \hat{x}$ *as* $j \to \infty$ *with* $0 \in \partial f(\hat{x})$, *then* $\varepsilon(x_j) \to \varepsilon(\hat{x}) = 0$ *as* $j \to \infty$.

Proof. a) Let $\bar{x}$ be such that $0 \notin \partial f(\bar{x})$. Then, since $\partial f(\cdot)$ is u.s.c., there exists an $\varepsilon_1 > 0$ such that $\|h_{\varepsilon_1}(\bar{x})\|^2 \geqq \frac{1}{2}\|h_0(\bar{x})\|^2 > 0$. Hence, since $\varepsilon' < \varepsilon''$ implies that $\|h_{\varepsilon'}(\bar{x})\|^2 \geqq \|h_{\varepsilon''}(\bar{x})\|^2$, it follows that

$$\varepsilon(\bar{x}) \geqq \max\left\{\varepsilon \in \mathscr{E} \,|\, \varepsilon \leqq \min\left\{\varepsilon_1, \frac{1}{2\delta}\|h_0(\bar{x})\|^2\right\}\right\} > 0. \tag{2.21}$$

Next, since by the maximum theorem in [B1], $\|h_{\varepsilon(\bar{x})}(\cdot)\|^2$ is l.s.c., and $\|h_{\varepsilon(\bar{x})}(\bar{x})\|^2 \geqq \delta\varepsilon(\bar{x})$, there exists a $\rho_3(\bar{x}) > 0$ such that

$$\|h_{\nu\varepsilon(\bar{x})}(x_i)\|^2 \geqq \|h_{\varepsilon(\bar{x})}(x_i)\|^2 \geqq \delta\nu\varepsilon(\bar{x}) \quad \text{for all } x_i \in B(\bar{x}, \rho_3(\bar{x})), \tag{2.22}$$

and hence (2.20) follows directly.

b) Suppose that $0 \in \partial f(\hat{x})$. Then $\|h_0(\hat{x})\|^2 = 0$ and for any $\varepsilon > 0$ $\|h_\varepsilon(\hat{x})\|^2 = 0$. Hence $\varepsilon(\hat{x}) = 0$. Next, suppose that $x_j \to \hat{x}$ as $j \to \infty$ and that $\lim \varepsilon(x_j) \geqq \hat{\varepsilon} > 0$. Since we must have $\hat{x} \in B(x_j, \hat{\varepsilon})$ for all j sufficiently large, we must have that $0 \in \partial_{\hat{\varepsilon}} f(x_j)$, for all j sufficiently large and hence $\|h_{\varepsilon(x_j)}(x_j)\|^2 = 0 < \varepsilon(x_j)$ for all $j \in K$ sufficiently large. But this contradicts the definition of $\varepsilon(x_j)$, and hence we are done. □

The final version of the progressively smeared steepest descent method is sufficiently important to be stated formally:

ALGORITHM 2.1 (conceptual).
Parameters: $\alpha, \beta, \nu \in (0, 1)$, $\varepsilon_0 > 0$, $\delta > 0$.
Data: $x_0 \in \mathbb{R}^n$.
Step 1: Set $i = 0$.
Step 2: Compute $h_i \triangleq h_{\varepsilon(x_i)}(x_i)$.
Step 3: Compute

$$\lambda_i = \arg\max\{\beta^k \,|\, f(x_i + \beta^k h_i) - f(x_i) \leqq \alpha\beta^k d^0_{\varepsilon(x_i)}(x_i; h_i)\}. \tag{2.23}$$

Step 4: Set $x_{i+1} = x_i + \lambda_i h_i$, set $i = i + 1$ and go to step 2.

THEOREM 2.4. *Suppose that* $\{x_i\}_{i=0}^{\infty}$ *is a sequence constructed by Algorithm* 2.1. *Then any accumulation point* $\hat{x}$ *of* $\{x_i\}$ (*if it exists*) *satisfies* $0 \in \partial f(\hat{x})$.

Proof. We only need to show that the assumptions of Theorem 2.3 are satisfied. Clearly, we must set $N_3(x) = \varepsilon(x)$, and by Proposition 2.1, it has the required properties. Next, we set $N_1(x) \triangleq \|h_{\varepsilon(x)}(x)\|^2$. Then the required properties of $N_1(\cdot)$ follow from those of $\varepsilon(\cdot)$ (with $\rho_1(x) = \rho_3(x)$) and, by inspection,

$$d^0_{\varepsilon(x)} f(x; h_{\varepsilon(x)}(x)) = N_1(x).$$

Finally, we set $N_2(x) \triangleq \arg\max\{\|h\| \,|\, h \in \partial_{\varepsilon_0} f(x)\}$. Since $N_2(\cdot)$ is u.s.c. by the maximum theorem in [B1], we are done. □

Next we turn to implementable algorithms. These are characterized by the fact that they approximate the sets $\partial_\varepsilon f(x)$ by means of finite operations while retaining a great resemblance to the conceptual algorithms from which they are derived. It does not appear to be possible to construct a truly useful general convergence theorem of

the form of Theorem 2.3 for such algorithms. Instead, it seems simplest to use a minor modification of [P8, Thm. (1.3.10)], as follows.

THEOREM 2.5. *Consider algorithms of the form* (2.2a, b). *If for every* $\bar{x} \in \mathbb{R}^n$ *such that* $0 \notin \partial f(\bar{x})$ *there exist a* $\bar{k} \in \mathbb{N}^+$, *a* $\bar{\delta} > 0$ *and a* $\bar{\rho} > 0$ *such that for all* $x_i \in B(\bar{x}, \bar{\rho})$,

$$f(x_i + \beta^{\bar{k}} h_i) - f(x_i) \leqq -\alpha\beta^{\bar{k}}\delta_i \leqq -\alpha\beta^{\bar{k}}\bar{\delta}, \tag{2.24}$$

then any accumulation point $\hat{x}$ *of a sequence* $\{x_i\}_{i=0}^{\infty}$ *constructed by such an algorithm satisfies* $0 \in \partial f(\hat{x})$.

Proof. Suppose $x_i \to^K \hat{x}$ and $0 \notin \partial f(\hat{x})$. Then there exists an i_0 such that for all $i \in K$, $i \geqq i_0$, $\lambda_i \geqq \beta^{\bar{k}}$, and hence

$$f(x_{i+1}) - f(x_i) \leqq -\alpha\beta^{\bar{k}}\bar{\delta} \quad \forall i \geqq i_0, \quad i \in K. \tag{2.25}$$

But $\{f(x_i)\}$ is monotonically decreasing and $f(\cdot)$ is continuous; hence $f(x_i) \to f(\hat{x})$ as $i \to \infty$. But, clearly, this contradicts (2.25), and we are done. □

At the present time, we only know how to construct implementable algorithms for optimization problems in which the function $f(\cdot)$ is semismooth (see [M1]).

DEFINITION 2.3 [M1]. A locally Lipschitz continuous function $f(\cdot)$ is said to be *semismooth* if it is directionally differentiable and if for any x, $h \in \mathbb{R}^n$, and for any sequences $\{\lambda_k\} \subset \mathbb{R}^+$, $\{z_k\}$, $\{v_k\} \subset \mathbb{R}^n$, such that $\lambda_k \to 0$, $(1/\lambda_k)v_k \to 0$ and $z_k \in \partial f(x + \lambda_k h + v_k)$, the sequence $\{\langle z_k, h\rangle\}$ converges to $df(x; h)$.

From our point of view, the most important property of semismooth functions, which does not appear in the definition, is the following one:

PROPOSITION 2.2. *Suppose that* $f: \mathbb{R}^n \to \mathbb{R}^1$ *is semismooth. then, given any* x, h, $\{\lambda_k\}$, $\{v_k\}$ *as in Definition* 2.3,

$$\lim_{k \to \infty} df(x + \lambda_k h + v_k; h) = df(x; h). \tag{2.26}$$

We assume, until the end of this section, that $f(\cdot)$ is semismooth. We are now ready to construct an implementation for Algorithm 2.1, which satisfies the assumptions of Theorem 2.5. The implementation is based on the following observations derived from results of Lemarechal [L2] and Wolfe [W1], [W2]. Suppose that $x_i \in \mathbb{R}^n$, $\varepsilon > 0$ are given and that $0 \notin \partial_\varepsilon f(x_i)$. Let $Y_s \subset \partial_\varepsilon f(x_i)$ be the convex hull of a finite number of points in $\partial_\varepsilon f(x_i)$ and let

$$\eta_s = -\mathrm{Nr}\,(Y_s). \tag{2.27}$$

Now, let $k_s \in \mathbb{N}^+$ be such that

$$\beta\varepsilon < \beta^{k_s}\|\eta_s\| \leqq \varepsilon. \tag{2.28}$$

Then, either

$$f(x_i + \beta^{k_s}\eta_s) - f(x_i) \leqq -\alpha\beta^{k_s}\|\eta_s\|^2 \leqq -\alpha\beta^{k_s}\|\mathrm{Nr}\,(\partial_\varepsilon f(x_i))\|^2 \tag{2.29}$$

holds or not. If (2.29) does hold, then $h_i = \eta_s$ turns out to be an adequate approximation to $h_\varepsilon(x_i)$, as far as convergence is concerned. If (2.29) does not hold (see Fig. 1), then there must be a point $\bar{\mu} \in [0, \beta^{k_s}]$ such that

$$f(x_i + \bar{\mu}\eta_s) - f(x_i) = -\bar{\mu}\alpha\|\eta_s\|^2 \tag{2.30}$$

and

$$df(x_i + \bar{\mu}\eta_s) \geqq -\alpha\|\eta_s\|^2. \tag{2.31}$$

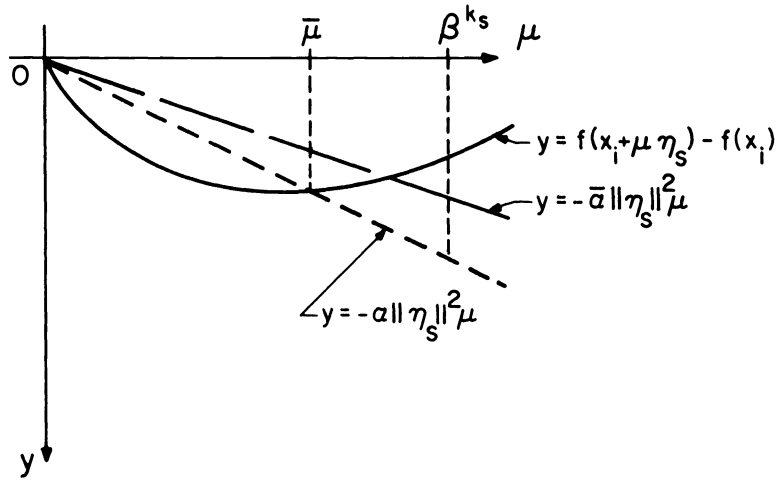


FIG. 1

Now suppose that $\mu_j \in [0, \beta^{k_s}]$, $j = 1, 2, \cdots$, are such that $\mu_j \searrow \bar{\mu}$ and that $y_j \in \partial f(x_i + \mu_j \eta_s)$, for $j = 1, 2, \cdots$. Then, because $f(\cdot)$ is semismooth,

$$\langle y_j, \eta_s \rangle \to df(x_i + \bar{\mu}\eta_s) \quad \text{as } j \to \infty, \tag{2.32}$$

and, consequently, given an $\bar{\alpha} \in (\alpha, 1)$, there exists a j_0 such that

$$\langle y_j, \eta_s \rangle \geqq -\bar{\alpha}\|\eta_s\|^2 \quad \forall j \geqq j_0. \tag{2.33}$$

Referring to Fig. 2, we see that if we set $Y_{s+1} = \text{co}\,(Y_s \cup \{y_j\})$, $\eta_{s+1} = -\text{Nr}\,(Y_{s+1})$ is smaller than η_s in norm. We can now replace η_s by η_{s+1} and return to the test in (2.28), and so forth. This cycle of operations cannot continue indefinitely, because, as shown in [P7], if $s \to \infty$, then $\eta_s \to 0$, which contradicts the obvious fact that $\|n_s\| \geqq \|h_s(x_i)\| > 0$. Hence the test (2.29) will be passed in a finite number of operations. Note also that β^{k_s} is locally (w.r.t. x) bounded both from below and from above.

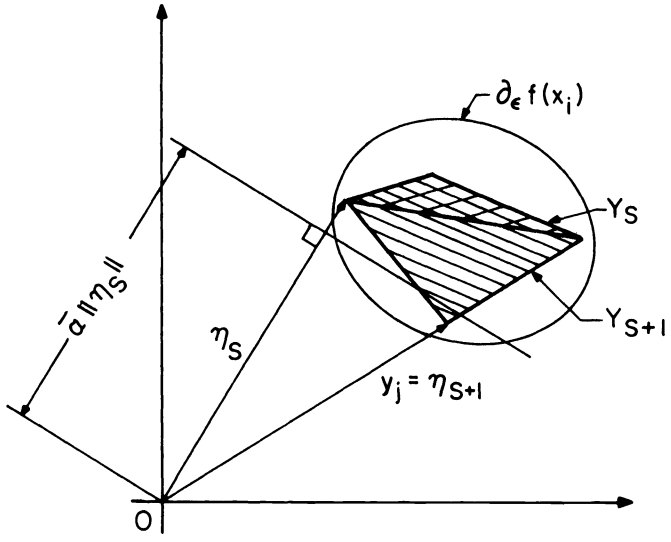


FIG. 2

Hence the convergence of the algorithm below is very easily deduced from the preceding results. Note that the algorithm below uses a bisection procedure for finding $\bar{\mu}$ and for constructing μ_j.

ALGORITHM 2.2.
Parameters: $\varepsilon_0 > 0$, $\alpha, \beta, \nu \in (0, 1)$, $\bar{\alpha} \in (\alpha, 1)$.
Data: $x_0 \in \mathbb{R}^n$.
Step 0: Set $i = 0$.
Step 1: Set $\varepsilon = \varepsilon_0$, $s = 0$.
Step 2: Compute $Y_s \subset \partial_\varepsilon f(x_i)$, a convex hull of a finite number of points in $\partial_\varepsilon f(x_i)$.
Step 3: Compute $\eta_s = -\mathrm{Nr}\,(Y_s)$ and $k_s \in \mathbb{N}^+$ such that $\beta\varepsilon \leqq \beta^{k_s}\|\eta_s\| \leqq \varepsilon$.
Step 4: If $\|\eta_s\| < \varepsilon$, set $\varepsilon = \nu\varepsilon$ and go to step 2.
Step 5: If

$$f(x_i + \beta^{k_s}\eta_s) - f(x_i) \leqq -\alpha\beta^{k_s}\|\eta_s\|^2, \tag{2.34a}$$

(i) set $h_i = \eta_s$ and compute the smallest $k_i \in \mathbb{N}^+$ such that

$$f(x_i + \beta^{k_i}h_i) - f(x_i) \leqq -\alpha\beta^{k_i}\|h_i\|^2; \tag{2.34b}$$

(ii) set $x_{i+1} = x_i + \beta^{k_i}h_i$;
(iii) set $i = i + 1$;
(iv) go to step 1.
Step 6: Set $j = 0$.
Step 7: Set $l_0 = 0$, $r_0 = \beta^{k_s}\|\eta_s\|^2$, $\mu_0 = r_0/2$.
Step 8: Compute a $y_{j+1} \in \partial_\varepsilon f(x_i + r_j\eta_s)$.
Step 9: If

$$\langle y_{j+1}, \mu_s\rangle \geqq -\bar{\alpha}\|\eta_s\|^2, \tag{2.35}$$

set

$$Y_{s+1} = \mathrm{co}\,(\{y_{j+1}\} \cup Y_s), \tag{2.36}$$

set $s = s + 1$ and go to step 3.
Step 10: If

$$f(x_i + \mu_j\eta_s) - f(x_i) \geqq -\alpha\mu_j\|\eta_s\|^2, \tag{2.37}$$

set $r_{j+1} = \mu_j$, $l_{j+1} = l_j$, $\mu_{j+1} = (r_{j+1} + l_{j+1})/2$.
Else set $r_{j+1} = r_j$, $l_{j+1} = \mu_j$, $\mu_{j+1} = (r_{j+1} + l_{j+1})/2$.
Step 11: Set $j = j + 1$ and go to step 8.

The success of Algorithm 2.2 depends on the following fact, due to Wolfe [W1], [W2] (see also [P7]).

PROPOSITION 2.3. *Let S' be a compact, convex subset of a compact convex set S and let $\bar{\alpha} \in (0, 1)$. Let $h' = \mathrm{Nr}(S')$ and let $g \in S$ be such that*

$$\langle g, h'\rangle \leqq \bar{\alpha}\|h'\|^2. \tag{2.38a}$$

Then $h'' = \mathrm{Nr}\,(\mathrm{co}\,\{g, S'\})$ satisfies

$$\|h''\|^2 \leqq \max\{\bar{\alpha}, 1 - (1-\bar{\alpha})^2\|h'\|^2/4C^2\}\|h'\|^2 \tag{2.38b}$$

where $C \geqq \max\{\|g\| \mid g \in S\}$.

THEOREM 2.6. a) *If Algorithm 2.2 generates a finite sequence $\{x_i\}_{i=0}^N$, jamming at x_N, i.e., with construction stopping and the algorithm cycling in the loop defined by steps 2–4, or steps 3–9 or steps 8–11, then $0 \in \partial f(x_N)$.*

b) *If Algorithm* 2.2 *generates an infinite sequence* $\{x_i\}_{i=0}^{\infty}$, *then every accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies* $0 \in \partial f(\hat{x})$.

Proof. a) Suppose that the sequence $\{x_i\}$ is finite with the algorithm jamming up at x_N, cycling indefinitely in one of the loops defined by steps 2 to 4 or steps 3 to 9 or steps 8 to 11. Suppose that $0 \notin \partial f(x_N)$.

(i) Consider the loop defined by steps 2 to 4. Since $0 \notin \partial f(x_N)$, $\varepsilon(x_N) > 0$ (see (2.19)) and hence for all $\varepsilon \leqq \varepsilon(x_N)$, $Y_s \subset \partial_\varepsilon f(x_N)$, $\|\mathrm{Nr}\,(Y_s)\| \geqq \|\mathrm{Nr}\,(\partial_\varepsilon f(x_N))\| \geqq \|\mathrm{Nr}\,(\partial_{\varepsilon(x_N)} f(x_N))\| \geqq \varepsilon(x_N) \geqq \varepsilon$, and hence no infinite cycling can occur in this loop.

(ii) Consider the loop defined by steps 8 to 11. This loop is always finite because $f(\cdot)$ is semismooth and (2.34a) is not satisfied.

(iii) Consider the loop defined by steps 3–9. Since $0 \notin \partial f(x_N)$, $\varepsilon \geqq \varepsilon(x_N)$ while in this loop. Hence by Proposition 2.3,

$$\|\eta_{s+1}\| \leqq \max\{\bar{\alpha}, 1-(1-\bar{\alpha})\|\eta_s\|^2/4C^2\}\|\eta_s\|^2 \tag{2.39}$$

where $C = \max\{\|\eta\| \,|\, \eta \in \partial_{\varepsilon_0} f(x_N)\}$. Since $\|\eta_s\| \geqq \varepsilon \geqq \varepsilon(x_N)$ for all s, it is clear from (2.39) that the sequence $\{\eta_s\}$ must be finite, i.e., the loop defined by steps 3–9 is exited after a finite number of operations. Consequently, the algorithm jams up at x_N only if $0 \in \partial f(x_N)$.

b) Now suppose that the sequence $\{x_i\}$ is infinite. Suppose that $x_i \to^K \hat{x}$, with $K \subset \{0, 1, 2, \cdots\}$ infinite and that $0 \notin \partial f(\hat{x})$. Then, by Proposition 2.1, there exists an i_0 such that for all $i \in K$, $i \geqq i_0$, $\varepsilon(x_i) \geqq \nu\varepsilon(\hat{x}) > 0$. Consequently, for all $i \notin K$, $i \geqq i_0$ (2.34a) is satisfied with $\|\eta_s\| \geqq \nu\varepsilon(\hat{x})$ and $\beta_s^k\|\eta_s\| \geqq \beta\varepsilon(x_i) \geqq \beta\nu\varepsilon(\hat{x})$. Hence, by (2.34b), for all $i \in K$, $i \geqq i_0$,

$$f(x_{i+1}) - f(x_i) \leqq -\alpha\beta^{k_i}\|h_i\|^2 \leqq -\alpha\beta(\nu\varepsilon(\hat{x}))^2. \tag{2.40}$$

Now $f(x_i) \to^K f(\hat{x})$ by continuity and $\{f(x_i)\}$ is monotonic decreasing. hence, we must have $f(x_i) \to f(\hat{x})$, which contradicts (2.40). This completes our proof. □

3. Constrained optimization: Conceptual algorithms. We begin by examining the easiest case, viz., problems of the form

$$\min\{f(x) \,|\, g^j(x) \leqq 0, j \in \mathbf{m}\}, \tag{3.1}$$

where f, $g^j : \mathbb{R}^n \to \mathbb{R}^1$ are locally Lipschitz continuous. For the purpose of conceptual algorithms, it is convenient to define the function

$$\psi(x) \triangleq \max_{j \in \mathbf{m}} g^j(x) \tag{3.2}$$

and to treat problem (3.1) in the simpler form

$$\min\{f(x) \,|\, \psi(x) \leqq 0\}. \tag{3.3}$$

In implementable algorithms, since we may not be able to obtain a formula for the set $\partial_\varepsilon\psi(x)$, we may have to use the possibly bigger set

$$M_\varepsilon(x) \triangleq \mathrm{co}\left\{\bigcup_{j \in I_\varepsilon(x)} \partial_\varepsilon g^j(x)\right\} \tag{3.4a}$$

with

$$I_\varepsilon(x) \triangleq \{j \in \mathbf{m} \,|\, g^j(x) \geqq \psi(x) - \varepsilon\}. \tag{3.4b}$$

It is quite easy to construct an appropriate counterpart to Theorem 2.3, for algorithms which generate sequences $\{x_i\}$ by a construction of the phase-I–phase-II feasible directions type [P3], using parameters $\alpha, \beta \in (0, 1)$, viz.:

$$x_{i+1} = x_i + \lambda_i h_i, \qquad i = 0, 1, 2, \cdots, \tag{3.5a}$$

$$\lambda_i = \begin{cases} \arg\max \{\beta^k \mid \psi(x_i + \beta^k h_i) - \psi(x) \leqq \alpha\beta^k \delta_i < 0;\ k \in \mathbb{N}^+\} & \text{if } \psi(x_i) > 0 \\ \arg\max \{\beta^k \mid f(x_i + \beta^k h_i) - f(x_i) \leqq \alpha\beta^k \delta_i < 0;\ \psi(x_i + \beta^k h_i) \leqq 0;\ k \in \mathbb{N}^+\} & \text{if } \psi(x_i) \leqq 0. \end{cases} \tag{3.5b}$$

Since ε-smearing was needed for the unconstrained case, it is a foregone conclusion that it is also needed for the constrained case and we shall not go into any further justifications of "ε-smearing." Also, for the phase-I part of the algorithms to work, we need the following:

Assumption 3.1. For all $x \in \mathbb{R}^n$ such that $\psi(x) \geqq 0$, $0 \notin \partial\psi(x)$.

This assumption ensures that a feasible point can be computed by means of an unconstrained optimization algorithm in a finite number of iterations.

THEOREM 3.1 (conceptual algorithms). *Suppose that Assumption* 3.1 *holds and that there exist three functions*

$$N_1, N_2, N_3 : \mathbb{R}^n \to \mathbb{R}^+ \tag{3.6}$$

such that:

(i) *if* $N_1(x)N_2(x)N_3(x) = 0$, *then either* $\psi(x) = 0$ *and* $0 \in \operatorname{co}(\partial f(x) \cup \partial\psi(x))$, *or* $\psi(x) < 0$ *and* $0 \in \partial f(x)$,

(ii) *for every* $x \in \mathbb{R}^n$ *such that* $N_1(x)N_2(x)N_3(x) > 0$, *there exist a* $\rho(x) > 0$ *and* $b_i(x) > 0$, $i = 1, 2, 3$, *such that for all* $x' \in B(x, \rho(x))$,

$$N_1(x') \geqq b_1(x), \tag{3.7a}$$

$$N_2(x') \leqq b_2(x), \tag{3.7b}$$

$$N_3(x') \geqq b_3(x). \tag{3.7c}$$

Now consider the process (3.5a, b) *and suppose that for all i,*

$$d^0_{N_3(x_i)} \psi(x_i; h_i) \leqq \delta_i \leqq -N_1(x_i) \quad \text{if } \psi(x_i) \geqq -N_3(x_i), \tag{3.7d}$$

$$d^0_{N_3(x_i)} f(x_i; h_i) \leqq \delta_i \leqq -N_1(x_i) \quad \text{if } \psi(x_i) \leqq 0, \tag{3.7e}$$

$$\|h_i\| \leqq N_2(x_i). \tag{3.7f}$$

If $\{x_i\}_{i=0}^{\infty}$ *is an infinite sequence constructed by this process, then any accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies either* $\psi(\hat{x}) < 0$ *and* $0 \in \partial f(\hat{x})$ *or* $\psi(\hat{x}) = 0$ *and* $0 \in \operatorname{co}\{\partial f(\hat{x}) \cup \partial\psi(\hat{x})\}$.

Proof. We note that we can distinguish between two cases: a) $\psi(x_i) > 0$ for all i, and b) there exists an i_0 such that $\psi(x_i) \leqq 0$ for all $i \geqq i_0$.

a) Suppose that $\psi(x_i) > 0$ for all i, that $x_i \to^K \hat{x}$, with $K \subset \{0, 1, 2, 3, \cdots\}$, and that $N_1(\hat{x})N_2(\hat{x})N_3(\hat{x}) > 0$. Then, the process (3.5a, b) reduces to the one considered in Theorem 2.3, and hence we conclude that $\psi(x_i) \searrow -\infty$. But this contradicts the fact that, by continuity of ψ, $\psi(\hat{x}) \geqq 0$, and hence this case is impossible.

b1) Suppose that $\psi(x_i) \leqq 0$ for all $i \geqq i_0$ and that $x_i \to^K \hat{x}$, with $\psi(\hat{x}) < 0$, and $N_1(\hat{x})N_2(\hat{x})N_3(\hat{x}) > 0$. Then, because of our assumptions, there exist $i_1, \bar{k} \in \mathbb{N}^+$, $i_1 \geqq i_0$, such that $\psi(x_i + \beta^{\bar{k}} h_i) \leqq 0$ for all $i \geqq i$, $i \in K$. Similarly, as in the proof of Theorem 2.3,

there exist i_2, $\hat{k} \in \mathbb{N}^+$, with $i_2 \geqq i_1$ and $\hat{k} \geqq \bar{k}$, such that $\lambda_i \geqq \beta^{\hat{k}}$ for all $i \geqq i_2$, $i \in K$. Hence, for all $i \in K$, $i \geq i_2$,

$$f(x_{i+1}) - f(x_i) \leqq \alpha\beta^{k_i}\delta_i \leqq -\alpha\beta^{\hat{k}} b_3(\hat{x}) < 0. \tag{3.8}$$

But $f(x_i) \searrow$ for $i \geqq i_0$ and hence (3.8) implies that $f(x_i) \searrow -\infty$, which contradicts our assumption that $x_i \to \hat{x}$. Hence this case is not possible.

b2) Suppose that $\psi(x_i) \leqq 0$ for all $i \geqq i_0$ and that $x_i \to_K \hat{x}$, with $\psi(\hat{x}) = 0$ and $N_1(\hat{x})N_2(\hat{x})N_3(\hat{x}) > 0$. Then our assumptions lead us to the conclusion that there exists an $i_1 \geqq i_0$ and a $\hat{k} \in \mathbb{N}^+$ such that

$$f(x_i + \beta^{\hat{k}} h_i) - f(x_i) \leqq \alpha\beta^{\hat{k}} d^0_{N_3(x_i)} f(x_i; h_i) \leqq \alpha\beta^{\hat{k}}\delta_i, \tag{3.9a}$$

$$\psi(x_i + \beta^{\hat{k}} h_i) - \psi(x_i) \leqq \alpha\beta^{\hat{k}} d^0_{N_3(x_i)} \psi(x_i; h_i) \leq \alpha\beta^{\hat{k}}\delta_i, \tag{3.9b}$$

and consequently, $\lambda_i \geqq \beta^{\hat{k}}$. Therefore, (3.8) holds for all $i \geqq i_1$, $i \in K$ and the contradiction follows exactly as for case b1). We have thus shown that if $x_i \to^K \hat{x}$, then $N_1(\hat{x})N_2(\hat{x})N_3(\hat{x}) = 0$ must hold and hence the desired conclusion follows from assumption i) on N_1, N_2, N_3. □

We are now ready to apply this theorem to two phase-I–phase-II methods in the class of the ones presented in [P3] for differentiable optimization. We begin with the simpler one. We shall need the following definitions. Let

$$\psi(x)_+ \triangleq \max\{0, \psi(x)\}. \tag{3.10}$$

Let $\varepsilon_0 > 0$ and $\nu \in (0, 1)$ be given and let

$$\mathscr{E} = \{\varepsilon \mid \varepsilon = \varepsilon_0 \nu^k, k \in \mathbb{N}^+\} \cup \{0\}. \tag{3.11}$$

Next, let $\gamma > 0$ be given and let $\Gamma: \mathbb{R}^n \to \mathbb{R}^1$ be defined by

$$\Gamma(x) \triangleq \exp(-\gamma\psi(x)_+). \tag{3.12}$$

Finally, for any $\varepsilon \geqq 0$, $\delta > 0$, we define

$$\partial^+_\varepsilon \psi(x) \triangleq \begin{cases} \partial_\varepsilon \psi(x) & \text{if } \psi(x) \geqq -\varepsilon, \\ \phi & \text{if } \psi(x) < -\varepsilon, \end{cases} \tag{3.13a}$$

$$h^f_\varepsilon(x) \triangleq -\mathrm{Nr}\,(\mathrm{co}\,\{\partial_\varepsilon f(x), \partial^+_\varepsilon \psi(x)\}), \tag{3.13b}$$

$$h^\psi_\varepsilon(x) \triangleq -\mathrm{Nr}\,(\partial_\varepsilon \psi(x)), \tag{3.13c}$$

$$\theta^1_\varepsilon(x) \triangleq -\max\{\|\Gamma(x)h^f_\varepsilon(x)\|^2, \|(1-\Gamma(x))h^\psi_\varepsilon(x)\|^2\}, \tag{3.13d}$$

$$h^1_\varepsilon(x) \triangleq \Gamma(x)h^f_\varepsilon(x) + (1-\Gamma(x))h^\psi_\varepsilon(x), \tag{3.13e}$$

$$\varepsilon^1(x) \triangleq \max\{\varepsilon \in \mathscr{E} \mid \theta^1_\varepsilon(x) \leqq -\delta\varepsilon\}. \tag{3.13f}$$

We recognize $h^\psi_\varepsilon(x)$ as a "steepest descent" direction for $\psi(\cdot)$ at an infeasible point, while $h^f_\varepsilon(x)$ is a "usable" feasible direction when x is feasible. The vector $h^1_\varepsilon(x)$ moves from $h^\psi_\varepsilon(\cdot)$ to $h^f_\varepsilon(\cdot)$ as x moves from the infeasible into the feasible region. This type of construction is the essence of the algorithms presented in [P3] and ensures that the possible increase in cost is kept in check as the feasible region is approached.

ALGORITHM 3.1 (conceptual).
Parameters: $\alpha, \beta, \nu \in (0, 1)$, $\varepsilon_0, \delta, \gamma > 0$.
Data: $x_0 \in \mathbb{R}^n$.
Step 0: Set $i = 0$.
Step 1: Compute $h_i = h^1_{\varepsilon^1(x_i)}(x_i)$. Stop if $h_i = 0$.

Step 2: If $\psi(x_i)>0$, compute the largest stepsize β^{k_i}, $k_i \in \mathbb{N}^+$ such that

$$\psi(x_i+\beta^{k_i}h_i)-\psi(x_i)\leqq -\alpha\beta^{k_i}\|h_i\|^2. \tag{3.14a}$$

If $\psi(x_i)\leqq 0$, compute the largest step size β^{k_i}, $k_i\in\mathbb{N}^+$, such that

$$f(x_i+\beta^{k_i}h_i)-f(x_i)\leqq -\alpha\beta^{k_i}\|h_i\|^2 \tag{3.14b}$$

and

$$\psi(x_i+\beta^{k_i}h_i)\leqq 0 \tag{3.14c}$$

Step 3: Set $x_{i+1}=x_i+\beta^{k_i}h_i$, set $i=i+1$ and go to step 1.

To bring this algorithm into correspondence with Theorem 3.1, we define

$$N_1(x)\triangleq -\theta^1_{\varepsilon^1(x)}(x), \tag{3.15a}$$

$$N_2(x)\triangleq \arg\max\{\|h\|\,|h\in \operatorname{co}\{\partial_{\varepsilon_0}f(x),\ \partial^+_{\varepsilon_0}\psi(x)\}\}, \tag{3.15b}$$

$$N_3(x)\triangleq \varepsilon^1(x), \tag{3.15c}$$

and we set

$$\delta_i\triangleq -\|h_i\|^2 \quad \text{for } i=0,1,2,\cdots. \tag{3.15d}$$

LEMMA 3.1. *For every* $\varepsilon\geqq 0$ *and any* $x\in\mathbb{R}^n$,

$$\|h^1_\varepsilon(x)\|^2\geqq -\theta^1_\varepsilon(x). \tag{3.16}$$

Proof. Case 1. Suppose that $\psi(x)<-\varepsilon$. Then $\|h^1_\varepsilon(x)\|^2=-\theta^1_\varepsilon(x)$.

Case 2. $\psi(x)\geqq-\varepsilon$. Consider the function $g\colon[0,1]\to\mathbb{R}^1$ defined by

$$g(t)\triangleq \|th^f_\varepsilon(x)+(1-t)h^\psi_\varepsilon(x)\|^2-(1-t)^2\|h^\psi_\varepsilon(x)\|^2. \tag{3.17}$$

Then $g(0)=0$, $g(1)=\|h^f_\varepsilon(x)\|^2\geqq 0$ and

$$\begin{aligned}\frac{d^2}{dt^2}g(t)&=2\{\|h^f_\varepsilon(x)-h^\psi_\varepsilon(x)\|^2-\|h^\psi_\varepsilon(x)\|^2\}\\&=2\{\|h^f_\varepsilon(x)\|^2-2\langle h^f_\varepsilon(x),h^\psi_\varepsilon(x)\rangle\}\leqq 0,\end{aligned} \tag{3.18}$$

because $\langle h^f_\varepsilon(x),h^\psi_\varepsilon(x)\rangle\geqq\|h^f_\varepsilon(x)\|^2$, by construction of $h^f_\varepsilon(x)$ and $h^\psi_\varepsilon(x)$. Hence $g(\cdot)$ is concave on $[0,1]$ and, since $g(0)=0$ and $g(1)\geqq 0$, $g(t)\geqq 0$ for all $t\in[0,1]$. Consequently,

$$\|h^1_\varepsilon(x)\|^2\geqq(1-\Gamma(x))^2\|h^\psi_\varepsilon(x)\|^2. \tag{3.19}$$

Similar reasoning gives that

$$\|h^1_\varepsilon(x)\|^2\geqq\Gamma(x)^2\|h^f_\varepsilon(x)\|^2, \tag{3.20}$$

and the proof is complete. □

COROLLARY 3.1. *With* δ_i *defined by* (3.15d) *and* $N_1(x_i)$ *defined by* (3.15a), *we have* $\delta_i\leqq -N_1(x_i)$ *for all* i.

PROPOSITION 3.1. *Consider the functions* $\theta^1_\varepsilon(\cdot)$ *defined in* (3.13d).

a) *For any* $x\in\mathbb{R}^n$, *if* $\varepsilon'>\varepsilon''\geqq 0$, *then* $\theta^1_{\varepsilon'}(x)\geqq\theta^1_{\varepsilon''}(x)$.

b) *For any* $\varepsilon\geqq 0$, $\theta^1_\varepsilon(\cdot)$ *is u.s.c.*

Proof. a) Since $\varepsilon'>\varepsilon''$ implies that $\partial_{\varepsilon'}\psi(x)\supset\partial_{\varepsilon''}\psi(x)$ and $\partial_{\varepsilon'}f(x)\supset\partial_{\varepsilon''}f(x)$, this part is obvious.

b) Since for any $\varepsilon\geqq 0$, $\partial^+_\varepsilon\psi(\cdot)$ and $\partial_\varepsilon f(\cdot)$ are both u.s.c., it follows from the maximum theorem in [B1] that $\|h^f_\varepsilon(\cdot)\|^2$ and $\|h^\psi_\varepsilon(\cdot)\|^2$ are l.s.c. Hence $\theta^1_\varepsilon(\cdot)$ is u.s.c. □

LEMMA 3.2. *For every* $\bar{x} \in \mathbb{R}^n$ *such that* $\theta_0^1(\bar{x}) \neq 0$, $\varepsilon^1(x) > 0$ *and there exists a* $\bar{\rho} > 0$ *such that*

$$N_3(x') \triangleq \varepsilon^1(x') \geqq \nu\varepsilon^1(\bar{x}) \triangleq b_3(\bar{x}) > 0 \quad \text{for all } x' \in B(\bar{x}, \bar{\rho}). \tag{3.21}$$

Proof. First, because the set valued maps $\partial f(\cdot)$ and $\partial\psi(\cdot)$ are u.s.c., and $\theta_0^1(x) < 0$, there must exist an $\bar{\varepsilon} \in \mathscr{E}$, $\bar{\varepsilon} > 0$, such that $\theta_{\bar{\varepsilon}}^1(\bar{x}) \leqq -\delta\bar{\varepsilon}$. Hence $\varepsilon^1(\bar{x}) > 0$. Now, for the sake of contradiction, suppose that there is no $\bar{\rho} > 0$ such that (3.21) holds. Then there must exist a sequence $\{x_i\}$, $x_i \to \bar{x}$ such that

$$\theta^1_{\nu\varepsilon^1(x)}(x_i) > -\delta\nu\varepsilon^1(\bar{x}) \quad \text{for all } i. \tag{3.22}$$

Since by Lemma 3.2 $\theta^1_{\nu\varepsilon^1(x)}(\cdot)$ is u.s.c., we conclude from (3.22) that

$$-\delta\nu\varepsilon^1(\bar{x}) \leqq \overline{\lim}\, \theta^1_{\nu\varepsilon^1(\bar{x})}(x_i) \leqq \theta^1_{\nu\varepsilon^1(\bar{x})}(\bar{x}) \tag{3.23a}$$

But, by Lemma 3.2, $\theta^1_{\varepsilon^1(\bar{x})}(\bar{x}) \geqq \theta^1_{\nu\varepsilon^1(\bar{x})}(\bar{x})$, and hence (3.22a) implies that

$$-\delta\varepsilon^1(\bar{x}) < \theta^1_{\varepsilon^1(\bar{x})}(\bar{x}), \tag{3.23b}$$

which contradicts the definition of $\varepsilon^1(\bar{x})$. □

THEOREM 3.2. *Let* $\{x_i\}_{i=0}^{\infty}$ *be any sequence constructed by Algorithm* 3.1. *Then any accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies* $\psi(\hat{x}) \leqq 0$ *and* $0 \in \text{co}\{\partial f(\hat{x}) \cup \partial\psi(\hat{x})\}$.

Proof. With N_1, N_2, N_3, δ_i defined as in (3.15a–d), we see that at any $\bar{x}$ such that $N_1(\bar{x})N_2(\bar{x})N_3(\bar{x}) \neq 0$, by Lemma 3.2, there exists a $\bar{\rho} > 0$ such that $b_1(\bar{x}) = b_3(\bar{x}) = \nu\varepsilon^1(\bar{x}) > 0$ satisfy (3.7a) and (3.7c) for all $x' \in B(\bar{x}, \bar{\rho})$. Since $\partial_{\varepsilon_0} f(x)$ and $\partial_\varepsilon f(x)$ are both u.s.c., it is clear that a required $b_2(\bar{x}) > 0$ exists for (3.7b) to hold in $B(\bar{x}, \bar{\rho})$. Finally, by Corollary 3.1, we have that $\delta_i \leqq -N_1(x_i)$ for all i. Furthermore, Assumption 3.1 and Lemma 3.2 ensure that $N_1(x)N_2(x)N_3(x) = 0$ implies that condition (i) of Theorem 3.1 is satisfied. Consequently, the desired result follows directly from Theorem 3.1. □

Our second algorithm has exactly the same structure as Algorithm 3.1, except that h_i is computed by evaluating a different optimality function, $\theta^2_{\varepsilon(x)}(x)$. It is a direct extension of the most efficient phase-I–phase-II method of feasible directions known [P3]. We need the following notation. Given $\gamma > 0$, for any $\varepsilon \geqq 0$ and $x \in \mathbb{R}^n$ we define

$$\theta_\varepsilon^2(x) \triangleq \min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}\|h\|^2 + \max\{\langle \xi_f, h\rangle - \gamma\psi_+(x),\ \xi_f \in \partial_\varepsilon f(x); \langle \xi_\psi, h\rangle,\ \xi_\psi \in \partial_\varepsilon^+\psi(x)\}\} \tag{3.24a}$$

and

$$h_\varepsilon^2(x) = \arg\min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}\|h\|^2 + \max\{\langle \xi_f, h\rangle - \gamma\psi_+(x),\ \xi_f \in \partial_\varepsilon f(x); \langle \xi_\psi, h\rangle,\ \xi_\psi \in \partial_\varepsilon^+\psi(x)\}\}. \tag{3.24b}$$

It follows by duality that when $\psi_+(x) = 0$, for all $\varepsilon \geqq 0$, $\theta_\varepsilon^1(x) = \theta_\varepsilon^2(x)$ and $h_\varepsilon^2(x) = h_\varepsilon^f(x)$. Hence, the behavior of the two algorithms can differ only in the infeasible region. We now define

$$\varepsilon^2(x) \triangleq \max\{\varepsilon \in \mathscr{E} \mid \theta_\varepsilon^2(x) \leqq -\delta\varepsilon\} \tag{3.25}$$

where ε and δ are as in (3.13f).

Not surprisingly, the conclusions of Lemma 3.1, Proposition 3.1, Lemma 3.2 and Corollary 3.1 remain valid when $\varepsilon^2(x)$, $h_\varepsilon^2(x)$ and $\theta_\varepsilon^2(x)$ are substituted for $\varepsilon^1(x)$, $h_\varepsilon^1(x)$ and $\theta_\varepsilon^1(x)$ in the appropriate definitions. Consequently, we may state without proof the following:

THEOREM 3.3. *Suppose that Algorithm* 3.1 *is modified so that* $h_i = h^2_{\varepsilon^1(x_i)}(x_i)$ *in Step* 1. *If* $\{x_i\}_{i=0}^{\infty}$ *is an infinite sequence constructed by this modified algorithm, then any accumulation point* $\hat{x}$ *of* $\{x_i\}_{i=0}^{\infty}$ *satisfies* $\psi(\hat{x}) \leqq 0$ *and* $0 \in \text{co}\{\partial f(\hat{x}) \cup \partial_0^+\psi(\hat{x})\}$.

Finally we turn to problems with both inequality and equality constraints, i.e., problems of the form

$$\text{P:}\quad \min\{f(x)|g^i(x)\leqq 0, i\in\mathbf{m}; h^j(x)=0, j\in\mathbf{l}\} \tag{3.26}$$

where $f, g^i, i\in\mathbf{m}$ and $h^j, j\in\mathbf{l}$, from $\mathbb{R}^n$ into $\mathbb{R}$ are all locally Lipschitz continuous. In the differentiable case, i.e., when f, g^i and $h^j, i\in\mathbf{m}, j\in\mathbf{l}$ are all continuously differentiable, there are two major approaches, based on exact penalty functions, for solving (3.26). The first is due to Mayne and Polak ([M3]). It replaces the problem P with P_c^1, below, $c>0$

$$\text{P}_c^1\text{:}\quad \min\left\{f(x)-c\sum_{j\in\mathbf{l}}h^j(x)|g^i(x)\leqq 0, i\in\mathbf{m}; h^j(x)\leqq 0, j\in\mathbf{l}\right\} \tag{3.27}$$

and, under mild assumptions, computes a finite $\bar{c}$ which makes P_c^1 and P locally equivalent in the vicinity of Kuhn–Tucker points of P for all $c\geqq\bar{c}$. The second approach, see e.g. [C5], [P14], replaces P with P_c^2, below (with $c>0$):

$$\text{P}_c^2\text{:}\quad \min_{x\in\mathbb{R}^n} f_c(x), \tag{3.28a}$$

where

$$f_c(x)\triangleq f(x)+c\max\left\{\max_{i\in\mathbf{m}} g^i(x)_+, \max_{j\in\mathbf{l}}|h^j(x)|\right\}. \tag{3.28b}$$

Again, it can be shown that, under mild assumptions, P and P_2^c are locally equivalent or c sufficiently large, in the vicinity of feasible Kuhn–Tucker points of P (see [P14]).

In the nondifferentiable case, both approaches tend to break down when equality constraints are present, because stationary points of P_c^1 or P_c^2 which are feasible for P cannot be shown to be also stationary for P. Furthermore, arbitrary feasible points of P may be stationary for P^c or P_2^c. Thus, consider the problem P_2^c. Suppose, for simplicity, that there are no inequality constraints in P, and that $l=1$, i.e., that there is only one equality constraint. Then (3.26) and (3.28a) become

$$\text{P:}\quad \min_{x\in\mathbb{R}^n}\{f(x)|h(x)=0\} \tag{3.29}$$

and

$$\text{P}_c^2\text{:}\quad \min_{x\in\mathbb{R}^n}\{f(x)+c|h(x)|\}, \tag{3.30}$$

respectively. Suppose that for some $c>0$, $\hat{x}\in\mathbb{R}^n$ satisfies the necessary optimality condition for P_c^2, and that $h(\hat{x})=0$. Then

$$0\in\partial f(\hat{x})+c\ \text{co}\{\partial h(\hat{x})\cup-\partial h(\hat{x})\}. \tag{3.31}$$

Now, from (3.31) we would like to conclude that (1.8) holds, i.e., that either

$$0\in\text{co}\{\partial f(\hat{x})\cup\partial h(\hat{x})\} \tag{3.32a}$$

or

$$0\in\text{co}\{\partial f(\hat{x})\cup-\partial h(\hat{x})\}. \tag{3.32b}$$

While in the differentiable case (3.32a) or (3.32b) follows directly from (3.31), a similar conclusion does not hold in general in the nondifferentiable case, as can be

seen from the following example. Let $x = (x^1, x^2)^T \in \mathbb{R}^2$, let $f(x) = -\frac{1}{2}x^1$, let

$$h(x) = \begin{cases} (x^1)^2 + (x^2)^2 - 5 & \text{if } x^1 \leqq 1, \\ x^1 + (x^2)^2 - 5 & \text{if } x^1 \geqq 1, \end{cases}$$

and let $\hat{x} = (1, 2)^T$. Then $\hat{x}$ is feasible for P in (3.29) and $\partial h(\hat{x}) = \text{co}\{(1,4)^T, (2,4)^T\}$. It is easily seen that for all $c \geqq 1$ (3.31) holds, but neither (3.32a) nor (3.32b) does.

This example shows that when $\partial h(\hat{x})$ is not contained in a one-dimensional subspace of $\mathbb{R}^2$ and $h(\hat{x}) = 0$, then co $\{\partial h(\hat{x}) \cup -\partial h(\hat{x})\}$ can be "blown up" by increasing c so that $\hat{x}$ becomes a stationary point for $f_c(\cdot)$, i.e., arbitrary feasible points of P become stationary points of P_2^c. Hence it seems that an exact penalty function method can be generalized to the nondifferentiable case only when the generalized gradients of all the equality constraints are each contained in a one-dimensional subspace of $\mathbb{R}^n$, so that co $\{\partial h^j(\hat{x}) \cup -\partial h^j(\hat{x})\}$ does not have an interior point in any multi-dimensional space. In the presence of inequality constraints alone, exact penalty methods should work, for the following reason. Suppose that $m = 1$ and that $\hat{x}$ satisfies $g(\hat{x}) = 0$ and $0 \in \partial f(\hat{x}) + c\,\partial g(\hat{x})_+$ for some $c > 0$. Then we have that

$$\xi_f + c\alpha\xi_\psi = 0 \tag{3.33}$$

for some $\xi_f \in \partial f(\hat{x})$, $\xi_\psi \in \partial g(\hat{x})$ and $\alpha \in [0, 1]$. Consequently,

$$(1 + c\alpha)\left[\frac{1}{1 + c\alpha}\xi_f + \frac{c\alpha}{1 + c\alpha}\xi_\psi\right] = 0, \tag{3.34}$$

i.e., $0 \in \text{co}\{\partial f(\hat{x}), \partial g(\hat{x})\}$. Hence it should be possible to solve P by exact penalty function methods, provided the following assumption holds:

Assumption 3.2. For all $j \in \mathbf{l}$, the functions $h^j(\cdot)$ are continuously differentiable.

For the differentiable case the approach based on P_c^1 is considerably more attractive, since it permits the use of a broad class of algorithms for solving P, in particular, second order algorithms. However, this advantage is not obvious for the nondifferentiable case. We will therefore consider here the more traditional approach based on P_c^2. Although it is not possible to precompute a satisfactory penalty $\hat{c}$ for P_c^2, the theory in [P10] on abstract exact penalty methods shows that such a penalty can be computed adaptively, provided an appropriate test function can be constructed. We shall exhibit such a test function for the problems in question.

We now define

$$\eta(x) \triangleq \max_{j \in l} |h^j(x)|, \tag{3.35a}$$

$$\psi(x)_+ \triangleq \max_{i \in \mathbf{m}} g^i(x)_+, \tag{3.35b}$$

$$\phi(x) \triangleq \max\{\eta(x), \psi(x)_+\}. \tag{3.35c}$$

Next we establish a number of properties of the problem P_c^2. The first one is obvious.

PROPOSITION 3.2. *Suppose $\hat{x}$ is a local minimizer for P_c^2 such that $\phi(\hat{x}) = 0$. Then $\hat{x}$ is also a local minimizer for* P.

PROPOSITION 3.3. *Suppose that Assumption* 3.2 *holds and that $\hat{x} \in \mathbb{R}^n$ is feasible for* P *and for some $c > 0$ satisfies*

$$0 \in \partial f(\hat{x}) + c\ \text{co}\{\partial g^i(\hat{x}), i \in I(\hat{x}); \partial|h^j(\hat{x})|, j \in l\}. \tag{3.36}$$

Then $\hat{x}$ satisfies (1.8).

Proof. By assumption, there exist: (i) $\xi_f \in \partial f(\hat{x})$, (ii) $\xi_i \in \partial g^i(\hat{x})$ and $t_i \in [0, 1]$ for all $i \in I(\hat{x})$, (iii) $s_j \in [-1, 1]$ for all $j \in \mathbf{l}$, such that

$$\xi_f + c \sum_{i \in I(\hat{x})} t_i \xi_i + c \sum_{j \in \mathbf{l}} s_j \nabla h^j(\hat{x}) = 0. \tag{3.37}$$

By dividing each element of (3.37) by $1 + c(\sum_{i \in I(\hat{x})} t_i + \sum_{j \in \mathbf{l}} |t_j|)$ we get (1.8). □

Before we can establish the existence of finite penalties, we must invoke the following commonly used hypothesis.

Assumption 3.3. For every $x \in \mathbb{R}^n$ and any $t_1, t_2, \cdots, t_l \in \{-1, 1\}$,

$$0 \notin \operatorname{co}\{\partial g^i(x), i \in K_g^0(x); t_j \nabla h^j(x), j \in K_h^0(x)\} \tag{3.38a}$$

where for any $\varepsilon \geqq 0$

$$K_g^\varepsilon(x) \triangleq \{i \in \mathbf{m} | g^i(x) \geqq \phi(x) - \varepsilon\}, \tag{3.38b}$$

$$K_h^\varepsilon(x) \triangleq \{i \in \mathbf{l} | |h^j(x)| \geqq \phi(x) - \varepsilon\}. \tag{3.38c}$$

We are now ready to establish the existence of exact penalties.

PROPOSITION 3.4. *Suppose that $\hat{x}$ satisfies $\phi(\hat{x}) \leqq 0$ and* (1.10). *Then there exists a $\hat{c} > 0$ such that*

$$0 \in \partial f(\hat{x}) + c \operatorname{co}\{\partial g^i(\hat{x})_+, i \in K_g^0(\hat{x}); \partial |h^j(\hat{x})|, j \in K_h^0(\hat{x})\} \tag{3.39}$$

for all $c > \hat{c}$, i.e., $\hat{x}$ is stationary for P_c.

Proof. By Theorem 1.1 and Assumption 3.3, there exist

$$\xi_f \in \partial f(\hat{x}), \quad \mu^i \geqq 0, \qquad \xi_{\psi,i} \in \partial g^i(\hat{x}) \cap \partial g^i(\hat{x})_+, \quad i \in I(\hat{x}),$$

and $\lambda^j \in \mathbb{R}$, $j \in \mathbf{l}$, such that

$$\xi_f + \sum_{i \in I(\hat{x})} \mu^i \xi_{\psi,i} + \sum_{j \in \mathbf{l}} \lambda^j \nabla h^j(\hat{x}) = 0. \tag{3.40a}$$

Therefore, for all $c > 0$,

$$\xi_f + c \sum_{i \in K_g^0(\hat{x})} \frac{\mu^i}{c} \xi_{\psi,i} + c \sum_{j \in K_h^0(\hat{x})} \frac{\lambda^j}{c} \nabla h^j(\hat{x}) = 0, \tag{3.40b}$$

since $K_g^0(\hat{x}) = I(\hat{x})$ and $K_h^0(\hat{x}) = l$. Clearly, there exists a $\hat{c} > 0$ such that for all $c \geqq \hat{c}$,

$$\frac{1}{c}\left[\sum_{i \in K_g^0(\hat{x})} \mu^i + \sum_{j \in K_h^0(\hat{x})} |\lambda^j|\right] \leqq 1,$$

and hence

$$\left(\sum_{i \in K_g^0(\hat{x})} \frac{\mu^i}{c} \xi_{\psi,i} + \sum_{j \in K_h^0(\hat{x})} \frac{\lambda^j}{c} \nabla h^j(\hat{x})\right) \in \operatorname{co}\{\partial g^i(\hat{x})_+, i \in K_g^0(\hat{x}); \partial |h^j(\hat{x})|, j \in K_h^0(\hat{x})\},$$

which leads to (3.39). □

The following proposition is a direct corollary of Assumption 3.3.

PROPOSITION 3.5. *Suppose that $\hat{x} \in \mathbb{R}^n$ is such that $\phi(\hat{x}) > 0$. Then there exists a $\hat{c} > 0$ such that for all $c \geqq \hat{c}$*

$$0 \notin [\partial f(\hat{x}) + c \operatorname{co}\{\partial g^i(\hat{x})_+, i \in K_g^0(\hat{x}); \partial |h^j(\hat{x})|, j \in K_h^0(\hat{x})\}]. \tag{3.41}$$

Proof. Since $\phi(\hat{x}) > 0$, $\partial g^i(\hat{x})_+ = \partial g^i(\hat{x})$ for all $i \in K_g^0(\hat{x})$ and $\partial |h^j(\hat{x})| = \nabla |h^j(\hat{x})|$ for all $j \in K_h^0(\hat{x})$.

By Assumption 3.3 there exists a $\delta > 0$ such that for any $\xi^i \in \partial g^i(\hat{x})$ with $i \in I_0(\hat{x})$

$$\mathrm{Nr}\,(\mathrm{co}\,\{\partial g^i(\hat{x}),\ i \in K_g^0(\hat{x});\ \nabla |h^j(\hat{x})|,\ j \in K_h^0(\hat{x})\}).$$

Consequently, Proposition 3.5 holds with

$$\hat{c} = \frac{1}{\delta} \cdot \max\{\|\xi_f\| \,|\, \xi_f \in \partial f(\hat{x})\}. \qquad \square$$

We now construct an exact penalty function method which computes the required penalty parameter c adaptively, making use of the scheme proposed in [P10]. This scheme uses a test function $t_c(\cdot)$ to determine whether c should be increased or not. As in (2.18) and (2.19), we define, for $\varepsilon \geqq 0$ and any $x \in \mathbb{R}^n$,

$$h_{c,\varepsilon}(x) \triangleq -\mathrm{Nr}\,(\partial_\varepsilon^0 f_c(x)) \tag{3.42a}$$

where

$$\partial_\varepsilon^0 f_c(x) \triangleq \partial_\varepsilon f(x) + c\ \mathrm{co}\,\{\partial_\varepsilon g^i(x)_+,\ i \in K_g^\varepsilon(x);\ \partial_\varepsilon |h^j(x)|,\ j \in K_h^\varepsilon(x)\} \tag{3.42b}$$

and (with $\delta > 0$)

$$\varepsilon_c(x) \triangleq \max\{\varepsilon \in \mathscr{E} \,|\, \|h_{c,\varepsilon}(x)\|^2 \geqq \delta\varepsilon\}. \tag{3.42c}$$

Then for any $c > 0$, $x \in \mathbb{R}^n$ we define

$$t_c(x) \triangleq -\varepsilon_c(x) + \frac{1}{c}\phi(x). \tag{3.42d}$$

In accordance with [M3], we therefore propose the following conceptual algorithm:

ALGORITHM 3.2.

Parameters: $\alpha, \beta, \nu \in (0, 1)$, $\varepsilon_0 > 0$, $\delta > 0$, and a sequence $\{c_j\}_{i=0}^{\infty} \subset \mathbb{R}^+$, $c_j \nearrow \infty$.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$, $j = 0$.

Step 1: If $t_{c_j}(x_i) > 0$, set $z_j = x_i$ and increase j to the first j^* such that $t_{c_{j^*}}(x_i) \leqq 0$. Set $j = j^*$.

Step 2: If $0 \in \partial f_{c_j}(x_i)$, stop. Else compute x_{i+1} by applying Algorithm 2.1 to $f_{c_j}(\cdot)$, from x_i, using the parameters supplied. Set $i = i + 1$ and go to Step 1.

THEOREM 3.4. (i) *If* $\{z_j\}$ *is finite, with last element* $z_j{}^*$, *then either the sequence* $\{x_i\}$ *is finite and its least element, say* x_k, *satisfies* $\phi(x_k) = 0$ *and* (1.8), *or it is infinite and any accumulation point of* $\{x_i\}$, *say* $\hat{x}$, *satisfies* $\phi(\hat{x}) = 0$ *and* (1.8).

(ii) *If* $\{z_j\}$ *is infinite, then it has no accumulation points.*

Proof. (i) Suppose that both $\{x_i\}$ and $\{z_j\}$ are finite, $\{x_i\}$ terminating at x_k. Then for some $j = j^*$, we must have $0 \in \partial f_{c_{j^*}}(x_k)$ and $t_{c_{j^*}}(x_k) \leqq 0$. Since $\varepsilon_{c_{j^*}}(x_k) = 0$, it follows that $\phi(x_k) = 0$, and since $0 \in \partial f_{c_{j^*}}(x_k)$, it follows from Proposition 3.3 that

$$0 \in \mathrm{co}\,\{\partial f(\hat{x});\ \partial g^i(\hat{x})_+,\ i \in I(\hat{x});\ \partial |h^j(\hat{x})|,\ j \in \mathbf{I}\}.$$

Next, suppose that $\{x_i\}$ is infinite, with $x_i \to^K \hat{x}$, $K \subset \mathbb{N}^+$ and that $\{z_j\}$ is finite, terminating at j^*. Let i_{j^*} be such that $x_{i_{j^*}} = z_{j^*}$. Then for all $i > i_{j^*}$ we have that

$$t_{c_{j^*}}(x_i) = -\varepsilon_{c_{j^*}}(x_i) + \frac{1}{c_{j^*}}\phi(x_i) \leqq 0. \tag{3.43}$$

But as in the proof of Theorem 2.1 we have that $\varepsilon_{c_j*}(x_i) \to^K \varepsilon_{c_j*}(\hat{x}) = 0$, and hence from (3.36) and the continuity of ϕ we get that $\phi(\hat{x}) = 0$. Finally, since $\varepsilon_{c_j*}(\hat{x}) = 0$ implies that $0 \in \partial_0 f_{c_j*}(\hat{x})$, it follows that $0 \in \text{co}\{\partial f(\hat{x}); \partial g^i(\hat{x})_+, i \in I(\hat{x}); \partial |h^j(\hat{x})|, j \in \mathbf{I}\}$.

(ii) Suppose that $\{z_j\}$ is infinite. To simplify notation, we assume that j assumes all the values in $\mathbb{N}^+$. Now suppose that $z_j \to^K \hat{z}$, for some subsequence indexed by $K \subset \mathbb{N}^+$. Since $t_{c_j}(z_j) > 0$, by construction of c_{j+1}, we must have that $\phi(z_j) > 0$ for all $j \in \mathbb{N}^+$, and since $\phi(\cdot)$ is continuous, it follows that $\phi(\hat{z}) \geqq 0$. Also, since $\{\phi(z_j)\}_{j \in K}$ is bounded, it follows that $(1/c_j)\phi(z_j) \to 0$ as $j \to \infty$. We distinguish two cases.

Case 1. Either $\phi(\hat{z}) > 0$, or $\phi(\hat{z}) = 0$ and (1.10) does not hold. In either event, it follows from Propositions 3.3 and 3.5 that there exist a $c^* > 0$ such that $0 \notin \partial_0^0 f_c(\hat{z})$ for all $c > c^*$.

First suppose that $\phi(\hat{z}) > 0$. Then for all $c > 0$,

$$\partial_0^0 f_c(\hat{z}) = \partial f(\hat{z}) + c \text{ co}\{\partial g^i(\hat{z}), i \in K_g^0(\hat{z}); \partial |h^j(\hat{z})|, j \in K_h^0(\hat{z})\} \tag{3.44}$$

and, hence, because of Assumption 3.3, $\|\text{Nr}(\partial_0^0 f_c(z))\| \to \infty$, as $c \to \infty$. Next suppose that $\phi(\hat{z}) = 0$ and that (1.10) does not hold. Then, because of Proposition 3.4, $0 \notin \partial_0^0 f_c(z)$ for all $c > 0$. It now follows from the fact that $0 \notin \partial f(\hat{z})$ that there exists a $\delta > 0$ such that $\|\text{Nr}(\partial f_c^0(\hat{z}))\| \geqq \delta$ for all $c > 0$.

Consequently, in both cases considered, there must exist a $\hat{\rho} > 0$ and an $\hat{\varepsilon} > 0$, such that for all $c > c^*$, for all $z \in B(\hat{z}, \hat{\rho})$, $\varepsilon(z) > \hat{\varepsilon}$. Now let $M = \sup_K \phi(z_j)$. Then there exists a $j^* \in K$ such that for all $j \geqq j^*$, $j \in K$, $z_j \in B(\hat{z}, \hat{\rho})$ and $-\hat{\varepsilon} + (1/c_j)M \leqq 0$. Hence $t_{c_j}(z_j) \leqq 0$, for all $j \in K$, $j \geqq j^*$, which contradicts the construction of c_{j+1}. Thus case 1 is not possible.

Case 2. Suppose that $\phi(\hat{z}) = 0$ and (1.10) holds, Then, by Proposition 3.4, there exists a $j^* \in K$ such that $0 \in \partial_0^0 f_c(z)$ for all $c \geqq c_{j*}$. Since $\phi(z_j) > 0$ for all j, and $z_j \to^K \hat{z}$, it follows from Assumption 3.3 that $\|\text{Nr}(\partial_0^0 f_{c_j}(z_j)\| \to^K \infty$ as $j \to^K \infty$. Now there exist a $\hat{\rho} > 0$ and a Lipschitz constant $L \in (0, \infty)$ such that for any $z_j, z' \in B(z, 2\hat{\rho})$, satisfying $\phi(z') = 0$, $\|z_j - z'\| \geqq \phi(zj)/L$ must hold. Hence $0 \notin \partial_\varepsilon^0 f_{c_j}(z_j)$ for all $\varepsilon > 0$ such that $\varepsilon < \phi(z_j)/L$, $j > j^*$, $j \in K$. Clearly, since $\|\text{Nr}(\partial_0^0 f_{c_j}(z))\| \to \infty$ as $j \to \infty$, for all $z \in B(\hat{z}, 2\hat{\rho})$ such that $\phi(z) > 0$, there exists a $j^{**} > j^*$, such that for all $j \geqq j^{**}$, $j \in K$, $\varepsilon_{c_j}(z_j) \geqq \nu\phi(z_j)/L \geqq \phi(z_j)/c$. Hence $t_{c_j}(z_j) \leqq 0$ for all $j \in K, j \geqq j^{**}$, which contradicts the construction of c_{j+1}.

To conclude this discussion, we must point out that one could also construct a similar exact penalty function method in which each constraint is penalized individually, by setting

$$f_c(x) \triangleq f(x) + \sum_{i=1}^{l+m} c^i g^i(x)_+ \tag{3.45}$$

with $g^{m+j} \triangleq |h^j|$ for $j = 1, 2, \cdots l$. The penalties c^i must be then be increased individually when $t_{c_j}^i(x) > 0$, with

$$t_c^i(x) \triangleq \varepsilon_c(x) + \frac{1}{c} g^i(x)_+. \tag{3.46}$$

4. Constrained optimization: Implementable algorithms. We shall consider only the problem (3.1) and the implementation of phase-I–phase-II methods, since the implementation of exact penalty function methods is essentially the same as in Algorithm 2.2.

We shall consider problem 3.1 in the compact form

$$\min\{f(x) | \psi(x) \leqq 0\} \tag{4.1}$$

with $f, \psi: \mathbb{R}^n \to \mathbb{R}^1$ locally Lipschitz and *semismooth*. Furthermore, we shall assume that $0 \notin \partial\psi(x)$ for all x such that $\psi(x) \geqq 0$. We shall make repeated use of the *bisection method* described in § 2 (eqs. (2.26)–(2.30)) which can be used (for semismooth functions) to find a $\xi \in \partial_\varepsilon f(x)$ (or $\xi \in \partial_\varepsilon \psi(x)$) such that $\langle \xi, h\rangle \leqq \bar{\alpha}\|h\|^2$ whenever $h \in \mathbb{R}^n$ is such that

$$f(x-\lambda h)-f(x) > -\alpha\lambda\|h\|^2 \tag{4.2a}$$

or

$$\psi(x-\lambda h)-\psi(x) > -\alpha\lambda\|h\|^2, \tag{4.2b}$$

with $\lambda\|h\| \leqq \varepsilon$ and $0 < \alpha < \bar{\alpha} < 1$.

We now present an implementation of Algorithm 3.1.

ALGORITHM 4.1 (implementable).

Data: $\varepsilon_0 > 0$, $\delta > 0$, $\alpha, \beta, \nu \in (0, 1)$, $\bar{\alpha} \in (\alpha, 1)$, $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$.

Step 1: Set $\varepsilon = \varepsilon_0$.

Step 2: If $\psi(x_i) \geqq -\varepsilon$, go to step 7.

CASE 1: $\psi(x_i) < -\varepsilon$.

Step 3: Set $j = 0$ and compute an $h_0^f \in \partial_\varepsilon f(x_i)$.

Step 4: If $\|h_j^f\|^2 < \delta\varepsilon$, set $\varepsilon = \nu\varepsilon$ and go to step 3.

Else, proceed.

Step 5: Set $s_j = \arg\max\{\beta^k | \beta^k \leqq (\varepsilon/\|h_j^f\|), k \in \mathbb{N}^+\}$.

Step 6: If

$$f(x_i - s_j h_j^f) - f(x_i) \leqq -s_j\alpha\|h_j^f\|^2, \tag{4.3a}$$

set $h_i = h_j^f$ and go to step 13.

Else, i) use the bisection method to compute a $\xi_j^f \in \partial_\varepsilon f(x_i)$ such that

$$\langle \xi_j^f, h_j^f\rangle \leqq \bar{\alpha}\|h_j^f\|^2, \tag{4.3b}$$

ii) compute $h_{j+1}^f = \mathrm{Nr\ co}\{\xi_j^f, h_j^f\}$, set $j = j+1$ and go to step 4.

Step 7: If $\psi(x_i) > 0$ go to step 14.

CASE 2: $\psi(x_i) \in [-\varepsilon, 0]$.

Step 8. Set $j = 0$. Compute $\xi_0^f \in \partial_\varepsilon f(x_i)$, $\xi_0^\psi \in \partial_\varepsilon \psi(x_i)$ and $h_0^f = \mathrm{Nr}\,(\mathrm{co}\{\xi_0^f, \xi_0^\psi\})$.

Step 9: If $\|h_j^f\|^2 < \delta\varepsilon$, set $\varepsilon = \nu\varepsilon$ and go to step 2.

Step 10: Set $s_j = \arg\max\{\beta^k | \beta^k \leqq (\varepsilon/\|h_j^f\|), k \in \mathbb{N}^+\}$.

Step 11: If

$$\psi(x_i - s_j h_j^f) - \psi(x_i) \leqq -s_j\alpha\|h_j^f\|^2, \tag{4.4a}$$

set $h_{j+1}^\psi = h_j^\psi$ and go to step 12.

Else, i) use the bisection method to compute a $\xi_{j+1}^\psi \in \partial_\varepsilon \psi(x_i)$ such that

$$\langle \xi_{j+1}^\psi, h_j^f\rangle \leqq \bar{\alpha}\|h_j^f\|^2, \tag{4.4b}$$

ii) compute $h_{j+1}^f = \mathrm{Nr}\,(\mathrm{co}\{\xi_j^f, \xi_i^\psi, \xi_{i+1}^\psi\})$, set $j = j+1$ and go to step 9.

Step 12: If

$$f(x_i - s_j h_j^f) - f(x_i) \leqq -s_j\alpha\|h_j^f\|^2, \tag{4.5a}$$

set $h_i = -h_j^f$ and go to step 13.

Else, i) use the bisection method to compute a $\xi_{j+1}^f \in \partial_\varepsilon f(x_i)$ such that

$$\langle \xi_{j+1}, h_j^f\rangle \leqq \bar{\alpha}\|h_j^f\|^2, \tag{4.5b}$$

ii) compute $h_{j+1}^f = \mathrm{Nr}\,(\mathrm{co}\,\{\xi_j^f, \xi_{j+1}^f, \xi_j^\psi\})$,
set $j = j+1$ and go to step 9.

Step 13: Compute

$$\lambda_i = \arg\max\{\beta^k \mid f(x_i+\beta^k h_i) - f(x_i) \leqq \alpha\beta^k \|h_i\|^2;\ \psi(x_i+\beta^k h_i) \leqq 0,\ k \in \mathbb{N}^+\}, \tag{4.6}$$

set $x_{i+1} = x_i + \lambda_i h_i$, set $i = i+1$ and go to step 1.

CASE 3: $\psi(x_i) > 0$.

Step 14: Set $j = 0$. Compute $\xi_0^f \in \partial_\varepsilon f(x_i)$, $\xi_0^\psi \in \partial_\varepsilon \psi(x_i)$, $\Gamma(x_i) = e^{-\psi(x_i)}$, $h_0^f = \mathrm{Nr}\,(\mathrm{co}\,\{\xi_0^f, \xi_0^\psi\})$. Set $h_0^\psi = \xi_0^\psi$, $h_0^\Gamma = \Gamma(x_i)h_0^f + (1-\Gamma(x_i))h_0^\psi$.

Step 15: If $\max\{\|\Gamma(x_i)h_j^f\|^2, \|(1-\Gamma(x_i))h_j^f\|^2\} < \delta\varepsilon$ set $\varepsilon = \nu\varepsilon$ and go to step 14.

Step 16: Set $s_j = \arg\max\{\beta^k \mid \beta^k \leqq \varepsilon/\|h_j^\Gamma\|,\ k \in \mathbb{N}^+\}$.

Step 17: If

$$\psi(x_i - s_j h_j^\Gamma) - \psi(x_i) \leqq -s_j\alpha\|h_j^\Gamma\|^2, \tag{4.7a}$$

set $h_i = -h_j^\Gamma$ and go to step 20.

Else, use the bisection method to compute a $\xi_{j+1}^\psi \in \partial_\varepsilon\psi(x_i)$ such that

$$\langle \xi_{j+1}^\psi, h_j^\Gamma\rangle \leqq \bar{\alpha}\|h_j^\Gamma\|^2 \tag{4.7b}$$

and proceed.

Step 18: If $j \leqq [\Gamma(x_i)^{-1}]$ (the integer part of) and $f(x_i - s_j h_j^\Gamma) - f(x_i) > s_j\alpha\|h_j^\Gamma\|^2$, use the bisection method to compute a $\xi_{j+1}^f \in \partial_\varepsilon f(x_i)$ such that

$$\langle \xi_{j+1}^f, h_j^f\rangle \leqq \bar{\alpha}\|h_g^f\|^2. \tag{4.8}$$

Else set $\xi_{j+1}^f = \xi_j^f$.

Step 19: Compute

$h_{j+1}^\psi = \mathrm{Nr}\,(\mathrm{co}\,\{h_j^\psi, \xi_{j+1}^\psi\})$, $h_{j+1}^f = \mathrm{Nr}\,(\mathrm{co}\,\{\xi_{j+1}^f, \xi_j^f, h_j^\psi, \xi_{j+1}^\psi\})$,
$h_{j+1}^\Gamma = \Gamma(x_i)h_{j+1}^f + (1-\Gamma(x_i))h_{j+1}^\psi$.
Set $j = j+1$ and go to step 15.

Step 20: Compute

$$\lambda_i = \arg\max\{\beta^k \mid \psi(x_i+\beta^k h_i) - \psi(x_i) \leqq -\beta^k\alpha\|h_i\|^2;\ k \in \mathbb{N}^+\}, \tag{4.9}$$

set $x_{i+1} = x_i + \lambda_i h_i$.
set $i = i+1$ and go to step 1.

THEOREM 4.1. *Suppose that Algorithm* 4.1 *constructs a sequence* $\{x_i\}$. *If* $\{x_i\}$ *is finite, with last element* x_N (*i.e., the algorithm jams at* x_N) *then* $\psi(x_N) \leqq 0$ *and* $0 \in \mathrm{co}\,\{\partial f(x_N) \cup \partial_0^+\psi(x_N)\}$. *If* $\{x_i\}$ *is infinite, then any accumulation point* $\hat{x}$ *of* $\{x_i\}$ *satisfies* $\psi(\hat{x}) \leqq 0$, $0 \notin \mathrm{co}\,\{\partial f(\hat{x}) \cup \partial_0^+\psi(\hat{x})\}$.

Proof. a) Suppose that $\{x_i\}$ is finite, terminating at x_N. Suppose that either $\psi(x_N) > 0$ or that $\psi(x_N) \leqq 0$ and $0 \notin \mathrm{co}\,\{\partial f(x_N) \cup \partial_0^+\psi(x_n)\}$.

Case 1. Suppose that $\psi(x_N) \leqq 0$ and $0 \notin \mathrm{co}\,\{\partial f(x_N) \cup \partial_0^+\psi(x_N)\}$. Then referring to (3.3f), $\varepsilon^1(x_N) > 0$ and we can consider two subcases:

Subcase 1*a*. The algorithm is cycling between steps 3 and 6. In this case, because f Proposition 2.3, we must have $\|h_j^f\| \to 0$ as $j \to \infty$ and hence $\varepsilon \searrow 0$ as $j \to \infty$. Consequently, there exists a j_0 such that $\varepsilon \leqq \varepsilon^1(x_N)$ for all $j \geqq j_0$ and hence (see 3.13b) we must have that $\|h_j^f\|^2 \geqq \|h^f_{\varepsilon^1(x_N)}(x_N)\|^2 > \delta\varepsilon^1(x_N)$ for all $j \geqq j_0$, which is clearly a contradiction.

Subcase 1b. The algorithm is cycling between steps 2, 7 and 12. Since by Proposition 2.3 $h_j^f \to 0$ as $j \to \infty$, $\varepsilon \searrow 0$. If $\psi(x_N) < 0$, then there exists a j such that

$\psi(x_N) < -\varepsilon$, and hence the algorithm transfers permanently into the loop defined by steps 3 to 6. But we have already shown that the algorithm cannot jam up in this loop. Hence, suppose that $\psi(x_N) = 0$. In this case, there exists a j_0 such that $\varepsilon \geqq \varepsilon^1(x_N)$ for all $j \geqq j_0$ and hence, $\|h_j^f\|^2 \geqq \|h^f_{\varepsilon^1(x_N)}(x_N)\|^2 \geqq \delta\varepsilon^1(x_N) > 0$ and, again, we have a contradiction.

Case 2. Suppose that $\psi(x_N) > 0$. Then, by Assumption 3.2, $0 \notin \partial\psi(x_N)$ and $\varepsilon^1(x_N) > 0$. Now, if $j \to \infty$, then, by Proposition 2.3, we must have $h_j^\psi \to 0$ as $j \to \infty$ and hence, by construction in step 19, $h_j^f \to 0$ as $j \to \infty$. Consequently, $\varepsilon \searrow 0$ as $j \to \infty$, so that there exists j_0 such that $\varepsilon \leqq \varepsilon^1(x_N)$ for all $j \geqq j_0$. But then, for all $j \geqq j_0$ we must have that $\|h_j^\psi\|^2 \geqq h^\psi_{\varepsilon(x_N)}(x_N)\|^2$ and $\|h_j^f\|^2 \geqq \|h^f_{\varepsilon(x_N)}(x_N)\|^2$. Consequently, $\max\{\|\Gamma(x_N)h_j^f\|^2, \|(1-\Gamma(x_N))h_j^\psi\|^2\} \geqq -\theta^1_{\varepsilon(x_N)}(x_N) \geqq \delta\varepsilon(x_N)$ which contradicts the conclusion that $\varepsilon \searrow 0$.

We have thus shown that the algorithm cannot jam up at a point x_N such that $\psi(x_n) > 0$ or $\psi(x_N) \leqq 0$ and $0 \notin \text{co}\{\partial f(x_N) \cup \partial_0^+\psi(x_N)\}$.

b) Suppose that the sequence $\{x_i\}$ is infinite and that $x_i \to^K \hat{x}$, with $K \subset \{0, 1, 2, \cdots\}$ and either $\psi(\hat{x}) > 0$ or $\psi(\hat{x}) \leqq 0$ and $0 \notin \text{co}\{\partial f(\hat{x}) \cup \partial_0^+\psi(x)\}$.

Case 1. $\psi(x_i) > 0$ for all i. In this case $\psi(\hat{x}) \geqq 0$ and $\varepsilon^1(\hat{x}) > 0$. By Lemma 3.2, there exists an i_0 such that $\varepsilon^1(x_i) \geqq \nu\varepsilon^1(\hat{x}) > 0$ for all $i \geqq i_0$, $i \in K$. Consequently, since the test value of ε in the implementable algorithm is always greater than or equal to that in the conceptual algorithm, we must have, by Lemma 3.1, that $\|h_i\|^2 \geqq \delta\nu\varepsilon^1(\hat{x}) > 0$ for all $i \geqq i_0$, $i \in K$. Also, there exists a $b < \infty$ such that $\|h_i\| \leqq b$ for all $i \in K$. Consequently, in (4.7a), for all $i \in K$, $i \geqq i_0$ and $j = 0, 1, 2, \cdots$, we must have $s_j \geqq \beta\varepsilon^1(x_i) \geqq \beta\nu^2\varepsilon^1(\hat{x})/b$. Hence, by (4.9)

$$\psi(x_{i+1}) - \psi(x_i) \leqq -[\beta\nu\varepsilon^1(\hat{x})/b]\delta\alpha\nu\varepsilon^1(\hat{x}) = -\delta\alpha\beta\nu^2\varepsilon^1(\hat{x})^2/b < 0 \tag{4.10}$$

for all $i \in K$, $i \geqq i_0$. However, by continuity, $\psi(x_i) \to^K \psi(\hat{x})$ and hence, since $\psi(x_i) \searrow$, we must have that $\psi(x_i) \to \psi(\hat{x})$. But this is contradicted by (4.10) and hence the theorem is proved for the case where $\psi(x_i) > 0$ for all i.

Case 2. There exists an i_0 such that $\psi(x_{i_0}) \leqq 0$. Then, by construction, we must have $\psi(x_i) \leqq 0$ for all $i \geqq i_0$. Now suppose that $x_i \to^K \hat{x}$, $K \subset \{0, 1, 2, \cdots\}$, with $\psi(\hat{x}) \leqq 0$ and $0 \notin \text{co}\{\partial f(\hat{x}) \cup \partial_0^+\psi(\hat{x})\}$. Then $\varepsilon^1(\hat{x}) > 0$ and there exists an $i_1 \geqq i_0$ such that $\varepsilon^1(x_i) \geqq \nu\varepsilon^1(\hat{x}) > 0$ for all $i \in K$, $i \geqq i_1$. Consequently, with $b = \sup\{\|h_i\| \mid i \in K\} < \infty$, we have once more that $s_j \geqq \nu\varepsilon^1(x_i)/b \geqq \beta\nu\varepsilon^1(\hat{x})/b$ for all $i \in K$, $i \geqq i_0$, and $\|h_i\|^2 \geqq \|h^f_{\varepsilon^1(x_i)}(x_i)\|^2 \geqq \delta\varepsilon^1(x_i) \geqq \delta\nu\varepsilon^1(\hat{x})$ for all $i \in K$, $i \geqq i_0$. It now follows from (4.3a) and (4.5a) that

$$f(x_{i+1}) - f(x_i) \leqq -\alpha\beta\nu^2\varepsilon^1(\hat{x})^2/b \tag{4.11}$$

for all $i \in K$, $i \geqq i$. Now, $f(x_i) \to^K f(\hat{x})$ by continuity and $f(x_i)$, hence $f(x_i) \to f(\hat{x})$. But this is contradicted by (4.11) and hence the proof is complete. □

5. Conclusion. We have presented in this paper a systematic approach to the extension of differentiable optimization algorithms to algorithms for the solution of nondifferentiable optimization problems. We have assumed very little structure in the functions defining the problem: in particular we have implicitly assumed that there is no easy way of determining that one is near a point of nondifferentiability. However, there is an important class of problems, (see e.g. the problems with eigenvalue constraints in [C6], [P16] and the semi-infinite problems in [G3]) where a clear and simple indication of approaching nondifferentiability is obtained by a trivial calculation. For these problems one can develop much more efficient algorithms than the ones discussed in this paper, but following a very similar approach, as we see in [P16].

There remain a number of open questions to be dealt with: a) What is the rate of convergence of these new nondifferentiable optimization algorithms? b) How can

one extend second order optimization algorithms to the nondifferentiable case? and c) Do outer approximation methods, such as those described in [G2], offer a preferable alternative to the ones discussed in this paper? The last question is particularly interesting, in view of [M4] and [P15] where we see that two extremely complex nondifferentiable optimization problems can be decomposed into infinite sequences of differentiable problems.

We hope that this paper and the questions we raised will stimulate further research in nondifferentiable optimization algorithms.

Acknowledgment. We wish to thank Prof. F. Clarke for supplying us with a proof for Theorem 1.1. (He has subsequently proved this result without requiring that the set $\{x|F(x)=0\}$ have measure zero; see [C4].)

REFERENCES

[A1] L. ARMIJO, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[B1] C. BERGE, *Topological Spaces*, Macmillan, New York, 1963.

[B2] D. P. BERTSEKAS AND S. K. MITTER, *A descent numerical method for optimization problems with nondifferentiable cost functionals*, this journal; 11 (1973), pp. 637–652.

[C1] F. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[C2] ———, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.

[C3] ———, *Optimal control and the true Hamiltonian*, SIAM Rev., 21 (1979), pp. 157–167.

[C4] ———, *Nonsmooth Analysis and Optimization*, Wiley-Interscience, New York, 1983.

[C5] A. R. CONN, *Constrained optimization using a nondifferentiable penalty function*, SIAM J. Numer. Anal., 10 (1973), pp. 760–784.

[C6] J. CULUM, W. E. DONATH AND P. WOLFE, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, in Mathematical Programming Studies 3, M. L. Balinski and R. Wolfe, eds., North-Holland, Amsterdam, 1975, pp. 35–55.

[D1] J. M. DANSKIN, *The theory of maxmin with applications*, SIAM J. Appl. Math., 14 (1966), pp. 641–655.

[D2] V. F. DEMJANOV, *Algorithms for some minimax problems*, J. Comput. System Sci., 2 (1968), pp. 342–380.

[D3] V. F. DEMJANOV AND V. N. MALOZEMOV, *Introduction to Minimax*, John Wiley, New York, 1974.

[D4] V. F. DEMJANOV, *Differentiability of a maxmin function*, I, U.S.S.R. Computational Math. and Math. Phys. (1968), pp. 1–15.

[F1] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[F2] R. FLETCHER, *An exact penalty function method for nonlinear programming with inequalities*, Math Programming, 5 (1973), pp. 129–150.

[G1] A. A. GOLDSTEIN, *Optimization of Lipschitz continuous functions*, Math. Programming, 13 (1977), pp. 14–22.

[G2] C. GONZAGA AND E. POLAK, *On constraints dropping schemes and optimality functions for a class of outer approximations algorithms*, this journal, 17 (1979), pp. 477–493.

[G3] C. GONZAGA, E. POLAK AND R. TRAHAN, *An improved algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automat. Control, AC-25 (1979), pp. 49–54.

[L1] G. LEBOURG, *Valeur moyenne pour gradient généralisé*, C. R. Acad. Sci. Paris, 281 (1975), pp. 795–797.

[L2] C. LEMARECHAL, *Extensions divérses des méthodes de gradient et applications*, Thesis, Univ. of Paris VIII, 1980.

[L3] ———, *Non smooth optimization: Toward a synthesis*, Abstracts journées de l'optimisation, Montréal, 1978, pp. 69–70.

[L4] ———, *Minimization of nondifferentiable functions with constraints*, Proc. 12th Allerton Conference on Circuit Theory, Univ. of Illinois, Urbana, 1974, pp. 16–24.

[L5] ———, *Nondifferentiable optimization, subgradient and ε subgradient methods*, in Optimization and Operations Research, Lecture Notes in Mathematics 117, Springer-Verlag, New York, 1976.

[L6] ———, *An extension of Davidon methods to nondifferential problems*, in Mathematical Programming Studies 3, M. L. Balinski and R. Wolfe, eds., North-Holland, Amsterdam, 1975, pp. 95–109.

[M1] R. MIFFLIN, *Semismooth and semiconvex functions in constrained optimization*, this Journal, 15 (1977), pp. 959–972.

[M2] D. Q. MAYNE, E. POLAK AND R. TRAHAN, *On outer approximation algorithms for computer aided design problems*, J. Optim. Theory Appl., 28 (1979), pp. 331–352.

[M3] D. Q. MAYNE AND E. POLAK, *Feasible directions algorithms for optimization problems with equality and inequality constraints*, Math. Programming, 11 (1976), pp. 67–80.

[M4] ———, *Algorithms for the design of control systems subject to singular value inequalities*, presented at NSF workshop on Numerical Methods in Engineering, Lexington, KY, June 12–16, 1980, Math. Programming Stud., to appear.

[P1] O. PIRONNEAU AND E. POLAK, *On the rate of convergence of certain methods of centers*, Math. Programming, 2 (1972), pp. 230–257.

[P2] E. POLAK AND D. Q. MAYNE, *On the solution of singular value inequalities over a continuum of frequencies*, IEEE Trans. Automat Control, AC-26 (1981), pp. 690–695.

[P3] E. POLAK, R. TRAHAN AND D. Q. MAYNE, *Combined phase-I–phase-II methods of feasible directions*, Math. Programming, 17 (1979), pp. 61–73.

[P4] E. POLAK AND D. Q. MAYNE, *An algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 184–193.

[P5] E. POLAK, *On a class of computer aided design problems*, in A Link Between Science and Applications of Automatic Control, A. Niemi, ed., Pergamon Press, Oxford, New York, 1979.

[P6] E. POLAK AND R. TRAHAN, *An algorithm for computer aided design of control problems*, Proc. IEEE Conf. on Decision and Control, 1976.

[P7] E. POLAK AND A. SANGIOVANNI-VINCENTELLI, *Theoretical and computational aspects of optimal design centering, tolerancing and tuning problems*, IEEE Trans. Automat. Control, AS-26 (1979), 9, pp. 295–318.

[P8] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1971.

[P9] E. POLAK, R. W. H. SARGENT AND D. J. SEBASTIAN, *On the convergence of sequential minimization algorithms*, J. Optim. Theory. Appl., 14 (1974), pp. 439–442.

[P10] E. POLAK, *On the global stabilization of locally convergent algorithms*, Automatica, 12 (1976), pp. 337–342.

[P11] B. T. POLJAK, *A general method for solving extremum problems*, Soviet Math. Dokl., 8 (1966), pp. 593–597.

[P12] ———, *Minimization of nonsmooth functionals*, U.S.S.R. Computational Math. and Math. Phys., 9 (1969), pp. 14–29.

[P13] B. N. PSHENICHNYI, *Necessary Conditions for an Extremum* (translation edited by L. W. Neustadt) K. Makovski, transl., Marcel Dekker Inc., New York, 1971.

[P14] T. PIETRZYKOWSKI, *The potential method for conditional maxima in the locally compact metric spaces*, Numer. Math., 14 (1970), pp. 325–329.

[P15] E. POLAK, *An implementable algorithm for the design centering, tolerancing and tuning problem*, in Computing Methods in Applied Science and Engineering, R. Glowinski and J. L. Lions, eds., North-Holland, Amsterdam, 1980, pp. 499–517.

[P16] E. POLAK AND Y. WARDI, *A nondifferentiable optimization algorithm for the design of control systems subject to singular value inequalities over a frequency range*, Univ. of California Electronics Research Lab. Memo. UCB/ERL M80/41, 1981, to appear in Automatica.

[R1] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Pr., Princeton, N.J., 1970.

[R2] W. RUDIN, *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1964.

[S1] N. Z. SHOR, *Utilization of the operation of space dilatation in the minimization of convex functions*, Cybernetics, 1 (1970), pp. 7–15.

[S2] N. Z. SHOR AND L. P. SHABASHOVA, *Solution of minmax problems by the method of generalized gradient descent with dilatation of space*, Cybernetics, 1 (1972), pp. 88–94.

[T1] S. TISHYADHIGAMA, E. POLAK AND R. KLESSIG, *A comparative study of several general convergence conditions for algorithms modeled by point-to-set maps*, Math. Programming Study, 10, North-Holland, Amsterdam, 1979, pp. 172–190.

[W1] P. WOLFE, *Finding the nearest point in a polytope*, Math. Programming , 11, (1976), pp. 128–149.

[W2] ———, *A method of conjugate subgradients for minimizing nondifferentiable functions*, in Nondifferentiable Optimization, M. L. Balinski and P. Wolfe, eds., Math. Programming Study 3, North-Holland, Amsterdam, 1975, pp. 145–173.

[Z1] W. I. ZANGWILL, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

[Z2] G. ZOUTENDIJK, *Methods of Feasible Directions*, Elsevier, Amsterdam, 1960.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0003 $01.25/0

PIECEWISE-LINEAR HOMOTOPY ALGORITHMS FOR SPARSE SYSTEMS OF NONLINEAR EQUATIONS*

MICHAEL J. TODD†

Abstract. When piecewise-linear homotopy algorithms are applied to the problem of approximating a zero of a sparse function $f: R^n \to R^n$, a large piece of linearity can be traversed in one step by using a suitable linear system. The linear system has n rows and $n+1$ columns, but is subject to a number of inequalities depending on the sparsity pattern of f. We show how an algorithm can be implemented using these large pieces; in particular, we demonstrate how to update the linear system corresponding to one large piece to obtain the appropriate system for an adjacent large piece. One measure of the complexity of such an implementation is the number of inequalities that may be required for any one piece. We prove that there can be no more than $O(n^{3/2})$ such inequalities, and that this bound is essentially tight; the argument is purely combinatorial. Finally, we provide guidelines on when such a "large-piece implementation" should be used instead of much simpler "small-piece implementations" for piecewise-linear homotopy algorithms.

Key words. piecewise-linear homotopy algorithms, solving nonlinear equations, sparsity

1. Introduction. We are concerned with piecewise-linear homotopy algorithms for approximating a zero of a function $f: R^n \to R^n$; see [1], [2], [4], [9]. Our interest is in the case where f is sparse, i.e., each component function of f depends on only a few components of the argument. As in [11], we confine ourselves to Merrill's restart algorithm [7], though our approach can also be applied to other such algorithms, in particular those of Eaves–Saigal [3] and van der Laan–Talman [5], [6]. Merrill's algorithm starts by choosing a simple function $f^0: R^n \to R^n$—we will always take $f^0(x) = G(x - x^0)$ where G is $n \times n$ and nonsingular—and defining $h: R^n \times [0, 1] \to R^n$ by $h(x, t) = tf(x) + (1-t)f^0(x)$. Next we choose a triangulation T of $R^n \times [0, 1]$ and let l be the piecewise-linear approximation to h with respect to T. One major cycle of the algorithm generates a sequence of simplices of T that contains the piecewise-linear path that is the connected component of $l^{-1}(0)$ containing $(x^0, 0)$ (perturbation may be necessary). If this sequence is finite, we obtain a point $(x^1, 1) \in l^{-1}(0)$; either x^1 is accepted as an approximate zero or another major cycle is initiated after updating T and f^0.

With each major cycle, each simplex that meets $l^{-1}(0)$ is associated with a certain linear system with $n+1$ rows and $n+2$ columns subject to $n+2$ nonnegativities. Moving from one simplex to its successor requires the evaluation of f or f^0 at the new vertex and a linear programming pivot step in the linear system.

In [10] we showed that, because of the linearity of f^0, the pieces of linearity of l were much larger than individual simplices for commonly used triangulations T; moreover, if f enjoyed special structure, these pieces were larger still. However, [10] proposed only a "local" method of exploiting this property: whenever it was determined that the next simplex lay in the same piece of linearity as the current one, one function evaluation was saved and the linear programming pivot step could be executed trivially. Later, we derived in [11] linear systems that allowed pieces of linearity of l to be traversed in one step, when f was general, separable or partially separable. These systems again had $n+1$ rows and $n+2$ columns, with up to $2n+1$ inequalities.

* Received by the editors January 19, 1981, and in revised form November 27, 1981. This research was partially supported by a Guggenheim Fellowship, and by the National Science Foundation under grant ECS-7921279.

† School of Operations Research and Industrial Engineering, College of Engineering, Cornell University, Ithaca, New York 14853.

However, the important case of a sparse function was not treated in [11]. Finally, [12] showed how these accelerated methods could use basis factorizations rather than inverses for numerical stability and preservation of sparsity and suggested that for the case of a sparse function f, the global idea of [11] for general f could be combined with the local method of [10] for sparsity.

In this paper we give in § 2 a linear system that allows the whole piece of linearity of l to be traversed in one step when f is sparse. This linear system has n rows and $n+1$ columns. However, the number of inequalities depends on the sparsity structure of f. Section 3 indicates how a piecewise-linear homotopy algorithm can be implemented using these linear systems. One measure of the complexity of such an implementation is the number of inequalities that may be required for any one piece. Section 4 shows that there can be no more than $O(n^{3/2})$ such inequalities, and gives an example demonstrating that $O(n^{3/2})$ may be necessary. Because of the large number of inequalities possible, it may not always be advisable to use this linear system. We discuss this issue in § 5. We conclude that if f is very sparse, use of the new system is worthwhile, and describe an alternative method of implementation for use in other cases.

Since this paper was written, Saigal [8] has proposed another piecewise-linear homotopy algorithm that exploits sparsity.

2. The large pieces. In this section we describe the pieces of linearity of the homotopy l when f is sparse and show how, when such a piece $\check{\sigma}$ meets $l^{-1}(0)$ in a line segment, this line segment can be traversed by considering a linear system of the form $Aw=b$, $Cw \geqq d$. Here A, b, C and d depend on the piece $\check{\sigma}$. A has n rows and $n+1$ columns, and the solutions to $Aw=b$ form a line in R^{n+1} whose intersection with $\check{\sigma}=\{w \in R^{n+1}: Cw \geqq d\}$ is the desired line segment. The matrix C has $n+1$ columns, but the number of rows (equal to the number of facets of the piece $\check{\sigma}$) depends on the sparsity pattern of f. We trace the segment numerically by generating a particular solution $\bar{w}$ to $Aw=b$, $Cw \geqq d$ (corresponding to where the piece $\check{\sigma}$ is entered) and a vector z in the null space of A. Then $\{w: Aw=b\}=\{\bar{w}+\lambda z\}$. Then we find where the segment leaves the piece by finding the range of λ for which $C(\bar{w}+\lambda z) \geqq d$; this corresponds to a minimum ratio test in linear programming. While each inequality in $Cw \geqq d$ is very simple (involving at most two components of w) the complexity of the algorithm clearly depends on the number of rows of C.

To make this section self-contained, we start by describing the triangulation $\tilde{J}_1=\{j_1(v, \pi, s)\}$ of $R^n \times [0, 1]$ on which our large pieces are based. It was shown in [10], [11] that this triangulation induces large pieces of linearity for several types of structure. We suppress dependence throughout on the grid size ε. Each individual simplex $\sigma \equiv j_1(v, \pi, s)$ depends on the starting vertex $v \in R^n \times \{1\}$, the permutation π of $\{1, 2, \cdots, n+1\}$ and the vector $s \in R^n \times \{-1\}$, where each component v_i of v is an odd multiple of ε for $i=1, 2, \cdots, n$ and each component of s is ± 1. Define j by $\pi(j)=n+1$ and let e^p denote the pth unit vector of appropriate dimension. Then the vertices of σ are $v^0, v^1, \cdots, v^{n+1}$, where

$$\begin{aligned} &v^0=v, \qquad v^i=v^{i-1}+\varepsilon s_{\pi(i)} e^{\pi(i)}, \qquad 1 \leqq i<j, \\ &v^j=v^{j-1}-e^{n+1}, \qquad v^k=v^{k-1}+\varepsilon s_{\pi(k)} e^{\pi(k)}, \qquad j<k \leqq n+1. \end{aligned} \tag{1}$$

Alternatively, σ can be described by its facets: it is the set of all $w \in R^{n+1}$ satisfying

$$\begin{aligned} &\varepsilon \geqq s_{\pi(1)}(w_{\pi(1)}-v_{\pi(1)}) \geqq \cdots \geqq s_{\pi(j-1)}(w_{\pi(j-1)}-v_{\pi(j-1)}) \geqq \varepsilon(1-w_{n+1}) \\ &\geqq s_{\pi(j+1)}(w_{\pi(j+1)}-v_{\pi(j+1)}) \geqq \cdots \geqq s_{\pi(n+1)}(w_{\pi(n+1)}-v_{\pi(n+1)}) \geqq 0. \end{aligned} \tag{2}$$

Henceforth, l is the piecewise-linear approximation to h with respect to $\tilde{J}_1$. However, because f^0 is affine—recall that we chose $f^0(x) = G(x - x^0)$—even for general f the pieces of linearity of l are larger than the simplices of $\tilde{J}_1$. In [11] we showed that these pieces formed a (polyhedral) subdivision $\hat{J}_1 = \{\hat{j}_1(v, \pi, s)\}$ of $R^n \times [0, 1]$. The individual piece $\hat{\sigma} \equiv \hat{j}_1(v, \pi, s)$ contains the simplex $j_1(v, \pi, s)$ and several other simplices; thus several triples (v, π, s) yield the same piece $\hat{\sigma}$. Suppose again that $\pi(j) = n+1$. Then $\hat{\sigma}$ is the set of all $w \in R^{n+1}$ satisfying

$$
\begin{aligned}
&\varepsilon \geqq s_{\pi(1)}(w_{\pi(1)} - v_{\pi(1)}) \geqq \cdots \geqq s_{\pi(j-1)}(w_{\pi(j-1)} - v_{\pi(j-1)}) \geqq \varepsilon(1 - w_{n+1}), \\
&\varepsilon(1 - w_{n+1}) \geqq |w_{\pi(k)} - v_{\pi(k)}|, \qquad j < k \leqq n+1, \\
&\varepsilon(1 - w_{n+1}) \geqq 0 \quad \text{if } j = n+1.
\end{aligned} \tag{3}
$$

Note that $\hat{\sigma}$ has $2n - j + 2$ facets if $j \leqq n$, and $n+2$ if $j = n+1$.

Since we will be much concerned with systems of inequalities similar to those in (3), we now introduce some very useful notation. The inequalities in (3) relate certain fundamental affine functions of w. Suppressing the dependence on v and s, we denote these by

$$
\begin{aligned}
&\gamma^0 w - \delta_0 \equiv \varepsilon, \\
&\gamma^p w - \delta_p \equiv s_p(w_p - v_p) \quad \text{for } 1 \leqq p \leqq n, \\
&\gamma^{n+1} w - \delta_{n+1} \equiv \varepsilon(1 - w_{n+1}), \\
&\gamma^{n+2} w - \delta_{n+2} \equiv -\varepsilon(1 - w_{n+1}).
\end{aligned}
$$

For $p, q \in N_+ \equiv \{0, 1, \cdots, n+1, n+2\}$ define $c^{pq} = \gamma^p - \gamma^q$ and $d_{pq} = \delta_p - \delta_q$, and note that each inequality of (3) is of the form $c^{pq} w \geqq d_{pq}$ for certain p, q. The appropriate pairs (p, q) are most easily identified by defining a subset Γ'_π of $N_+ \times N_+$. It is convenient to identify such subsets $\Gamma \subseteq N_+ \times N_+$ with the corresponding directed graphs (N_+, Γ). Then Γ'_π consists of a path from 0 through $\pi(1), \pi(2), \cdots, \pi(j-1)$ to $n+1$, together with edges $(n+1, \pi(k))$ and $(\pi(k), n+2)$ for $j < k \leqq n+1$; if $j = n+1$, Γ'_π also contains the edge $(n+1, n+2)$.

For any directed graph Γ on N_+, we define the $|\Gamma| \times (n+1)$ matrix $C(\Gamma)$ to have as rows the vectors c^{pq} corresponding to $(p, q) \in \Gamma$, and the $|\Gamma|$-vector $d(\Gamma)$ similarly. It is then easy to verify that the inequalities (3) can be rewritten as $C(\Gamma'_\pi) w \geqq d(\Gamma'_\pi)$.

Note that the same piece $\hat{\sigma}$ is defined by $C(\Gamma_\pi) w \geqq d(\Gamma_\pi)$, where, in addition to the edges in Γ'_π, Γ_π contains $(0, \pi(i))$ and $(\pi(i), n+1)$ for all $1 \leqq i < j$ and $(\pi(i), \pi(i'))$ for all $1 \leqq i < i' < j$. Clearly this new system contains a number of redundant inequalities, and the unique minimal set is as given in (3), with at most $2n+1$ inequalities. Summarizing, we have

LEMMA 1. *The piece $\hat{\sigma}$ is exactly $\{w \in R^{n+1}: Cw \geqq d\}$, where either $C = C(\Gamma'_\pi)$ and $d = d(\Gamma'_\pi)$ or $C = C(\Gamma_\pi)$ and $d = d(\Gamma_\pi)$.*

Let $A_{\hat{\sigma}}$ denote the derivative matrix of the affine function from R^{n+1} to R^n that agrees with l on $\hat{\sigma}$. Then clearly, for any $w, \bar{w} \in \hat{\sigma}$,

$$
l(w) = l(\bar{w}) + A_{\hat{\sigma}}(w - \bar{w}). \tag{4}
$$

Note that, because of the form of $\hat{\sigma}$, $A_{\hat{\sigma}}$ can be obtained very simply from the function values of f and f^0 at the vertices of $\hat{\sigma}$. Let us write y^i for the projection of $v^i \in R^n \times [0, 1]$ on R^n, so that $l(v^i) = f(y^i)$ for $0 \leqq i < j$, and $y^j = y^{j-1}$. If a^i denotes the ith column of

$A_{\hat{\sigma}}$, we find

$$
\begin{aligned}
& a^{\pi(i)} = (f(y^i) - f(y^{i-1}))/\varepsilon s_{\pi(i)}, \qquad 1 \leqq i < j, \\
& a^{\pi(k)} = g^{\pi(k)}, \qquad j < k \leqq n+1, \\
& a^{n+1} = f(y^j) - f^0(y^j),
\end{aligned} \tag{5}
$$

with g^k the kth column of the matrix G used to define f^0. From Lemma 1, (4) and (5) we obtain:

THEOREM 1. *The set of w lying in $l^{-1}(0) \cap \hat{\sigma}$ is the set of solutions to $Aw = b$, $Cw \geqq d$, where $A = A_{\hat{\sigma}}$ is given by* (5), *$b = A\bar{w} - l(\bar{w})$ for any $\bar{w} \in \hat{\sigma}$ and $C = C(\Gamma'_\pi)$, $d = d(\Gamma'_\pi)$.*

Now we introduce the sparsity of f. We say that coordinates p and q interact if there is some component of f that depends on both x_p and x_q. If f is differentiable, p and q interact if the pth and qth columns of the derivative matrix $Df(x)$ have nonzeroes in some common row. Naturally, we choose f^0 to have the same sparsity pattern as f. Note that, except for its final dense column, the matrix $A_{\hat{\sigma}}$ has the same sparsity pattern as G and $Df(x)$.

Suppose $p = \pi(i)$ and $q = \pi(i+1)$, $1 \leqq i < j-1$, do not interact. Then if $\tilde{\pi}$ denotes π with the positions of p and q interchanged, we find that $\tilde{\sigma} = \hat{j}_1(v, \tilde{\pi}, s)$ differs only slightly from $\hat{\sigma}$; indeed, just one vertex v^i of $\hat{\sigma}$ changes, and $A_{\tilde{\sigma}}$ coincides with $A_{\hat{\sigma}}$ except possibly in its pth and qth columns, which become

$$
\frac{f(y^i + \varepsilon s_q e^q) - f(y^{i-1} + \varepsilon s_q e^q)}{\varepsilon s_p} \quad \text{and} \quad \frac{f(y^{i+1} - \varepsilon s_p e^p) - f(y^i - \varepsilon s_p e^p)}{\varepsilon s_q},
$$

respectively. But because p and q do not interact, these vectors are precisely the same if the terms $+\varepsilon s_q e^q$ and $-\varepsilon s_p e^p$ are deleted everywhere. Thus $A_{\tilde{\sigma}} = A_{\hat{\sigma}}$. It follows that l is linear in $\hat{\sigma} \cup \tilde{\sigma}$. Continuing in this way we may obtain a much larger piece of linearity than $\hat{\sigma}$ if f is sparse. To describe this piece, which we denote $\check{\sigma} = \check{j}_1(v, \pi, s)$, we define certain subgraphs of Γ_π. Let Σ (corresponding to sparsity) be the graph consisting of all edges $(0, p)$, $(p, n+1)$, $(n+1, p)$ and $(p, n+2)$ for $1 \leqq p \leqq n$, as well as $(0, n+1)$ and $(n+1, n+2)$, together with edges (p, q) and (q, p) for $1 \leqq p, q \leqq n$ if p and q interact.

Next let $\Delta_\pi = \Gamma_\pi \cap \Sigma$. We can then define $\check{\sigma}$ as $\{w \in R^{n+1}: Cw \geqq d\}$ for $C = C(\Delta_\pi)$, $d = d(\Delta_\pi)$. However, just as the system $C(\Gamma_\pi)w \geqq d(\Gamma_\pi)$ contained many superfluous inequalities, so does this new system. We therefore define the "minimum cover" of Δ_π as $\Delta'_\pi = \{(p, r) \in \Delta_\pi$: there is no q with $(p, q) \in \Delta_\pi$ and $(q, r) \in \Delta_\pi\}$. By considering a piecewise-linear path joining any two of the points in $\check{\sigma}$ we can then easily prove:

LEMMA 2. *l is linear on the polyhedron $\check{\sigma}$, which is the set of all $w \in R^{n+1}$ satisfying $Cw \geqq d$, where $C = C(\Delta'_\pi)$ and $d = d(\Delta'_\pi)$ or $C = C(\Delta_\pi)$ and $d = d(\Delta_\pi)$.*

While Δ_π is always smaller than Γ_π, the same is not always true of Δ'_π with respect to Γ'_π. This implies that the number of inequalities defining $\check{\sigma}$, or, geometrically, the number of its facets, may be larger than $2n+1$. We shall investigate this further in § 4. However, we immediately have:

THEOREM 2. *$\check{\sigma} \cap l^{-1}(0)$ is $\{w \in R^{n+1}: Aw = b$ and $Cw \geqq d\}$, where A and b are as in Theorem* 1, *$C = C(\Delta'_\pi)$, and $d = d(\Delta'_\pi)$.*

In the next section, we discuss how a piecewise-linear homotopy algorithm using the large pieces $\check{\sigma}$ could be implemented.

3. Implementation. We have already given, at the beginning of § 2, an outline of how a result such as Theorem 2 can be used to traverse a given piece $\check{\sigma}$. Here we

are concerned with some details of an implementation of a piecewise-linear homotopy algorithm using these large pieces.

First, we describe what must be stored in such an algorithm. We maintain a point $\bar{w}$ with $l(\bar{w}) = 0$, which is the point at which the current piece $\check{\sigma}$ was entered. We keep the vectors y (the projection of v onto R^n), $f(y)$ and s, the matrix $A = A_{\hat{\sigma}}$ and some representation of its null space. In order to generate new columns of A, we also store, for each p, a vector t^p of 0's and ± 1's such that, if $1 \leqq \pi^{-1}(p) < j$, $a^p = (f(y + \varepsilon t^p + \varepsilon s_p e^p) - f(y + \varepsilon t^p))/\varepsilon s_p$ and $a^{n+1} = f(y + \varepsilon t^{n+1}) - f^0(y + \varepsilon t^{n+1})$. Finally the graph Δ'_π is stored. Note that this information is not enough to recover the permutation π, nor even its first $j-1$ elements, although j can easily be obtained by counting the edges emanating from node $n+1$.

In order to be able to generate vectors in the null space of A, we will maintain an LU factorization of some permutation of the columns of A. Thus $L^{-1}AP = U = [\bar{U}, u]$, where L^{-1} is a product of permutations and lower triangular elementary matrices, P is a permutation matrix and U is upper triangular, with $\bar{U}$ nonsingular. Thus

$$z = P^T \begin{pmatrix} \bar{U}^{-1} u \\ -1 \end{pmatrix}$$

is in the null space of A.

The graph $\Delta = \Delta'_\pi$ is stored by maintaining with each $q \in N_+$ the set $P(q) = \{p \in N_+ : (p, q) \in \Delta\}$ of predecessors and the set $S(q) = \{r \in N_+ : (q, r) \in \Delta\}$ of successors of q. We also store $I = \{0, 1, 2, \cdots, n+1\} \backslash S(n+1)$ in a doubly-linked list so that if $(p, q) \in \Delta \cap (I \times I)$, p is before q in the list. By "add (delete) (p, q) from Δ" we mean make the appropriate changes to $S(p)$ and $P(q)$.

Initially, we have $\bar{w} = (x^0, 0)$, y_i the nearest odd integer multiple of ε to x_i^0 and s_i arbitrary for $i = 1, 2, \cdots, n$. We evaluate $f(y)$ and thus obtain $A = [G, f(y) - f^0(y)]$; if an $L\bar{U}$ factorization of G is available or computed, we obtain easily a corresponding LU factorization of A. We set each $t^p = 0$. The graph Δ'_π is the graph with edges $(0, n+1)$ and $(n+1, p)$, $(p, n+2)$ for $1 \leqq p \leqq n$.

Each iteration is performed as follows. We use the factorization of A to obtain a vector z in its null space. Ignoring degeneracy, we find that $\bar{w} + \lambda z$ lies in $\check{\sigma}$ for $0 \leqq \lambda \leqq \bar{\lambda}$, some $\bar{\lambda} > 0$, or for $\bar{\lambda} \leqq \lambda \leqq 0$, some $\bar{\lambda} < 0$. We will choose the sign of z so that the former case occurs. Then for λ just larger than $\bar{\lambda}$, $\bar{w} + \lambda z$ lies in a new piece $\check{\sigma}$ (or has last component greater than one) and we must update everything stored.

The critical value $\bar{\lambda}$ is found by examining the inequalities $C(\Delta)(\bar{w} + \lambda z) \geqq d(\Delta)$ and (again under a nondegeneracy assumption) exactly one of these inequalities will be tight for $\lambda = \bar{\lambda}$. (Otherwise we make a perturbation, replacing $\bar{w}$ by $\bar{w}_\varepsilon = \bar{w} + d_\varepsilon$, where $Ad_\varepsilon = (\varepsilon, \varepsilon^2, \cdots, \varepsilon^n)^T$ for small positive ε.) Each inequality corresponds to an edge of Δ, and we analyze the update by considering each possibility. In each case $\bar{w} \leftarrow \bar{\bar{w}} = \bar{w} + \bar{\lambda} z$.

Case 1. The critical edge (leading to the tight inequality) is $(n+1, n+2)$. Then $\bar{\bar{w}} = (x^1, 1)$ for some x^1, and a major cycle is completed. We also have an LU factorization of a finite difference approximation to $Df(x^1)$, or at least an "LH" factorization. (That is, $L^{-1}DP = H$ where $D \approx Df(x^1)$ and H is upper Hessenberg so that $h_{ij} = 0$ for $i > j+1$. It is easy to obtain from this an LU factorization of D. Alternatively, we can maintain the permutation P so that a^{n+1} always corresponds to the last or penultimate column of U, in which case an LU factorization of D is immediate.)

Case 2. The critical edge is $(n+1, p)$ for some $1 \leqq p \leqq n$. Define $y^{j-1} = y + \varepsilon \sum_{i \in I} s_i e^i$ and calculate $f(y^{j-1}) = f(y) + \varepsilon \sum_{i \in I} s_i a^i$. Then set $s_p \leftarrow 1$ and evaluate

$f(y^{j-1}+\varepsilon s_p e^p)$ and replace $a^p = g^p$ by $(f(y^{j-1}+\varepsilon s_p e^p)-f(y^{j-1}))/\varepsilon s_p$ and then a^{n+1} by $a^{n+1}+\varepsilon s_p(a^p-g^p)$. The LU factorization of A is updated using a standard technique; see, e.g., [12]. (We may, if desired, keep a^{n+1} corresponding to one of the last two columns of U; then the second replacement is trivial.) The vectors y and $f(y)$ are unchanged, while s is only possibly changed in its pth component; we set $t^p = \sum_I s_i e^i$ and $t^{n+1} = t^p + s_p e^p$. Finally, we update Δ (i.e. the doubly-linked list I and the sets $S(q)$, $P(q)$ for $q \in N_+$) as follows.

Delete $(n+1, p)$, $(p, n+2)$ and (if present) $(0, n+1)$ from Δ.
Add $(0, p)$, $(p, n+1)$, and, if $S(n+1) = \varnothing$, $(n+1, n+2)$ to Δ.
Set $B = \varnothing$. (B will represent the set of nodes in I with a path to p.)
For $i \in I$, working backwards through the list from $n+1$ to 0 (not including $n+1$ or 0):
i) if $S(i) \cap B \neq \varnothing$, $B \leftarrow B \cup \{i\}$;
ii) if $S(i) \cap B = \varnothing$ and $(i, p) \in \Sigma$, add (i, p) to Δ;
$B \leftarrow B \cup \{i\}$;
delete (if present)
$(i, n+1)$ and $(0, p)$ from Δ.
Set $I \leftarrow I \cup \{p\}$ and add p to the doubly-linked list just before $n+1$.

Case 3. The critical edge is $(p, n+2)$ for some $1 \leqq p \leqq n$. Proceed exactly as in case 2 but with $s_p \leftarrow -1$.

Case 4. The critical edge is $(p, n+1)$. This is the reverse of Cases 2 or 3. We replace a^{n+1} by $a^{n+1}-\varepsilon s_p(a^p-g^p)$ and a^p by g^p and update the factorization of A. The vectors y, $f(y)$ and s remain unchanged; we set $t^{n+1} \leftarrow t^{n+1} - s_p e^p$. Δ is updated as follows. Set $I \leftarrow I \backslash \{p\}$ and remove p from the doubly-linked list. For $i \in P(p)$, delete (i, p) from Δ; if $S(i) = \{p\}$, add $(i, n+1)$ to Δ. Remove $(p, n+1)$ and add $(n+1, p)$ and $(p, n+2)$. If $S(0) = \varnothing$, add $(0, n+1)$ to Δ; if $(n+1, n+2)$ was present, delete it from Δ.

Case 5. The critical edge is $(0, p)$ for some $1 \leqq p \leqq n$. Then set $f_{\text{old}} \leftarrow f(y)$, $y \leftarrow y + 2\varepsilon s_p e^p$ and $s_p \leftarrow -s_p$. Evaluate $f(y)$ and set $a^p \leftarrow a^p + (f(y) - f_{\text{old}})/\varepsilon s_p$; update the factorization of A. For each q, set $t^q \leftarrow t^q - 2t_p^q e^p$. Δ is unchanged.

Case 6. The critical edge is (p, q) for some $1 \leqq p, q \leqq n$. Calculate $f(y+\varepsilon t^p) = f(y) + \varepsilon \bar{A} t^p$ ($\bar{A}$ is A with its final column a^{n+1} removed). Evaluate $f(y+\varepsilon t^p+\varepsilon s_q e^q)$. Set $a^q_{\text{old}} \leftarrow a^q$, $a^q \leftarrow (f(y+\varepsilon t^p+\varepsilon s_q e^q) - f(y+\varepsilon t^p))/\varepsilon s_q$ and $a^p \leftarrow a^p + (a^q_{\text{old}} - a^q)s_q/s_p$. Update the factorization of A. Set $t^q \leftarrow t^p$ and $t^p \leftarrow t^p + s_q e^q$. The vectors y, $f(y)$ and s are unchanged. Finally update Δ as follows.

1) Set $B_q \leftarrow \{q\}$. (B_q will represent the nodes before q, and not because of p.)
For $i \in I$, working backwards through the list from q to p (not including q or p)
if $S(i) \cap B_q \neq \varnothing$, $B_q \leftarrow B_q \cup \{i\}$.
Set $B_p \leftarrow \{p\}$. (B_p will represent the nodes before p and not necessarily before q.)
For $i \in I$, working backwards from p to 0 (not including p or 0)
if $S(i) \cap B_q \neq \varnothing$, $B_q \leftarrow B_q \cup \{i\}$,
if $p \in S(i)$, delete (i, p) from Δ;
else if $S(i) \cap B_p \neq \varnothing$ and $(i, q) \in \Sigma$
$B_q \leftarrow B_q \cup \{i\}$,
add (i, q) to Δ,
if $p \in S(i)$, delete (i, p) from Δ;
else if $S(i) \cap B_p \neq \varnothing$, $B_p \leftarrow B_p \cup \{i\}$.

Add (q, p) to and delete (p, q) from Δ.
If $(0, p)$ is present in Δ, delete it.
If $P(q) = \varnothing$, add $(0, q)$ to Δ.

2) Set $A_p \leftarrow \{p\}$. (A_p will represent the nodes after p, and not because of q.)
For $i \in I$, working forwards through the list from p to q (not including p or q)
if $P(i) \cap A_p \neq \varnothing$, $A_p \leftarrow A_p \cup \{i\}$.
Set $A_q \leftarrow \{q\}$. (A_q will represent the nodes after q and not necessarily after p.)
For $i \in I$, working forwards through the list from q to $n+1$ (not including q or $n+1$)
if $P(i) \cap A_p \neq \varnothing$, $A_p \leftarrow A_p \cup \{i\}$,
if $q \in P(i)$, delete (q, i) from Δ;
else if $P(i) \cap A_q \neq \varnothing$ and $(p, i) \in \Sigma$,
$A_p \leftarrow A_p \cup \{i\}$,
add (p, i) to Δ,
if $q \in P(i)$, delete (q, i) from Δ;
else if $P(i) \cap A_q \neq \varnothing$, $A_q \leftarrow A_q \cup \{i\}$.
If $(q, n+1)$ is present in Δ, delete it.
If $S(p) = \varnothing$, add $(p, n+1)$ to Δ.

This concludes our discussion of the implementation of the large-piece algorithm. Note that, in each case, updating the factorization of A requires the replacement of one or two columns of A and U: but that only in the final case are two general column exchanges required. The sparsity of A and its factors is likely to compensate for the possible increase in work over other implementations requiring only single column exchanges. Note also that the sparsity information in the graph Δ is used in performing the minimum ration test in determining $\bar{\lambda}$. The work involved in this test, and also in the rather complicated updating of Δ, is proportional to the cardinality of Δ, which we study in the next section.

4. A bound on the number of facets of a large piece $\check{\sigma}$. We have seen that the complexity of our large-piece implementation depends on the number of facets of such a piece $\check{\sigma}$, or equivalently on the number of edges of a graph Δ'_π. This section investigates how large this number can be. We can assume that $\pi = (1, 2, \cdots, j-1, n+1, j, \cdots, n)$ for some $1 \leqq j \leqq n+1$. Now consider the following combinatorial problem.

Let $m = j+1$ and let $S_1, S_2, \cdots, S_m$ be arbitrary subsets of $N = \{1, 2, \cdots, n\}$. Write $p < q$ if $1 \leqq p < q \leqq m$ and $S_p \cap S_q \neq \varnothing$. We wish to find an upper bound $\phi(m, n)$ on the cardinality of

$$E = \{(p, r): p < r \text{ and for no } q \text{ is } p < q < r\}. \tag{6}$$

The relationship with bounding $|\Delta'_\pi|$ is as follows. Define $S_1 = S_m = N$ and $S_{i+1} = \{p : f_p(x) \text{ depends on } x_i\}$, $1 \leqq i < j$. Then $c^{pq}w > d_{pq}$ is a row of $C(\Delta_\pi)w > d(\Delta_\pi)$ with $\min\{p, q\} < j$ if and only if $p+1 < q'+1$ where $q' = j$ if $q = n+1$ and $q' = q$ otherwise; moreover inequalities of $C(\Delta'_\pi)w \geqq d(\Delta'_\pi)$ with $\min\{p, q\} < j$ correspond in this way to pairs in the set E. Hence $\max\{1, 2(n+1-j)\} + \{\phi(j+1, n)\}$ gives an upper bound on $|\Delta'_\pi|$. (The first term comes from counting edges of the form $(n+1, p)$ or $(p, n+2)$.)

THEOREM 3. *$\phi(m, n) = \min\{m(m-1)/2,\ 2m\sqrt{(n+1)/3}\}$ is an upper bound on $|E|$.*

Proof. Since $p < q$ implies $p < q$, $|E| \leqq m(m-1)/2$. So assume that $2m\sqrt{(n+1)/3} < m(m-1)/2$, so that $\sqrt{3(n+1)} < m-1$.

Define variables x_{pq}, $1 \leqq p < q \leqq m$, to be 1 or 0 according as the pair (p, q) lies in E or not. Define

$$y_s = \sum_{p=1}^{m-s} x_{p,p+s} \quad \text{for } 1 \leqq s < m.$$

Then

$$|E| = \sum_{s=1}^{m-1} y_s.$$

Now $x_{pr} = 1$ implies that there is some index $i \in N$ with $i \in S_p \cap S_r$, $i \notin S_q$ for $p < q < r$. Thus in the $m \times n$ matrix of incidences of the sets S_p with N, there is a column with the pattern $(\cdots, 1, 0, \cdots, 0, 1, \cdots)$ with the ones in rows p and r. The sequence of zeros followed by the one in row r occupies $r-p$ positions, and cannot be associated with any other pair (p, r) in E. Thus

$$\sum_{s=1}^{m-1} s y_s \leqq mn, \tag{7}$$

where the left side gives the number of positions in the incidence matrix required to obtain the elements of E and the right side is the total number of entries in the incidence matrix.

Next, we have $x_{pq} + x_{qr} + x_{pr} \leqq 2$ for all $1 \leqq p < q < r \leqq m$, since if $(p, r) \in E$ we cannot have (p, q) and (q, r) in E. Thus for $1 \leqq s < u < m$

$$\sum_{p=1}^{m-u} (x_{p,p+s} + x_{p+s,p+u} + x_{p,p+u}) \leqq 2(m-u);$$

adding to this

$$\sum_{p=m-u+1}^{m-s} x_{p,p+s} \leqq u - s$$

and

$$\sum_{p=1-s}^{0} x_{p+s,p+u} \leqq s$$

gives

$$y_s + y_t + y_u \leqq 2m \tag{8}$$

whenever $s + t = u$, $1 \leqq s, t, u < m$.

Now for any odd $s \geqq 3$ we deduce from (8)

$$\begin{aligned}
y_1 & & & & + y_{s-1} + y_s &\leqq 2m, \\
y_2 & & & + y_{s-2} & + y_s &\leqq 2m, \\
\ddots & & \iddots & & \vdots & \\
& y_{(s-1)/2} + y_{(s+1)/2} & & & + y_s &\leqq 2m;
\end{aligned}$$

summing these inequalities gives

$$\sum_{r=1}^{s-1} y_r + \frac{s-1}{2} y_s \leqq 2m \frac{s-1}{2}. \tag{9}$$

This inequality is also valid for $s \geqq 3$ and even, using half of

$$2y_{s/2} + y_s \leqq 2m$$

as the final inequality. Then, by induction, (9) yields

$$\sum_{r=1}^{s} y_r \leqq \frac{2ms}{3} \tag{10}$$

for any $s \geqq 3$. Indeed, (10) is exactly (9) for $s = 3$, while $(t-2)/2$ times (10) for $s = t$ added to (9) for $s = t+1$ yields $t/2$ times (10) for $s = t+1$. Further, we have, trivially,

$$y_1 \leqq m \quad \text{and} \quad y_1 + y_2 \leqq 2m. \tag{11}$$

Then adding inequality (7) to the inequalities in (11) and inequality (10) for $s = 3, 4, \cdots, t-1$ we obtain

$$t \sum_{s=1}^{t-1} y_s + \sum_{s=t}^{m-1} s y_s \leqq mn + m + \frac{mt(t-1)}{3}$$

so that

$$|E| = \sum_{s=1}^{m-1} y_s \leqq \frac{m(n+1)}{t} + \frac{m(t-1)}{3}. \tag{12}$$

Now choose $t = \lceil \sqrt{3(n+1)} \rceil \leqq m-1$ where for any real λ, $\lceil \lambda \rceil$ denotes the least integer not less than λ and $\lfloor \lambda \rfloor$ the greatest integer not greater than λ. Then $m(n+1)/t \leqq m\sqrt{(n+1)/3}$ and $m(t-1)/3 \leqq m\sqrt{(n+1)/3}$ and the theorem is proved.

Next we show that the bound in Theorem 3 is essentially tight.

THEOREM 4. *The sets $S_1, \cdots, S_m$ can be chosen so that*

$$|E| \geqq \min\left\{\left\lfloor \frac{m^2}{4} \right\rfloor, \frac{m\lfloor \sqrt{n} \rfloor}{2}\right\}.$$

Proof. Suppose first $n \geqq \lfloor m^2/4 \rfloor$. Assume m is even—the construction is similar with m odd. Then we choose $S_1, \cdots, S_m$ so that $(p, q) \in E$ if and only if $p \leqq m/2 < q$. Indeed, since $n \geqq m^2/4$ we can assign to each pair (p, q) a different $i \in N$, and then let S_p and S_q but no other S_r contain this i. It is clear that $|E| = m^2/4$ with this construction.

Now suppose $n < \lfloor m^2/4 \rfloor$ so that $v \equiv \lfloor \sqrt{n} \rfloor < m/2$. Write $m = rv + s$, $0 \leqq s < v$. We then divide $M = \{1, 2, \cdots, m\}$ into $M_1, \cdots, M_r$, the first, second, $\cdots$, rth group of v elements, and M_{r+1}, the last s elements of M. Since $n \geqq v^2$ we may associate with each $i \in N$ a pair (t, u) with $1 \leqq t, u \leqq v$. This index i is then put into the $r+1$ sets S_t, $S_{v+\{t+u\}}$, $S_{2v+\{t+2u\}}$, $\cdots$, $S_{rv+\{t+ru\}}$ where $\{h\}$ denotes the integer between 1 and v that equals h modulo v. For certain t, u, the final S may have an index greater than m—arbitrarily change it to m. It can be checked that the resulting E contains all pairs (p, q) with $p \in M_k$, $q \in M_{k+1}$, $k = 1, 2, \cdots, r$; thus there are $(r-1)v^2 + vs = v(rv + s - v) = v(m - v) > mv/2 = m\lfloor \sqrt{n} \rfloor /2$ pairs in E. This completes the proof.

Theorem 3 implies that the number of facets of $\check{\sigma}$ can be no more than $1 + 2(n+2)\sqrt{(n+1)/3}$. Theorem 4 (by adding sets $S_0 = S_{m+1} = N$ to the extremes of the sets constructed there) shows that there may be as many as $1 + n\lfloor \sqrt{n} \rfloor /2$. That is, the bound is $O(n^{3/2})$ and is tight.

5. Discussion. Section 3 has described how a large-piece implementation can be designed, based on Theorem 2. Thus we use the linear system

$$A_{\hat{\sigma}} w = b, \qquad C_{\Delta} w \geqq d^{\Delta}, \tag{13}$$

where $C_\Delta = C(\Delta'_\pi)$, $d^\Delta = d(\Delta'_\pi)$. Some measure of the complexity of such an implementation is provided by the number of inequalities in (13), $|\Delta'_\pi|$, which we have seen can be $O(n^{3/2})$.

An alternative implementation would use the smaller pieces $\hat{\sigma}$ and Theorem 1 instead. This would therefore involve the linear system

$$A_{\hat{\sigma}} w = b, \qquad C_\Gamma w \geqq d^\Gamma, \tag{14}$$

where $C_\Gamma = C(\Gamma'_\pi)$, $d^\Gamma = d(\Gamma'_\pi)$. Note that (14) has at most $2n+1$ inequalities. If we move from $\hat{\sigma}$ into an adjacent piece $\tilde{\sigma}$ and l is linear on $\hat{\sigma} \cup \tilde{\sigma}$ (see the discussion below Theorem 1), then $A_{\tilde{\sigma}} = A_{\hat{\sigma}}$ and C_Γ and d^Γ have at most three new rows; only two new ratios need be computed. We can then search for the new minimum ratio (i.e., traverse $\tilde{\sigma}$) and continue in this way. If it is likely that several small pieces belonging to one large piece will be encountered, then it may be worthwhile maintaining the ratios in a heap instead of recomputing the minimum afresh each time.

To determine whether it is more efficient to use (13) than (14) and whether it is worthwhile to form a heap in the latter case, we need to answer two questions. Firstly, is it likely that (13) will have more than a small multiple of n inequalities; secondly, are the pieces $\check{\sigma}$ generally composed of many pieces $\hat{\sigma}$, or, more precisely, can line segments in pieces $\check{\sigma}$ meet many pieces $\hat{\sigma}$? We give some indication of the answer to these questions.

First note that the number of inequalities in (13) equals the number of facets of the piece $\check{\sigma}$. Since, from Lemma 1, each piece $\hat{\sigma}$ has at most $2n+1$ facets, we conclude that whenever (13) has $O(n^{3/2})$ inequalities, $\check{\sigma}$ consists of at least $O(n^{1/2})$ pieces $\hat{\sigma}$. A much more complete analysis can be made in particular cases.

Let us first consider the example constructed in Theorem 4 (we assume $m = n$). The number of pieces $\hat{\sigma}$ contained in one piece $\check{\sigma}$ is then the number of permutations that keep the first $\lfloor\sqrt{n}\rfloor$ positions ahead of the next $\lfloor\sqrt{n}\rfloor$, and so on. However each subgroup can be arbitrarily permuted, so that there are at least $(\lfloor\sqrt{n}\rfloor!)^{\lfloor\sqrt{n}\rfloor} > (n/9)^{n/2}$ such pieces. Secondly, it is possible to choose w and z so that w satisfies (14) but $w + \lambda z$, as λ increases, completely reverses the order of the first $\lfloor\sqrt{n}\rfloor$ terms, the second $\lfloor\sqrt{n}\rfloor$ terms, and so on before any of these subgroups cross. This implies that at least $\lfloor\sqrt{n}\rfloor^2(\lfloor\sqrt{n}\rfloor - 1)/2$ inequalities of (14) that do not occur in (13) are violated before encountering one from (13); thus there are line segments in $\check{\sigma}$ that meet $O(n^{3/2})$ pieces $\hat{\sigma}$. It is of course very likely that far fewer pieces $\hat{\sigma}$ within a given piece $\check{\sigma}$ will be met by most line segments. For this example it seems reasonable to avoid having $O(n^{3/2})$ inequalities by using (14) rather than (13), but to maintain the ratios in a heap to guard against sequences of a large number of pieces $\hat{\sigma}$ within one piece $\check{\sigma}$.

Let us now consider a much more reasonable example. Suppose that each component f_i of the function f depends on at most r components x_j of the argument x and that each x_j similarly affects at most c f_i's. Then each individual coordinate can interact with at most $c(r-1)$ other coordinates. It follows that (13) has at most $\max\{2, c(r-1)/2\}n$ inequalities. If cr is relatively small (e.g., $c < 5$, $r < 5$) then using (13) does not seem to exact too high a price. Let us examine the possible inefficiencies of using (14) when $Df(x)$ has small band width, i.e., f_i depends on x_j only if $|i-j| < k$ for some $k \ll n$. In this case $c = r = 2k-1$; suppose for simplicity that $n = (2k-1)t$. Let the permutation π take $(1, 2, \cdots, n+1)$ into $(1, c+1, \cdots, (t-1)c+1, 2, c+2, \cdots, n, n+1)$ and let $\check{\sigma} = \check{j}_1(v, \pi, s)$ for some v, s. Then, since the first, second, $\cdots$, cth group of t consecutive elements can be arbitrarily permuted and yet give rise to the same piece $\check{\sigma}$, we see that $\check{\sigma}$ contains exactly $(t!)^c$ pieces $\hat{\sigma}$; for example, if Df is tridiagonal, $c = r = 3$ and $(t!)^c = ((n/3)!)^3$. In addition, by choosing w and z

appropriately as in the previous example, we find that $\check{\sigma}$ contains line segments meeting at least $ct(t-1)/2 = n(t-1)/2 \cong n^2/4k$ pieces $\hat{\sigma}$. Hence this example suggests the use of (13), which can be far more efficient than using (14) without incurring an excessive number of inequalities.

The conclusion from these examples is that use of the new linear system (13) for traversing the large pieces $\check{\sigma}$ is recommended when the function f is very sparse so that Df has $O(n)$ nonzeros; otherwise the use of (3) is suggested, with the ratios maintained in a heap.

Acknowledgment. I am grateful to Mike Powell and the referees for several very helpful suggestions concerning this work.

REFERENCES

[1] E. ALLGOWER AND K. GEORG, *Simplicial and continuation methods for approximating fixed points*, SIAM Rev., 22 (1980), pp. 28–85.

[2] B. C. EAVES, *A short course in solving equations with PL homotopies*, in Nonlinear Programming, Proc. Ninth SIAM-AMS Symposium in Applied Mathematics, R. W. Cottle and C. E. Lemke, eds., Society for Industrial and Applied Mathematics, Philadelphia, 1976, pp. 73–143.

[3] B. C. EAVES AND R. SAIGAL, *Homotopies for computation of fixed points on unbounded regions*, Math. Programming, 3 (1972), pp. 225–237.

[4] B. C. EAVES AND H. E. SCARF, *The solution of systems of piecewise linear equations*, Math. Oper. Res., 1 (1976), pp. 1–27.

[5] LAAN, G. VAN DER AND A. J. J. TALMAN, *A restart algorithm for computing fixed points without an extra dimension*, Math. Programming, 17 (1979), pp. 74–84.

[6] ———, *A class of simplicial restart fixed point algorithms without an extra dimension*, Math. Programming, 20 (1981), pp. 33–48.

[7] O. H. MERRILL, *Applications and extensions of an algorithm that computes fixed points of certain upper semi-continuous point to set mappings*, Ph.D. Dissertation, Department of Industrial Engineering, University of Michigan, Ann Arbor, 1972.

[8] R. SAIGAL, *A homotopy for solving large sparse and structural fixed point problems*, Dept. Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, January 1981.

[9] M. J. TODD, *The Computation of Fixed Points and Applications*, Springer-Verlag, Berlin–Heidelberg–New York, 1976.

[10] ———, *Exploiting structure in piecewise-linear homotopy algorithms for solving equations*, Math. Programming, 18 (1980), pp. 223–247.

[11] ———, *Traversing large pieces of linearity in algorithms that solve equations by following piecewise-linear paths*, Math. Oper. Res., 5 (1980), pp. 242–257.

[12] ———, *Numerical stability and sparsity in piecewise-linear algorithms*, Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980, pp. 1–24.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0004 $01.25/0

WEAKEST CONDITIONS FOR EXISTENCE OF LIPSCHITZ CONTINUOUS KROTOV FUNCTIONS IN OPTIMAL CONTROL THEORY*

R. B. VINTER†

Abstract. Let Y be a metric space; let (y_*^-, y_*^+) be a point in the product space $Y \times Y$, and let v be an "admissible value function" on $Y \times Y$ taking values in the extended real line. We give a necessary and sufficient condition for existence of some Lipschitz continuous function $\phi: Y \to \mathbb{R}$ such that

$$\phi(y^+) - \phi(y^-) \leqq v(y^-, y^+), \quad \text{all } (y^-, y^+) \in Y \times Y,$$

$$\phi(y_*^+) - \phi(y_*^-) = v(y_*^-, y_*^+)$$

(such functions are called Krotov functions). These results provide conditions, which are in a certain sense weakest, for the applicability of methods of Carathéodory in the calculus of variations and optimal control theory concerning validation of extremals.

Notation and conventions. Let S be an abstract set. $\mathcal{F}(S)$ denotes the usual real linear space of functions $f: S \to \mathbb{R}$. $\mathcal{F}'(S)$ is the algebraic dual of $\mathcal{F}(S)$, that is, the usual linear space of linear functionals on $\mathcal{F}(S)$.

As is well known, f' is in $\mathcal{F}'(S)$ if and only if there exists a finite set of points $p_1, \cdots, p_k$ in S and real numbers $\alpha_1, \cdots, \alpha_k$ such that

$$f'(f) = \sum_i \alpha_i f(p_i), \quad \text{all } f \in \mathcal{F}(S).$$

$\bar{\mathbb{R}}$ denotes $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$. Given $a, b \in \bar{\mathbb{R}}$, we follow the customary rules for evaluating "$a+b$" when either a or b is finite, or when a and b are infinite but of the same sign. We set $a+b = b+a = -\infty$ if $a = +\infty$, $b = -\infty$.

1. Introduction. Let Y be a metric space and let $v: Y \times Y \to \bar{\mathbb{R}}$ be a given function which satisfies

$$v(y, y) = 0, \quad \text{all } y \in Y,$$

$$v(y_1, y_2) + v(y_2, y_3) \geqq v(y_1, y_3), \quad \text{all } y_1, y_2, y_3 \in Y$$

(such a function will be called an "admissible value function"). Suppose that (y_*^-, y_*^+) is a point in $Y \times Y$. Our main result is a necessary and sufficient condition (we shall call it "strong calmness") under which there exists a Lipschitz continuous function $\phi: Y \to \mathbb{R}$, satisfying

$$\phi(y^+) - \phi(y^-) \leqq v(y^-, y^+), \quad \text{all } (y^-, y^+) \in Y \times Y,$$

$$\phi(y_*^+) - \phi(y_*^-) = v(y_*^-, y_*^+).$$

Functions ϕ satisfying these conditions have previously been referred to as Krotov functions in the literature [5].

What is the significance of such a result? To answer this question we examine the optimal control problem: let a function $l: \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and points

* Received by the editors January 25, 1982. This research was supported in part by the National Research Council under grant A9082, while the author was visiting the Department of Mathematics, University of British Columbia, Canada V6T 1Y4.

† Department of Electrical Engineering, Imperial College of Science and Technology, London SW7 2BT, England.

(x_*^-, t_*^-), (x_*^+, t_*^+) in $\mathbb{R}^{n+1}$, $t_*^+ > t_*^-$, be given. We seek the infimum of the functional

$$\int_{t_*^-}^{t_*^+} l(x(t), t, \dot{x}(t))\, dt \tag{1.1}$$

over admissible arcs $x(\cdot)$ satisfying the endpoint conditions

$$x(t_*^-) = x_*^-, \qquad x(t_*^+) = x_*^+. \tag{1.2}$$

By "admissible arc" we mean an absolutely continuous function $x(\cdot)$ such that $t \to l(x(t), t, \dot{x}(t))$ is measurable (we adopt the usual conventions in evaluating the integral when either $t \to \max\{l(x(t), t, \dot{x}(t)), 0\}$ or $t \to \max\{-l(x(t), t, \dot{x}(t)), 0\}$ is integrable, and set the integral to $-\infty$ when neither is integrable). An admissible arc which satisfies (1.2) and achieves the infimum will be called a minimizing arc. The infimum of the integrals is the minimum cost.

Although this problem has the guise of a problem in the calculus of variations, we refer to it as an optimal control problem because l is permitted to take the value $+\infty$; this feature permits us to treat within our formulation constraints on $x(t)$ and $\dot{x}(t)$ normally seen as lying in the domain of optimal control theory.

An approach to solving this problem commonly associated with the name of Carathéodory (see [1] and also [6], [14] for historical background) is centered on the following simple result:

PROPOSITION 1.1. *Let $x(\cdot)$ be an admissible arc satisfying* (1.2) *and $\phi: \mathbb{R}^{n+1} \to \mathbb{R}$ a continuously differentiable function satisfying*

$$\phi_t(x, t) + \phi_x(x, t)\dot{x} \leqq l(x, t, \dot{x}) \quad \textit{for all } (x, t, \dot{x}), \tag{1.3}$$

$$\phi(x_*^+, t_*^+) - \phi(x_*^-, t_*^-) = \int_{t_*^-}^{t_*^+} l(x(t), t, \dot{x}(t))\, dt. \tag{1.4}$$

Then $x(\cdot)$ is a minimizing arc and $\phi(x_*^+, t_*^+) - \phi(x_*^-, t_*^-)$ is the minimum cost.

Proof. For any admissible arc $\bar{x}(\cdot)$ satisfying (1.2) we have

$$\begin{aligned}\int_{t_*^-}^{t_*^+} l(x(t), t, \dot{x}(t))\, dt &= \phi(x_*^+, t_*^+) - \phi(x_*^-, t_*^-) \\ &= \int_{t_*^-}^{t_*^+} [\phi_t(\bar{x}(t), t) + \phi_x(\bar{x}(t), t)\dot{\bar{x}}(t)]\, dt \\ &\leqq \int_{t_*^-}^{t_*^+} l(\bar{x}(t), t, \dot{\bar{x}}(t))\, dt.\end{aligned}$$

□

It is customary in the solution of specific problems to choose the function ϕ in Proposition 1.1 as the cost (1.1) associated with a "field of extremals" parameterized, say, by the right endpoints of the field elements (see [6] or [13]). We note however that ϕ, if it exists, may be determined as the solution of an optimization problem:

PROPOSITION 1.2. *Suppose there exists some continuously differentiable function ϕ satisfying* (1.3) *and* (1.4) *for some admissible arc $x(\cdot)$ satisfying* (1.2). *Then ϕ achieves the maximum of $\psi(x_*^+, t_*^+) - \psi(x_*^-, t_*^-)$ over the class of continuously differentiable functions ψ satisfying*

$$\psi_t(x, t) + \psi_x(x, t)\dot{x} \leqq l(x, t, \dot{x}) \quad \textit{for all } x, t, \dot{x}. \tag{1.5}$$

Proof. For any continuously differentiable ψ satisfying (1.5), we have

$$\begin{aligned}\phi(x_*^+, t_*^+) - \phi(x_*^-, t_*^-) &= \int_{t_*^-}^{t_*^+} l(x(t), t, \dot{x}(t))\, dt \\ &\geqq \int_{t_*^-}^{t_*^+} [\psi_t(x(t), t) + \psi_x(x(t), t)\dot{x}(t)]\, dt \\ &= \psi(x_*^+, t_*^+) - \psi(x_*^-, t_*^-). \quad \square\end{aligned}$$

While a number of problems have been solved by methods in the spirit of Proposition 1.1 (see [6], [7], [14]), there are other problems, notably many problems involving "state" and "control" constraints, to which Proposition 1.1 is not applicable. The approach fails here because, loosely speaking, the requirement that the function ϕ be continuously differentiable is too strong.

Such considerations motivated Ioffe to modify the optimality condition inherent in Proposition 1.1 so that it make sense even when the function ϕ is not continuously differentiable and thereby to extend its applicability. These modifications we now describe.

Let $\eta((x^-, t^-), (x^+, t^+))$ be the infimum of the integral (1.1) over admissible arcs $x(\cdot)$ satisfying $x(t^-) = x^-$, $x(t^+) = x^+$. (The value of the infimum is taken as $+\infty$ if no such admissible arcs exist.) It is clear that if a continuously differentiable function ϕ satisfies (1.3), (1.4) then ϕ also satisfies

$$\phi(x^+, t^+) - \phi(x^-, t^-) \leqq \eta((x^-, t^-), (x^+, t^+)), \quad \text{all } ((x^-, t^-), (x^+, t^+)) \in \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}, \tag{1.6}$$

$$\phi(x_*^+, t_*^+) - \phi(x_*^-, t_*^-) = \eta((x_*^-, t_*^-), (x_*^+, t_*^+)). \tag{1.7}$$

We shall refer to (1.6), (1.7) as the integrated versions of (1.3), (1.4). The noteworthy fact about the integrated versions is that they make sense for an arbitrary function ϕ. That an optimality condition can still be given in terms of them is obvious.

PROPOSITION 1.3. *Suppose that there exists a function* $\phi : \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \to \mathbb{R}$ *satisfying* (1.6), (1.7). *Then the admissible arc* $x(\cdot)$ *satisfying* (1.2) *is minimizing if and only if*

$$\phi(x_*^+, t_*^+) - \phi(x_*^-, t_*^-) = \int_{t_*^-}^{t_*^+} l(x(t), t, \dot{x}(t))\, dt.$$

The modified optimality condition (Proposition 1.3) is applicable, in principle, to a very wide class of problems even if ϕ is restricted to be continuous. This we conclude from the following notable result proved by Ioffe [5].

THEOREM 1.1. *Suppose the function* η *is finite at the point* $((x_*^-, t_*^-), (x_*^+, t_*^+))$. *Then*:

a) *there exists some function* ϕ *satisfying* (1.6), (1.7) *if and only if there exists some* ϕ *satisfying* (1.6), *and*

b) *if* η *is lower semicontinuous on* $\mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$, *then there exists some continuous function* ϕ *satisfying* (1.6), (1.7) *if and only if there exists some continuous function* ϕ *satisfying* (1.6).

These conditions assuring existence of ϕ satisfying (1.6), (1.7) are very mild. For example, the conditions under which a continuous function ϕ exists are often met in problems where l is nonnegative valued and either l is convex in its $\dot{x}$ dependence or we permit relaxation of the controls [13].

Thus we achieve in Proposition 1.3 our goal of significantly extending the range of applicability of Proposition 1.1. There is however a major drawback in these developments: the intractability of the integrated constraint (1.6). How should we test that a function ϕ generated by, say, a field of extremals, satisfies (1.6), which is expressed in terms of η, the value of which at $((x^-, t^-), (x^+, t^+))$ we are trying to find?

One way round the difficulty is to give an optimality condition in terms of a sequence of smooth functions ϕ such that the infinitesimal form (1.3) of the constraint is satisfied along the sequence [12]. If however we aim for a more precise optimality condition expressed in terms of a single function ϕ, we require a variant of Theorem 1.1 which asserts, under weakest conditions, the existence of a function ϕ which satisfies (1.6), (1.7) and for which (1.6) can be expressed in testable, infinitesimal form. (Of course the second condition (1.7) also depends on η but this is not cause for concern: when ϕ is generated by a field of extremals we may replace (1.7) by (1.4), in which $x(\cdot)$ is the field element satisfying the boundary conditions (1.2). We view (1.4) as testable since it does not involve η.) It would appear that the class of Lipschitz continuous functions is just about the largest of the simply described classes of functions ϕ for which condition (1.6) can be expressed in infinitesimal form.

It is against this background that our main result, which provides (in the context of optimal control theory) a weakest condition assuring existence of a Lipschitz continuous function ϕ satisfying (1.6), (1.7) should be viewed. Indeed, η is an example of an admissible value function. The foregoing suggests a sense in which our results provide the weakest condition under which the Carathéodory approach works.

In [10] (this issue, pp. 235–245) an analogue of Proposition 1.2 for Lipschitz continuous functions ϕ is proved for a related optimal control problem in which (1.5) is replaced by

$$\psi_t(x, t) + \psi_x(x, t)\dot{x} \leqq l(x, t, \dot{x}) \quad \text{for all } \dot{x} \text{ and a.e. } (x, t): \tag{1.8}$$

Equation (1.8) is the appropriate infinitesimal form of (1.6). We also show that our results on existence of a Lipschitz continuous function satisfying (1.6), (1.7) lead directly to new necessary conditions that an admissible arc be minimizing [10].

Our results, we believe, make a striking connection with another branch of control theory. It turns out that the condition we introduce, "strong calmness", is a condition intermediate in strength between calmness of the optimal control problem and a stronger, uniform notion of calmness. The condition of "calmness" introduced by Clarke (see, e.g., [2] or [3]), is the weakest known condition assuring normality of the control problem. (Normality is the condition that the Pontryagin maximum principle yield a nonzero multiplier associated with the functional to be minimized and, therefore, in a sense, yield nontrival information.) Actually the notions of calmness and strong calmness essentially coincide under mild compactness and Lipschitz continuity assumptions on the data (see [11]). It is intriguing to observe that the theory of necessary conditions in optimal control theory and the Carathéodory approach rest on the foundations of related hypotheses. Such connections were anticipated by L. C. Young [12, p. 264].

Various conditions are available elsewhere in the literature [4], [9], [14] from which may be deduced conditions under which the conclusion of our main theorem applies. The weakest appears to be the condition of Lipschitz continuity of η on its effective domain, hypothesised by Lewis [8]. This Lipschitz continuity condition is stronger than uniform calmness (which amounts to a kind of one-sided Lipschitz continuity assumption) which, in turn, is stronger than our strong calmness. All these papers

(and also [12]) make continuity and compactness assumptions on the data, which play no part here.

Finally some comments about our methods. In broad outline these pattern those of Ioffe [5], used in proving Theorem 1.1. As in [5] we interpret ϕ's satisfying (1.6), (1.7) as subgradients of a convex function, and thereby reduce the problem to one of establishing nonemptiness of a subdifferential. We too must confront the difficulty encountered by Ioffe in [5]—that this convex function has, possibly, effective domain with empty interior. Ioffe uses a criterion for nonemptiness of the subdifferential which applies in such circumstances; we get round the difficulty by constructing a convex cone whose interior is both nonempty and disjoint from the epigraph of the convex function, so that an elementary separation principle is applicable. The main difference between [5] and this paper is the choice of spaces. Since the subgradient is to be a Lipschitz continuous function, we must adopt as domain of the convex function some predual space of the Lipschitz continuous functions. What should this be? The answer is provided by concepts due to L. C. Young. Young's space of "(simplical) boundaries" with the boundary norm [12] is indeed such a predual space and is ideally suited to the present application. However, the results on boundaries in [14] must be extended and reworked since the continuity and compactness assumptions in [14] are not present in our development. We believe that the space of boundaries is exploited as a predual of the space of Lipschitz continuous functions here for the first time.

Concerning related research, we mention that the problem of maximizing $\phi(y_*^+) - \phi(y_*^-)$ over continuous functions which satisfy (1.6) can be interpreted as the dual problem (in the sense of convex programming) of a special case of Kantorovich's "mass transportation" problem. The mass transportation problem has been extensively studied by Levin and Milyutin [8]. These authors are primarily concerned with conditions under which the values of the mass transportation problem and its dual problem coincide with respect to both continuous and Lipschitz continuous ϕ's; in contrast, our concerns, viewed in relation to the mass transportation problem, are with conditions under which the dual problem has as solution a Lipschitz continuous function.

2. Admissible value functions. Let Y be a metric space. The metric on Y will be written m.

DEFINITION 2.1. A function $v\colon Y \times Y \to \bar{\mathbb{R}}$ is an *admissible value function* if and only if

$$v(y, y) = 0 \quad \text{for all } y \in Y,$$

$$v(y_1, y_2) + v(y_2, y_3) \geqq v(y_1, y_3) \quad \text{for all } y_1, y_2, y_3 \in Y.$$

We have given one example of an admissible value function in § 1 (the function η on $\mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$). A special feature of this example, which is common to all admissible value functions arising from optimal control problems in which one coordinate is identified as time (so called nonparametric problems), is that

$$\eta((x^-, t^-), (x^+, t^+)) = +\infty \quad \text{when } t^+ < t^-. \tag{2.1}$$

Note however that conditions such as (2.1) play no part in Definition 2.1, and our results on admissible value functions (to follow) are relevant to control problems not covered by the formulation in § 1, and in particular to parametric problems.

Notice that the metric m defines an admissible value function on $Y \times Y$. That this is so follows simply from the defining properties of the metric.

3. Strong calmness. Let v be an admissible value function, and let (y_*^-, y_*^+) be a point in $Y \times Y$. By analogy with the definition current in the optimal control and mathematical programming literature [2], [3], we say v is *calm* at (y_*^-, y_*^-) if there exists a real number c such that

$$v(y^-, y^+) - v(y_*^-, y_*^+) \geqq -c[m(y^-, y_*^-) + m(y^+, y_*^+)] \tag{3.1}$$

for all $(y^-, y^+) \in Y \times Y$ (m is the metric on Y, remember).

We have commented in § 1 on the significance of such a condition in optimal control theory as assuring nontriviality (i.e., normality) of the conclusions of the Pontryagin maximum principle. (It should be added however that calmness understood in the above sense, when specialized to apply to the function $\eta: \mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \to \bar{\mathbb{R}}$ of § 1, differs slightly, and is rather stronger, than the condition of the same name hypothesized by Clarke [2]. Clarke's hypothesis requires (3.1) to hold only locally, in a sense made precise in [2], and for perturbations only of the x-component of either y_*^- or y_*^+.)

We shall say that v is *uniformly calm* if there exists a real number c such that (3.1) holds for all $y^-, y^+, y_*^- y_*^+ \in Y$.

The notions of calmness and uniform calmness will not enter elsewhere into this paper. Our reasons for including them here is that they are natural, simply stated conditions on the stability of the associated minimization problem under data perturbations, which illuminate the concept in terms of which our main result is expressed, namely strong calmness.

Before introducing strong calmness we take note of the interpretation of calmness suggested by Fig. 1: we can view v as defined on directed line segments $\sigma(y^-, y^+)$ joining the arbitrary points y^-, y^+ in Y. v is calm at $\sigma^* = \sigma(y_*^-, y_*^+)$ when there exists some real number c such that, on replacing the segment σ^* by a new segment $\sigma = \sigma(y^-, y^+)$, the decrease in the value of v is not greater than c times the sum of the lengths of the gaps (the dashed lines) in any polygonal arc joining y_*^- and y_*^+ which includes the new segment σ.

The condition of strong calmness of v at (y_*^-, y_*^+) (with modulus of calmness c) imposes an essentially similar restriction on v, with the exception that now we consider replacement of σ^* by an arbitrary finite set of segments $\{\sigma_1, \cdots, \sigma_k\}$: the decrease in the sum of the values of v at $\sigma_1, \cdots, \sigma_k$, as compared with that at σ^* must not exceed c times the sum of the lengths of the gaps in any polygonal arc from y_*^- to y_*^+ which includes $\sigma_1, \cdots, \sigma_k$. (See Fig. 1 again.)

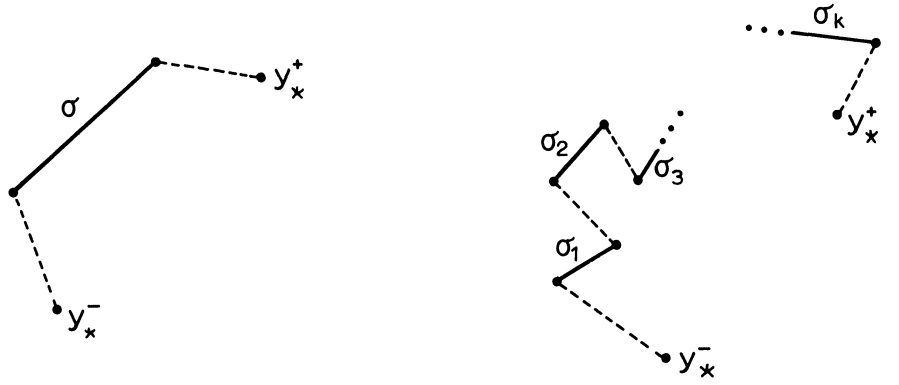


FIG. 1. *The strong calmness condition.*

DEFINITION 3.1. The admissible value function $v: Y \times Y \to \bar{\mathbb{R}}$ is *strongly calm* at $(y_*^-, y_*^+) \in Y \times Y$ if and only if $v(y_*^-, y_*^+)$ is finite and there exists a real number c such that for any collection of points $\{(y_1^-, y_1^+), \cdots, (y_k^-, y_k^+)\}$ in $Y \times Y$ we have:

$$(3.2)\quad \sum_{i=1}^{k} v(y_i^-, y_i^+) - v(y_*^-, y_*^+) \geqq -c\left[\sum_{i=1}^{k-1} m(y_{i+1}^-, y_i^+) + m(y_*^-, y_1^-) + m(y_k^+, y_*^+)\right]$$

and

$$(3.3)\quad \sum_{i=1}^{k} v(y_i^-, y_i^+) \geqq -c\left[\sum_{i=1}^{k-1} m(y_{i+1}^-, y_i^+) + m(y_k^+, y_1^-)\right].$$

The minimum of all nonnegative values of c satisfying (3.2), (3.3) for all collections is called the *modulus of calmness* of v at (y_*^-, y_*^+).

The qualification "essentially" in the paragraph preceding Definition 3.1 refers to the extra condition (3.3). This condition is very weak; when v is associated with some optimal control problem as in § 1, we may arrange that (3.3) holds whenever the function l in (1.1) is bounded below.

We have used terms "segment" and "arc" in this section in a loose, provisional sense to describe the ideas behind Fig. 1. These terms will be given a different, precise meaning below.

4. The main result. The class of Lipschitz continuous functions on Y, written Lip, is understood in the following sense:

DEFINITION 4.1. The set Lip is the family of functions $\phi: Y \to \mathbb{R}$ which satisfy

$$\sup_{y \neq y'} \{|\phi(y) - \phi(y')|/m(y, y')\} < \infty.$$

The Lipschitz constant of $\phi \in \text{Lip}$ is the number c,

$$(4.1)\quad c = \sup_{y \neq y'} \{|\phi(y) - \phi(y')|/m(y, y')\}.$$

Given an admissible value function v and a point (y_*^-, y_*^+) in $Y \times Y$, we examine conditions under which there exists an element $\phi \in \text{Lip}$ with the following properties:

$$(4.2)\quad \phi(y^+) - \phi(y^-) \leqq v(y^-, y^+) \quad \text{for all } (y^-, y^+) \in Y \times Y,$$

$$(4.3)\quad \phi(y_*^+) - \phi(y_*^-) = v(y_*^-, y_*^+).$$

It is a simple matter to show that there is some connection between the conditions (4.2), (4.3) and strong calmness (defined in § 3). Indeed suppose that $\phi \in \text{Lip}$, with Lipschitz constant c, satisfies (4.2), (4.3) for some $(y_*^-, y_*^+) \in Y \times Y$. From (4.2), (4.3) we deduce

$$\sum_{i=1}^{k} v(y_i^-, y_i^+) - v(y_*^-, y_*^+) \geqq \sum_{i=1}^{k} [\phi(y_i^+) - \phi(y_i^-)] - [\phi(y_*^+) - \phi(y_*^-)].$$

The right-hand side may be rearranged to give

$$\sum_{i=1}^{k-1} [\phi(y_i^+) - \phi(y_{i+1}^-)] + \phi(y_*^-) - \phi(y_1^-) + \phi(y_k^+) - \phi(y_*^+).$$

Inequality (3.2) now follows from (4.1).

Likewise we deduce from (4.2) that

$$\sum_{i=1}^{k} v(y_i^-, y_i^+) \geqq \sum_{i=1}^{k} [\phi(y_i^+) - \phi(y_i^-)]$$

$$= \sum_{i=1}^{k-1} [\phi(y_i^+) - \phi(y_{i+1}^-)] + \phi(y_k^+) - \phi(y_1^-),$$

which implies (3.3) by (4.1). Note that the modulus of calmness is at most c.

Thus, strong calmness is necessary for existence of $\phi \in \text{Lip}$ satisfying (4.2), (4.3). It is perhaps unexpected that strong calmness is also sufficient. This is our main result. We summarize:

THEOREM 4.1. *Let v be an admissible value function and (y_*^-, y_*^+) a point in $Y \times Y$. Then there exists $\phi \in \text{Lip}$ with Lipschitz constant c satisfying*

$$\phi(y^+) - \phi(y^-) \leqq v(y^-, y^+) \quad \textit{for all } (y^-, y^+) \in Y \times Y,$$

$$\phi(y_*^+) - \phi(y_*^-) = v(y_*^-, y_*^+)$$

if and only if v is strongly calm at (y_*^-, y_*^+) with modulus of calmness c.

In general terms, Theorem 4.1. has the form that Ioffe's Theorem 1.1 would lead one to expect. Ioffe showed that a continuous function $\phi\colon \mathbb{R}^{n+1} \to \mathbb{R}$ satisfying (1.6), (1.7) exists if, essentially, the admissible value function η of § 1 satisfies a kind of one-sided continuity condition (namely lower semicontinuity). Theorem 4.1 asserts that a Lipschitz continuous function $\phi\colon Y \to \mathbb{R}$ satisfying (4.2), (4.3) exists if v satisfies a kind of one-sided Lipschitz continuity condition (namely strong calmness).

The rest of this paper is given over to proving Theorem 4.1 and, prior to that, setting up the necessary machinery for the purpose.

5. Flows and their boundaries. In this and the next section, we treat Y as merely a nonempty abstract set.

An ordered pair of points $(s^-, s^+) \in Y \times Y$ defines a *segment* s. This is an element in $\mathscr{F}'(Y \times Y)$ defined by

$$s(g) = g(s^-, s^+) \quad \text{for } g \in \mathscr{F}(Y \times Y)$$

(the spaces $\mathscr{F}(Y \times Y)$, $\mathscr{F}'(Y \times Y)$ were defined in the opening section). We refer to the ordered pair (s^-, s^+) as the *endpair*, and to s^- and s^+ as the *left* and *right endpoints*, of s.

A *flow* f is a linear combination in $\mathscr{F}'(Y \times Y)$ of segments $\{s_i\}$

$$f = \sum_{i=1}^{k} \alpha_i s_i \tag{5.1}$$

in which the coefficients $\{\alpha_i\}$ are nonnegative. Linear combinations in $\mathscr{F}'(Y \times Y)$ with nonnegative coefficients are referred to as *mixtures*. For example, a flow is a mixture of segments.

An *arc* is a flow f expressible as (5.1), but in which $\alpha_i = 1$, $i = 1, \cdots, k$, and $s_i^+ = s_{i+1}^-$, $i = 1, \cdots, k-1$. Here s_i^+ and s_i^- are the right and left endpoints of s_i. If, further $s_k^+ = s_1^-$, then the arc is *closed*. The endpair and left and right endpoints of an arc which is not closed are defined in an obvious way. A *subarc* of an arc p is an arc whose constituent segments are also constituent segments of p. A *two-arc* is an arc expressible in terms of two segments.

Given a flow f, the *boundary* of f, written ∂f, is an element in $\mathscr{F}'(Y)$ defined by

$$\partial f(\phi) = f(\tilde{\phi}), \quad \text{for all } \phi \in \mathscr{F}(Y),$$

where $\tilde{\phi}(y^+, y^-) = \phi(y^+) - \phi(y^-)$. It is so called because it is defined through the restriction of f to some subspace of $\mathcal{F}(Y \times Y)$. The space of boundaries is written $\mathcal{B}$.

A point $b \in \mathcal{F}'(Y)$ is in $\mathcal{B}$ if and only if it may be expressed

$$b(\phi) = \sum_i \alpha_i [\phi(y_i^+) - \phi(y_i^-)], \quad \text{all } \phi \in \mathcal{F}(Y), \tag{5.2}$$

in which the α_i's are nonnegative and the y_i^+'s, y_i^-'s are points in Y. Indeed b given by (5.2) is the boundary of f given by

$$f(g) = \sum_i \alpha_i g(y_i^-, y_i^+) \quad \text{for all } g \in \mathcal{F}(Y \times Y). \tag{5.3}$$

Conversely, every flow f can be expressed (5.1) and, therefore, has boundary of the form (5.2).

It is clear from the preceding observation that $\mathcal{B}$ is a linear subspace of $\mathcal{F}'(Y)$. $\mathcal{B}$ is obviously a convex cone. However, given $b \in \mathcal{B}$ there is a corresponding $-b$, obtained by interchanging each y_i^- and the associated y_i^+ in the representation (5.2) of b. Thus, $\mathcal{B}$ is a linear subspace.

Another representation of boundaries is useful. A point $b \in \mathcal{F}'(Y)$ is in $\mathcal{B}$ if and only if it is zero or it can be expressed

$$b(\phi) = \sum_i \alpha_i^+ \phi(y_i^+) - \sum_j \alpha_j^- \phi(y_j^-), \quad \text{all } \phi \in \mathcal{F}(Y), \tag{5.4}$$

in which the α_i^+, α_j^-'s are positive, the y_i^+, y_j^-'s are distinct and

$$\sum_i \alpha_i^+ = \sum_j \alpha_j^-.$$

(The y_i^-, y_j^+'s and α_i^+, α_j^-'s in (5.2) and (5.4) are not necessarily the same.) Indeed, given (nonzero) b expressed as (5.2), we may obviously re-express it as (5.4) by coalescing repeating y_j^-, y_i^+'s and throwing out terms with zero coefficients. Conversely, given (nonzero) $b \in \mathcal{F}'(Y)$ satisfying (5.4), then b is the boundary of the flow:

$$f(g) = \left(\sum_i \alpha_i^+\right)^{-1} \sum\sum_{i,j} \alpha_i^- \alpha_j^+ s_{ij};$$

here s_{ij} is the segment with endpair (y_i^-, y_j^+).

Some terminology is required in connection with boundaries. The representation (5.4), which is obviously unique to within ordering of terms in the summation, is referred to as the *normal representation* of b. We denote by $\text{supp}^+\{b\}$ the set $\{y_i^+\}$ (in the representation (5.4)) and by $\text{supp}^-\{b\}$ the set $\{y_j^-\}$.

One may think of $\text{supp}^+\{b\}$ and $\text{supp}^-\{b\}$ as the supports of the positive and negative components, obtained from Jordan decomposition, of the measure associated with b.

A flow with zero boundary is called a *closed flow*. Obviously mixtures of closed arcs are closed flows. Of crucial importance is the converse of this result:

LEMMA 5.1. *Every closed flow is expressible as mixture of closed arcs.*

Proof. Let f be a closed flow. We may assume that f is nonzero since, otherwise, the result is trivial. By definition, f is expressible as a mixture of segments

$$f = \sum_i \alpha_i s_i \tag{5.5}$$

in which we may suppose that the coefficients are positive and segments are distinct. Let S be the set of segments in the mixture. Since the mixture (5.5) is expressed in

terms of distinct segments, we may associate with each $s_i \in S$, in a unique way, the coefficient of the corresponding term in (5.5).

Let L, R be sets comprising respectively left, right endpoints of the segments in S. Choose any $y \in Y$. Define $\phi \in \mathcal{F}(Y)$ to take the value 1 at y and zero elsewhere. Since the flow f is closed,

$$\left(\sum_i \alpha_i s_i\right)(\tilde{\phi}) = 0 \quad \text{where } \tilde{\phi}(a, b) = \phi(b) - \phi(a).$$

We conclude

$$\sum_{i \in S_L(y)} \alpha_i = \sum_{i \in S_R(y)} \alpha_i, \tag{5.6}$$

in which $S_L(y)$, $S_R(y)$ is the set of index values i for which y is respectively a left, right endpoint of the segment s_i. Since the α_i's are positive, (5.6) implies that if a point y is some right endpoint then y is some left endpoint and vice versa.

Now choose $y \in R$. We have just observed that we can choose a segment $d_1 \in S$ with left endpoint y. We can proceed and choose $d_2 \in S$ with left endpoint the right endpoint of d_1, and so on, thereby constructing a sequence of segments $d_1, \cdots, d_N$. We terminate this procedure when the right endpoint of a newly selected segment coincides, for the first time, with the left endpoint of a previously selected segment. Termination must occur since S is a finite set. By discarding initial terms in the sequence, we may arrange that $d_1 + d_2 + \cdots + d_N$, which we write p_1, is a closed arc.

Let c_1 be the smallest of the coefficients associated with $d_1, \cdots, d_N$. f may be expressed as

$$f = r + c_1 p_1,$$

in which r is a mixture of segments in some proper subset $S_1 \subset S$. The segments which are eliminated from S to form S_1 are those used in the construction of p_1 and whose coefficients are c_1.

In similar fashion we may decompose r as the sum of a mixture of segments in some proper subset $S_2 \subset S_1$ and a "weighted" closed arc $c_2 p_2$, and so on. Since S is a finite set, after a finite number of steps, M, the set S_M is empty. There follows the desired representation

$$f = \sum_{i=1}^{M-1} c_i p_i. \qquad \square$$

LEMMA 5.2. *Suppose that s is a flow and $b = \partial s$. Then s is expressible as*

$$s = g_1 + g_2,$$

in which g_1 is a mixture of closed arcs and g_2 is absent or a mixture of arcs with endpairs in $\text{supp}^- \{b\} \times \text{supp}^+ \{b\}$.

Proof. We may suppose that the boundary b is nonzero (otherwise the lemma reduces to Lemma 5.1). Let

$$b(\phi) = \sum_i \alpha_i^+ \phi(y_i^+) - \sum_j \alpha_j^- \phi(y_j^-), \quad \text{all } \phi \in \mathcal{F}(Y), \tag{5.7}$$

be the normal representation of b.

Define $Y^* = Y \times \mathbb{R}$. We embed Y in Y^*: $y \to (y, 0)$. Let $\{z_{ij}\}$ be a collection of distinct points in $Y^* \backslash (Y \times \{0\})$. Treating the y_j^-, y_i^+'s as embedded in Y^*, we may

define the segments in $\mathscr{F}'(Y^* \times Y^*)$:

$$s_{ij} = \sigma(y_j^+, z_{ji}) + \sigma(z_{ji}, y_i^-).$$

Here $\sigma(y^-, y^+)$ denotes the segment with endpair (y^-, y^+). Define also the flow in $\mathscr{F}'(Y^*, Y^*)$,

$$s_{\text{rev}} = \left(\sum_i \alpha_i^+\right)^{-1} \sum\sum_{i,j} \alpha_i^- \alpha_j^+ s_{ij}.$$

By embedding the endpoints of the constituent segments of s in Y^*, we may view s as a flow in $\mathscr{F}'(Y^* \times Y^*)$. It is easy to see that

$$\partial s_{\text{rev}} = -\partial s.$$

It follows that $s + s_{\text{rev}}$ is a closed flow (in $\mathscr{F}'(Y^* \times Y^*)$). By Lemma 5.1 then, $s + s_{\text{rev}}$ is expressible as a mixture of closed arcs; thus,

$$s + s_{\text{rev}} = \sum_i c_i d_i. \tag{5.8}$$

We may take the coefficients c_i positive since s, and so certainly $s + s_{\text{rev}}$, is a nonzero flow.

But the constituent elements of s_{rev} are defined through two-arcs with distinct midpoints, and the set of midpoints, which is in $Y^* \backslash (Y \times \{0\})$, is disjoint from the left and right endpoints of segments in $Y \times \{0\}$, a mixture of which represents s. These constituent elements cannot be broken up in the representation of $s + s_{\text{rev}}$ as a mixture (with positive coefficients) of closed arcs. It follows that we may separate out the constituent elements of s_{rev} and write

$$\sum_i c_i d_i = \sum_i c_i \bar{d}_i + s_{\text{rev}}. \tag{5.9}$$

Here the $\bar{d}_i$'s are d_i's from which all two-arcs s_{ij} have been removed. But then from (5.8) and (5.9),

$$s = \sum_i c_i \bar{d}_i. \tag{5.10}$$

The endpoints of the constituent segments of the $\bar{d}_i$'s are in $Y \times \{0\}$. We may therefore view (5.10) as providing a representation of s in $\mathscr{F}'(Y \times Y)$. Our construction expresses s as a mixture of arcs with endpairs in $\text{supp}^-\{b\} \times \text{supp}^+\{b\}$ and, possibly, closed arcs. However, arbitrary closed arcs, with zero coefficients, can always be included in the mixture. The lemma is proved. □

6. Functionals induced by admissible value functions. Let v be an admissible value function (see § 2) which takes value nowhere $-\infty$. Then v induces a functional $\tilde{v}$ on the space of flows, thus,

$$\tilde{v}(f) = \sum_i \alpha_i v(y_i^-, y_i^+), \tag{6.1}$$

where $\sum_i \alpha_i \sigma(y_i^-, y_i^+)$ is some representation of f as a mixture of segments $\sigma(y_i^-, y_i^+)$ with endpairs (y_i^-, y_i^+). In (6.1) we define $\alpha_i v(y_i^-, y_i^+)$ to be zero if $\alpha_i = 0$ and $v(y_i^-, y_i^+) = +\infty$. It is easy to see that, with this convention, the value of $\tilde{v}(f)$ does not depend on the particular representation of f chosen.

We may associate with v a functional ∂v on the boundary space $\mathscr{B}$, via the functional $\tilde{v}$,

$$\partial v(b) = \inf\{\tilde{v}(f) : f \text{ is a flow, } \partial f = b\}. \tag{6.2}$$

The following properties are simple consequences of the definition of "admissible value function" and are stated without proof.

LEMMA 6.1. *Let v be an admissible value function taking value nowhere $-\infty$. Then:*

(i) *For any* arc g *whose endpair coincides with that of the segment s, we have*

$$\tilde{v}(g) \geqq \tilde{v}(s).$$

(ii) *For any closed polygonal arc,*

$$\tilde{v}(g) \geqq 0.$$

LEMMA 6.2. *Let v be an admissible value function taking value nowhere $-\infty$, and let $b \in \mathcal{B}$. Then there exists a flow f such that*

$$\partial f = b, \qquad \partial v(b) = \tilde{v}(f).$$

If b is zero, f can be taken as the zero flow. If b is nonzero, f may be expressed as a mixture of segments with endpairs in $\operatorname{supp}^- \{b\} \times \operatorname{supp}^+ \{b\}$.

Proof. We deal first with the case $b = 0$. In this case any flow f such that $\partial f = b$ $(= 0)$ is expressible as a mixture of closed arcs (see Lemma 5.1). By Lemma 6.1, $\tilde{v}(f) \geqq 0$. But then $\tilde{v}$ achieves its minimum value $\partial v(b)$ over flows with zero boundary at the zero flow.

Now suppose that $b \neq 0$. We order the elements in $\operatorname{supp}^+ \{b\}$, $\operatorname{supp}^- \{b\}$ as $\{y_i^+\}$, $\{y_i^-\}$, respectively. Let the positive numbers $\{\alpha_i^+\}$, $\{\alpha_i^-\}$ be defined now through the normal representation of b:

$$b(\phi) = \sum_j \alpha_j^+ \phi(y_j^+) - \sum_i \alpha_i^- \phi(y_i^-), \quad \text{all } \phi \in \mathcal{F}(Y).$$

Let s_{ij} be the segment with endpair (y_i^-, y_j^+). It is easy to see that the set

$$\mathcal{S} = \{\text{flows } s\colon \partial s = b,\ s \text{ is a mixture of the } s_{ij}\text{'s}\}$$

is nonempty (indeed, $(\sum_j \alpha_j^+)^{-1} \sum \sum_{i,j} \alpha_i^- \alpha_i^+ s_{ij}$ is such a flow), and that the mixture $\sum \sum_{i,j} c_{ij} s_{ij}$ lies in $\mathcal{S}$ if and only if

$$\{c_{ij}\} \in P,$$

where

$$P = \{\{c_{ij}\}\colon \sum_j c_{ij} = \alpha_i^-, \sum_i c_{ij} = \alpha_j^+, c_{ij} \geqq 0\}.$$

Now consider

$$\inf \{\tilde{v}(s)\colon s \in \mathcal{S}\}, \tag{6.3}$$

which, in view of the foregoing, may be written

$$\inf \left\{ \tilde{v}\left(\sum_{ij} c_{ij} s_{ij} \right)\colon \{c_{ij}\} \in P \right\} = \inf \left\{ \sum_{i,j} c_{ij} \tilde{v}(s_{ij})\colon \{c_{ij}\} \in P \right\}.$$

If the value of the expression (6.3) is $+\infty$, then the infimum is achieved by any $s \in \mathcal{S}$ and, in particular, on some s with boundary b. If the value is finite, then the infimum is again achieved since its determination reduces in effect to minimization of a linear function over the nonempty compact polyhedron in the finite dimensional linear space of matrices (of a certain, fixed, dimension):

$$P \cap \{\{c_{ij}\}\colon c_{ij} = 0 \text{ if } \tilde{v}(s_{ij}) = +\infty\}.$$

We are interested in

$$\partial v(b) = \inf\{v(s): \partial s = b\}. \tag{6.4}$$

We have just shown that the infimum in (6.4) is achieved if the constraint $\partial s = b$ is supplemented by the requirement that s is a mixture of the elements s_{ij} each of which has endpair in $\operatorname{supp}^-\{b\} \times \operatorname{supp}^+\{b\}$. The proof will be complete then if we can that, given a flow f with $\partial f = b$, there exists $s \in \mathscr{S}$ such that

$$\tilde{v}(f) \geqq \tilde{v}(s),$$

for then the supplementary constraint can be added without reducing the infimum. This we proceed to show.

Let f be such that $\partial f = b$. By Lemma 6.2,

$$f = \sum_i \alpha_i g_i + \sum_i \beta_i h_i, \tag{6.5}$$

where the α_i's, β_i's are nonnegative, the g_i's are arcs with ends in $\operatorname{supp}^-\{b\} \times \operatorname{supp}^+\{b\}$ and the h_i' are closed arcs. Now define $\bar{g}_i$, $i = 1, 2, \cdots$ to be segments with endpairs those of g_i, $i = 1, 2, \cdots$, and set

$$s = \sum_i \alpha_i \bar{g}_i.$$

Clearly, $s \in \mathscr{S}$. In terms of the representation (6.5), $\tilde{v}(f)$ may be expressed

$$\tilde{v}(f) = \sum_i \alpha_i \tilde{v}(g_i) + \sum_i \beta_i \tilde{v}_i(h_i).$$

However, by Lemma 5.1

$$\tilde{v}(h_i) \geqq 0 \quad \text{and} \quad \tilde{v}(g_i) \geqq \tilde{v}(\bar{g}_i).$$

It follows that

$$\tilde{v}(f) \geqq \sum_i \alpha_i \tilde{v}(g_i) \geqq \sum_i \alpha_i \tilde{v}_i(\bar{g}_i) = \tilde{v}(s).$$

We have exhibited an element s in $\mathscr{S}$ with the desired properties. □

Lemma 6.2 specialized to boundaries of segments gives:

COROLLARY 6.1. *Let v be an admissible value function taking value nowhere $-\infty$ and let b be the boundary of a segment with endpair (y^-, y^+). Then*

$$\partial v(b) = v(y^-, y^+).$$

LEMMA 6.3. *Let v be an admissible value function taking values nowhere $-\infty$. Then, for b, b_1, $b_2 \in \mathscr{B}$,*

(i) $\partial v(b) > -\infty$,

(ii) $\partial v(0) = 0$,

(iii) $\partial v(\alpha b) = \alpha \partial v(b)$ for $\alpha > 0$, *and*

(iv) $\partial v(b_1 + b_2) \leqq \partial v(b_1) + \partial v(b_2)$.

Proof. (i) follows from our assumption that $\tilde{v}$ cannot take value $-\infty$ and the fact that the infimum in (6.2) is achieved (see Lemma (6.2)). (ii) and (iii) are direct consequences of the definitions of ∂v. Finally, to prove (iv), we choose f_1, f_2 such that

$$\partial v(b_1) = \tilde{v}(f_1), \quad \partial f_1 = b_1, \qquad \partial v(b_2) = \tilde{v}(f_2), \quad \partial f_2 = b_2$$

(this is permissible by Lemma 6.2). Then

$$\begin{aligned} \partial v(b_1 + b_2) = \inf\{\tilde{v}(f): \partial f = b_1 + b_2\} &\leqq \tilde{v}(f_1 + f_2) \\ &= \tilde{v}(f_1) + \tilde{v}(f_2) = \partial v(b_1) + \partial v(b_2). \end{aligned}$$

□

7. A topology on the space of boundaries and the associated dual space. In this section, we use the metric m on Y.

As we have observed (§ 2), m is an admissible value function. m therefore induces a function ∂m on $\mathcal{B}$ (see § 6) which, we recall, may be written

$$\partial m(b) = \inf\{\tilde{m}(f): \partial f = b\}$$

in which $\tilde{m}$ is the "extension" of m to the class of flows.

LEMMA 7.1. *∂m defines a norm on $\mathcal{B}$.*

Proof. ∂m takes values in $[0, \infty)$ since m takes values in $[0, \infty)$. By Lemma 6.3 and the fact that ∂m is finite valued, we have that $\partial m(b_1 + b_2) \leqq \partial m(b_1) + \partial m(b_2)$, and $\partial m(\alpha b) = \alpha \partial m(b)$ for all $\alpha \geqq 0$. In order to verify that ∂m is a norm, it remains to show that $\partial m(b) > 0$ when $b \neq 0$. However, if $b \neq 0$, there exists a mixture of segments $\sum_i \alpha_i s_i$ with endpairs in $\text{supp}^-\{b\} \times \text{supp}^+\{b\}$ such that the α_i's are positive and

$$\partial m(b) = \sum_i \alpha_i \tilde{m}_i(s_i)$$

(see Lemma 5.2). Since $\text{supp}^-\{b\}$ and $\text{supp}^+\{b\}$ are disjoint, and since m is a metric, it follows that the $\tilde{m}(s_i)$'s are positive. We conclude that $\partial m(b) > 0$ as required. □

The norm ∂m on $\mathcal{B}$ is called the boundary norm and is written $|\cdot|_{\mathcal{B}}$. We now identify Lip as the topological dual of $\mathcal{B}$ with boundary norm (Lip was defined in § 4).

PROPOSITION 7.1. *A linear functional l on $\mathcal{B}$ is continuous with respect to the norm $|\cdot|_{\mathcal{B}}$ if and only if there exists some $\phi \in \text{Lip}$ such that*

$$l(b) = \sum_i \alpha_i [\phi(y_i^+) - \phi(y_i^-)], \tag{7.1}$$

where the right-hand side is expressed in terms of α_i, (y_i^-, y_i^+), $i = 1, 2, \cdots$, taken from any representation of b of the form

$$b(\psi) = \sum_i \alpha_i [\psi(y_i^+) - \psi(y_i^-)], \quad \text{all } \psi \in \mathcal{F}(Y). \tag{7.2}$$

Proof. Suppose first that l is a continuous linear functional. Then there exists some nonnegative number K such that

$$|l(b)| \leqq K|b|_{\mathcal{B}} \quad \text{for all } b \in \mathcal{B}.$$

Let $y^* \in Y$ be an arbitrary point. We now define

$$\phi(y) = l(\delta(y) - \delta(y^*)), \quad \text{all } y \in Y.$$

Here and below $\delta(y)$ denotes the functional taking value $\phi(y)$ on $\phi \in \mathcal{F}(Y)$.

We observe that, for $y, y' \in Y$,

$$\begin{aligned} |\phi(y) - \phi(y')| &= |l(\delta(y) - \delta(y^*) - \delta(y') + \delta(y^*))| \\ &= |l(\delta(y) - \delta(y'))| \leqq K|\delta(y) - \delta(y')|_{\mathcal{B}} = Km(y, y'). \end{aligned}$$

The final equality follows from Corollary 6.1. We have shown $\phi \in \text{Lip}$.

Notice also that, if $b = \sum_i \alpha_i(\delta(y_i^+) - \delta(y_i^-))$, then

$$\begin{aligned} l(b) &= l\left(\sum_i \alpha_i(\delta(y_i^+) - \delta(y_i^-))\right) \\ &= l\left(\sum_i \alpha_i[\delta(y_i^+) - \delta(y^*)] - \sum_i \alpha_i[\delta(y_i^-) - \delta(y^*)]\right) \end{aligned}$$

$$= \sum_i \alpha_i l(\delta(y_i^+) - \delta(y^*)) - \sum_i \alpha_i l(\delta(y_i^+) - \delta(y^*))$$

$$= \sum_i \alpha_i [\phi(y_i^+) - \phi(y_i^-)].$$

We see that $l(\cdot)$ is expressed in terms of ϕ as stated.

This deals with the necessity of the condition for continuity of l. We now prove sufficiency. Suppose that l is defined by (7.1), in which $\phi \in \text{Lip}$. According to Lemma 6.2, we may suppose that the representation (7.2) is so chosen that

$$|b|_{\mathcal{B}} = \sum_i \alpha_i m(y_i^-, y_i^+).$$

From (7.1) then

$$|l(b)| \leqq \sum_i \alpha_i |\phi(y_i^+) - \phi(y_i^-)| \leqq \bar{K} \sum_i \alpha_i m(y_i^-, y_i^+) = \bar{K} |b|_{\mathcal{B}},$$

in which $\bar{K}$ is the Lipschitz constant of ϕ. It follows that l is continuous. □

8. Proof of the main result.

LEMMA 8.1. *Let v be an admissible value function which is calm at $(y_*^-, y_*^+) \in Y \times Y$ and which has modulus of calmness c. Let b_0 be the boundary of the segment with endpair (y_*^-, y_*^+). Then for any $b \in \mathcal{B}$*

$$\partial v(b_0 + b) - \partial v(b_0) \geqq -c|b|_{\mathcal{B}}.$$

Proof. Choose a flow f (in $\mathcal{F}'(Y \times Y)$), which we express as a mixture of segments

$$f = \sum_i c_i \sigma(a_i^-, a_i^+) \tag{8.1}$$

such that

$$\partial v(b_0 + b) = \tilde{v}(f) \tag{8.2}$$

and

$$\partial f = b_0 + b. \tag{8.3}$$

In (8.1), and throughout this proof, $\sigma(y^-, y^+)$ denotes the segment with endpair (y^-, y^+).

Choose also a flow f_1 (in $\mathcal{F}'(Y \times Y)$), expressed as a mixture of segments

$$f_1 = \sum_{i=1}^{K} d_i \sigma(b_i^-, b_i^+) \tag{8.4}$$

such that

$$|b|_{\mathcal{B}} = \tilde{m}(f_1) \tag{8.5}$$

and

$$\partial f_1 = b. \tag{8.6}$$

Such choices of f and f_1 are possible by Lemma 6.2.

Define $Y^* = Y \times \mathbb{R}$. We embed Y in Y^*: $y \to (y, 0)$. Viewing now the a_i^-, a_i^+, b_i^-, b_i^+'s of (8.1) and (8.4) as embedded in the larger space, we may regard f and f_1, given by (8.1) and (8.4), as flows in $\mathcal{F}'(Y^* \times Y^*)$. Let $e_1, \cdots, e_K$ be distinct points in

$Y^*\backslash(Y\times\{0\})$. We define a flow $\tilde{f}_1$ in $\mathscr{F}'(Y^*\times Y^*)$ with boundary $-b$:

$$\tilde{f}_1=\sum_{i=1}^{K} d_i[\sigma(b_i^+, e_i)+\sigma(e_i, b_i^-)] \tag{8.7}$$

(comparing (8.4) and (8.7), we see that $\tilde{f}_1$ is obtained from f_1 by replacing segments by "reversed two-arcs").

Since $\partial f=b_0+b$ and $\partial\tilde{f}_1=-b$, it follows that

$$\partial(f+\tilde{f}_1)=b_0. \tag{8.8}$$

Let us for the time being make the assumption

$$b_0\neq 0. \tag{8.9}$$

Then, by Lemma 5.2, the flow $f+\tilde{f}_1$ may be expressed

$$f+\tilde{f}_1=f_0+f_c. \tag{8.10}$$

In (8.10) f_0 is a mixture with positive coefficients of arcs with endpairs (y_*^-, y_*^+). f_0 can be so chosen because f_0, a flow with nonzero boundary b_0, is nonzero. The flow f_c is a mixture with positive coefficients of closed arcs (or f_c is the zero flow). We assume f_c is nonzero. The case f_c is zero is treated by deleting all terms below associated with f_c; the conclusions will be the same.

Now the constituent elements of $\tilde{f}_1$ are two-arcs with distinct midpoints. The set of these midpoints is in $Y^*\backslash(Y\times\{0\})$ and is therefore disjoint from the set $\{b_i^-, b_i^+, i=1, 2, \cdots\}$ of endpoints of the two-arcs, from the set of endpoints of the segments providing the representation (8.1) of f, and from y_*^-, y_*^+, since all these points are in $Y\times\{0\}$. Consequently these two-arcs cannot be broken up in the representation (8.10) of $f+\tilde{f}_1$ as a mixture with positive coefficients of arcs which are either closed or have endpairs (y_*^-, y_*^+). It follows that f_0, f_c can be expressed

$$f_0=\sum_i \alpha_i\left(\sum_j \bar{g}_{ij}\right), \tag{8.11}$$

$$f_c=\sum_i \beta_i\left(\sum_i \bar{\bar{g}}_{ij}\right). \tag{8.12}$$

In (8.11) and (8.12) each $\bar{g}_{ij}$, $\bar{\bar{g}}_{ij}$ is either a segment which is a constituent segment of f, or a two-arc of the form $\sigma(b_k^+, e_k)+\sigma(e_k, b_k^-)$ for some k. Further, for each i, $\{\bar{g}_{ij}, j=1, 2, \cdots\}$ defines an arc with endpair (y_*^-, y_*^+) and $\{\bar{\bar{g}}_{ij}, j=1, 2, \cdots\}$ defines a closed arc. The coefficients $\{\alpha_i\}$, $\{\beta_i\}$ are all positive.

By (8.8), (8.10), (8.11) and (8.12), we have, for ϕ taking value 1 at y_*^+ and zero elsewhere,

$$\sum_i \alpha_i=\partial(f+f_1)(\phi)=1. \tag{8.13}$$

For each i define

$$\bar{S}_i=\{j:\ \bar{g}_{ij} \text{ is a segment}\}, \qquad \bar{\bar{S}}_i=\{j:\ \bar{\bar{g}}_{ij} \text{ is a segment}\}.$$

Since the $\bar{g}_{ij}$'s which are segments can only arise as constituent segments of f and since the $\bar{g}_{ij}$'s which are two-arcs can only arise as constituent two-arcs of $\tilde{f}_1$ (the same is true of the $\bar{\bar{g}}_{ij}$'s), we may separate out contributions of f and $\tilde{f}_1$ in the expressions

(8.11) and (8.12). We obtain representations of f and $\tilde{f}_1$ as follows:

$$f=\sum_i \alpha_i\Big(\sum_{j\notin\bar{S}_i}\bar{g}_{ij}\Big)+\sum_i \beta_i\Big(\sum_{j\in\bar{\bar{S}}_i}\bar{\bar{g}}_{ij}\Big), \tag{8.14}$$

$$\tilde{f}_1=\sum_i \alpha_i\Big(\sum_{j\notin\bar{S}_i}\bar{g}_{ij}\Big)+\sum_i \beta_i\Big(\sum_{j\notin\bar{\bar{S}}_i}\bar{\bar{g}}_{ij}\Big). \tag{8.15}$$

Now the flow f_1 is recovered from $\tilde{f}_1$ by replacing two-arcs $\bar{g}_{ij}$, $\bar{\bar{g}}_{ij}$ in (8.15) by segments having endpairs the reverse of the endpairs of the two-arcs (see (8.4) and (8.7)). If we define $\bar{h}_{ij}=\sigma(b_k^+, b_k^-)$ when $\bar{g}_{ij}$ is a two-arc $\bar{g}_{ij}=\sigma(b_k^-, e_k)+\sigma(e_k, b_k^+)$, some k, and define $\bar{\bar{h}}_{ij}$ similarly in relation to $\bar{\bar{g}}_{ij}$, we may write

$$f_1=\sum_i \alpha_i\Big(\sum_{j\notin\bar{S}_i}\bar{h}_{ij}\Big)+\sum_i \beta_i\Big(\sum_{j\notin\bar{\bar{S}}_i}\bar{\bar{h}}_{ij}\Big). \tag{8.16}$$

Notice that the $\bar{g}_{ij}$'s, for $j\in\bar{S}_i$ and the $\bar{h}_{ij}$'s have endpoints in $Y\times\{0\}$ (an analogous property holds for the $\bar{\bar{g}}_{ij}$'s, $\bar{\bar{h}}_{ij}$'s). In fact, all flows considered in the remainder of this proof are expressible through segments with endpoints in $Y\times\{0\}$ (the flow $\tilde{f}_1$, of which this was not true, was merely a device for setting up the representations (8.14) and (8.16)). We revert, therefore, to considering flows as elements in $\mathscr{F}'(Y\times Y)$.

Equations (8.14) and (8.16) may be interpreted as follows: The flow f is expressible as a mixture of arcs which are either closed or have endpairs (y_*^-, y_*^+) and from which have been removed certain subarcs. The mixture associated with the subarcs, on reversal of these subarcs, provides a representation of f_1. (In the case of an arc in this mixture with endpair (y_*^-, y_*^+), we may assume that subarcs have been removed which incorporate initial and concluding segments of this arc; this may be arranged by including, if necessary, additional trivial segments of the form $\sigma(y_*^-, y_*^-)$, $\sigma(y_*^+, y_*^+)$ in the representation of f_1. These are trivial in the sense that $\tilde{v}$, $\tilde{m}$ take value zero on them. In the case of a closed arc, we may arrange, again by introduction of trivial segments, that at least one subarc is removed).

These conclusions may be stated as follows:

There exist positive integers $\bar{K}_i$, $i=1,2,\cdots,\bar{K}$ and $\bar{\bar{K}}_i$, $i=1,2,\cdots,\bar{\bar{K}}$; for $i=1,\cdots,\bar{K}$ there exist points $\{(y_{ij}^-, y_{ij}^+)\}_{j=1}^{\bar{K}_i}$ in $Y\times Y$ and collections of arcs $\{p_{ij}\}_{j=1}^{\bar{K}_i}$, $\{r_{ij}\}_{j=0}^{\bar{K}_i}$ such that

p_{ij} has endpair (y_{ij}^-, y_{ij}^+), $j=1,\cdots,\bar{K}_i$;

r_{ij} has endpair $(y_{i(j+1)}^-, y_{ij}^+)$, $j=1,\cdots,\bar{K}_i-1$;

r_{i0} has endpair (y_{i1}^-, y_*^-), $r_{i\bar{K}_i}$ has endpair $(y_*^+, y_{i\bar{K}_i}^+)$; and for $i=1,\cdots,\bar{\bar{K}}$ there exist points $\{(z_{ij}^-, z_{ij}^+)\}_{j=1}^{\bar{\bar{K}}_i}$ in $Y\times Y$, and collections of arcs $\{q_{ij}\}_{j=1}^{\bar{\bar{K}}_i}$, $\{s_{ij}\}_{j=1}^{\bar{\bar{K}}_i}$ such that

q_{ij} has endpair (z_{ij}^-, z_{ij}^+), $j=1,\cdots,\bar{\bar{K}}_i$;

s_{ij} has endpair $(z_{i(j+1)}^-, z_{ij}^+)$, $j=1,\cdots,\bar{\bar{K}}_i-1$;

$s_{i\bar{\bar{K}}_i}$ has endpair $(z_{i1}^-, z_{i\bar{\bar{K}}_i}^+)$;

with the following properties:

f may be expressed as

$$f=\sum_{i=1}^{\bar{K}}\alpha_i\Big(\sum_{j=1}^{\bar{K}_i}p_{ij}\Big)+\sum_{i=1}^{\bar{\bar{K}}}\beta_i\Big(\sum_{j=1}^{\bar{\bar{K}}_i}q_{ij}\Big). \tag{8.17}$$

Also f_1, modified possibly by addition of a mixture of "trivial segments", may be expressed as

$$f_1=\sum_{i=1}^{\bar{K}}\alpha_i\Big(\sum_{j=0}^{\bar{K}_i}r_{ij}\Big)+\sum_{i=1}^{\bar{\bar{K}}}\beta_i\Big(\sum_{j=1}^{\bar{\bar{K}}_i}s_{ij}\Big). \tag{8.18}$$

The coefficients α_i, β_i in (8.17), (8.18) are the same as in (8.14), (8.16).

Recall the properties (8.2), (8.3), (8.5), (8.6) of f and f_1, which may now be written in terms of the representations (8.17), (8.18):

$$\partial f = b_0 + b, \tag{8.19}$$

$$\partial v(b_0+b) = \sum_{i=1}^{\bar{K}} \alpha_i \sum_{j=1}^{\bar{K}_i} \tilde{v}(p_{ij}) + \sum_{i=1}^{\hat{K}} \beta_i \sum_{j=1}^{\hat{K}} \tilde{v}(q_{ij}), \tag{8.20}$$

$$\partial f_1 = b, \tag{8.21}$$

$$|b|_{\mathscr{B}} = \sum_{i=1}^{\bar{K}} \alpha_i \sum_{j=0}^{\bar{K}_i} \tilde{m}(r_{ij}) + \sum_{i=1}^{\hat{K}} \beta_i \sum_{j=1}^{\hat{K}_i} \tilde{m}(s_{ij}). \tag{8.22}$$

We may replace the flows f and f_1 by new flows in which the arcs $\{p_{ij}\}$, $\{q_{ij}\}$, $\{r_{ij}\}$, $\{s_{ij}\}$ are substituted by segments with the same endpairs. The new f, f_1 so obtained still satisfy (8.19), (8.21) since the changes do not affect the endpairs of the constituent arcs and, therefore, the boundaries, of f, f_1. Furthermore the changes cannot increase the values of the right-hand sides of (8.20), (8.21) since, for example, $\tilde{v}(p_{ij}) \geqq v(y^-, y^+)$, where (y^-, y^+) is the endpair of p_{ij} (see Corollary 6.1). It follows that the new f, f_1 still satisfy (8.20), (8.22) since, by the nature of the definition of $\partial v(b_0+b)$, $|b|_{\mathscr{B}}$ and by (8.19), (8.21), the values of the right-hand sides cannot be decreased.

After making these changes and on noting that, by Corollary 6.1, $\tilde{v}(\sigma(y^-, y^+)) = v(y^-, y^+)$, $\tilde{m}(\sigma(y^-, y^+)) = m(y^- m\, y^+)$ for $(y^-, y^+) \in Y \times Y$, we may write (8.20), (8.22) in terms of y_{ij}^-, y_{ij}^+, etc.:

$$\partial v(b_0+b) = \sum_{i=1}^{\bar{K}} \alpha_i \left[\sum_{j=1}^{\bar{K}_i} v(y_{ij}^-, y_{ij}^+) \right] + \sum_{i=1}^{\hat{K}} \beta_i \left[\sum_{j=1}^{\hat{K}} v(z_{ij}^-, z_{ij}^+) \right], \tag{8.23}$$

$$\begin{aligned} |b|_{\mathscr{B}} = {} & \sum_{i=1}^{\bar{K}} \alpha_i \left[\sum_{j=1}^{\bar{K}_i - 1} m(y_{i(j+1)}^-, y_{ij}^+) + m(y_{i1}^-, y_*^-) + m(y_*^+, y_{i\bar{K}_i}^+) \right] \\ & + \sum_{i=1}^{\hat{K}} \beta_i \left[\sum_{j=1}^{\hat{K}_i - 1} m(y_{i(j+1)}^-, y_{ij}^+) + m(y_{i1}^-, y_{i\hat{K}_i}^+) \right]. \end{aligned} \tag{8.24}$$

We have, by (8.13) and Corollary 6.1,

$$\partial v(b_0) = \left(\sum_i \alpha_i \right) \partial v(b_0) = \left(\sum_i \alpha_i \right) v(y_*^-, y_*^+).$$

Subtracting this equation from (8.23), we obtain

$$\partial v(b_0+b) - \partial v(b_0) = \sum_{i=1}^{\bar{K}} \alpha_i \left[\sum_{j=1}^{\bar{K}_i} v(y_{ij}^-, y_{ij}^+) - v(y_*^-, y_*^+) \right] + \sum_{i=1}^{\hat{K}} \beta_i \left[\sum_{j=1}^{\hat{K}_i} v(z_{ij}^-, z_{ij}^+) \right]. \tag{8.25}$$

However, since v is calm at (y_*^-, y_*^+) with modulus c, we have for each i

$$\begin{aligned} & \left[\sum_{j=1}^{\bar{K}_i} v(y_{ij}^-, y_{ij}^+) \right] - v(y_*^-, y_*^+) \\ & \qquad \geqq -c \left[\sum_{j=1}^{\bar{K}_i - 1} m(y_{i(j+1)}^-, y_{ij}^+) + m(y_{i1}^-, y_*^-) + m(y_*^+, y_{i\bar{K}_i}^+) \right] \end{aligned} \tag{8.26}$$

and

$$\sum_{j=1}^{\hat{K}_i} v(z_{ij}^-, z_{ij}^+) \geqq -c \left[\sum_{j=1}^{\hat{K}_i - 1} m(y_{i(j+1)}^-, y_{ij}^+) + m(y_{i1}^-, y_{i\hat{K}_i}^+) \right]. \tag{8.27}$$

It follows from (8.24), (8.25), (8.26) and (8.27) that

$$\partial v(b_0+b)-\partial v(b_0)\geqq -c|b|_{\mathscr{B}}. \tag{8.28}$$

Recall that (8.28) has been proved under the assumption (see (8.9)) that $b_0\neq 0$. We now sketch the proof of (8.28) when $b_0=0$. In this case $f+f_1$ is a closed flow (see (8.8)). We may assume that $f+f_1$ is a nonzero flow; otherwise, f_1 is also zero, b then is zero (by (8.6)) and our assertion (8.26) obviously holds. By Lemma 5.1 the closed, nonzero flow can be expressed as a mixture, with positive coefficients, of closed arcs or, in other words, as (8.10) in which the term f_0 is absent. We now follow through the previous arguments but making no reference to the now absent f_0. There results (8.24) and (8.25) in which the coefficients α_i are all zero. We conclude (8.28) from these equations as before. □

Proof of Theorem 4.1. In view of the results in § 4, it remains to show that strong calmness is a sufficient condition for existence of $\phi\in$ Lip satisfying (4.2), (4.3) and that, if v has modulus of calmness c, then ϕ may be chosen with Lipschitz constant not greater than c. Suppose that v is strongly calm at (y_*^-, y_*^+) with modulus of calmness c. By definition of strong calmness, $v(y_*^-, y_*^+)$ is finite. By (3.3) v takes values $-\infty$ nowhere.

Let b_0 be the boundary of the segment with endpair (y_*^-, y_*^+). By Corollary 6.1, $\partial v(b_0)$ is $v(y_*^-, y_*^+)$ and is therefore finite.

Consider the subsets in the linear space $\mathbb{R}\times\mathscr{B}$ with product norm

$$S_1=\{(\alpha, b)\colon \alpha\geqq\partial v(b_0+b)-\partial v(b_0)\}, \tag{8.29}$$

$$S_2=\{(\alpha, b)\colon -\alpha\geqq c|b|_{\mathscr{B}}\}. \tag{8.30}$$

We deduce from Lemma 6.3 that S_1 is a convex cone with apex the origin. S_2 too is a convex cone with apex the origin.

Now the interior of S_2 is nonempty and disjoint from S_1 by Lemma 8.1. The sets S_1 and S_2 may be separated then by a continuous linear functional on $\mathbb{R}\times\mathscr{B}$. There exist therefore an element l in the topological dual of $\mathscr{B}$ and a number α_0, not both zero, such that

$$\alpha_0\alpha-l(b)\geqq 0, \qquad (\alpha, b)\in S_1, \tag{8.31}$$

$$\alpha_0\alpha-l(b)\leqq 0, \qquad (\alpha, b)\in S_2. \tag{8.32}$$

The separation property may be expressed in this manner since the origin is common to S_1 and S_2.

From (8.30) and (8.32), we deduce that

$$-l(b)\leqq\alpha_0 c|b|_{\mathscr{B}} \quad \text{for all } b\in\mathscr{B}. \tag{8.33}$$

Since $\mathscr{B}$ is a linear space and l is linear, (8.33) implies

$$|l(b)|\leqq\alpha_0 c|b|_{\mathscr{B}} \quad \text{for all } b\in\mathscr{B}. \tag{8.34}$$

It is clear from this inequality that α_0 must be positive, for otherwise, α_0 and l would both be zero. We may assume then that l and α_0 have been scaled so that $\alpha_0=1$.

(8.34) may now be written

$$|l(b)|\leqq c|b|_{\mathscr{B}} \quad \text{for all } b\in\mathscr{B}. \tag{8.35}$$

We have from (8.29) and (8.31)

$$l(b)\leqq\partial v(b_0+b)-\partial v(b_0) \quad \text{for all } b\in\mathscr{B}.$$

We readily deduce from the positive homogeneity of ∂v (see Lemma 6.3) that

$$l(b_0) = \partial v(b_0), \tag{8.36}$$

$$l(b) \leqq \partial v(b) \quad \text{for all } b \in \mathcal{B}. \tag{8.37}$$

Now let $\phi \in \text{Lip}$ represent l (we refer to § 7). Suppose that $b \in \mathcal{B}$ is the boundary of a segment with endpair the arbitrary point (y^-, y^+) in $Y \times Y$. According to Proposition 7.1,

$$l(b) = \phi(y^+) - \phi(y^-). \tag{8.38}$$

However, by Corollary 6.1, $\partial v(b) = v(y^-, y^+)$ and $|b|_{\mathcal{B}} = m(y^-, y^+)$. We conclude then from (8.36), (8.37) and (8.38) that

$$\phi(y^+) - \phi(y^-) \leqq v(y^-, y^+) \quad \text{for all } (y^-, y^+) \in Y \times Y,$$

$$\phi(y_*^+) - \phi(y_*^-) = v(y_*^-, y_*^+).$$

From (8.35) we have

$$|\phi(y^+) - \phi(y^-)| \leqq cm(y^-, y^+) \quad \text{for all } (y^-, y^+) \in Y \times Y,$$

or in other words, ϕ has Lipschitz constant at most c. □

Acknowledgment. The author greatly benefitted from discussions with J.-P. Aubin, F. Clarke and A. D. Ioffe in the course of this research.

REFERENCES

[1] C. CARATHÉODORY, *Calculus of Variations and Partial Differential Equations of the First Order*, R. B. Dean and J. J. Brandstatter, trans., Holden-Day, San Francisco, 1965.

[2] F. H. CLARKE, *Extremal arcs and extended Hamiltonian systems*, Trans. Amer. Math. Soc., 231 (1977), pp. 349–367.

[3] ———, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.

[4] R. L. GONZALES, *Sur l'existence d'une solution maximale de l'equation de Hamilton-Jacobi*, CR Acad. Sci., 282 (1976), pp. 1287–1290.

[5] A. D. IOFFE, *Convex functions ocurring in variational problems and the absolute minimum problem*, Mat. Sb, 88 (1972), pp. 194–210 = Math. USSR-Sb (1972), pp. 191–208.

[6] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.

[7] V. F. KROTOV AND V. I. GURMAN, *Metody i Zadachi Optimal'novo Upravlenija*, Nauka, Moscow, 1973. (In Russian.)

[8] V. L. LEVIN AND A. A. MILYUTIN, *The problem of mass transfer with a discontinuous cost function and a mass statement of the duality theorem for convex extremal problems*, Uspekhi Mat. Nauk, 34, 3 (1979), pp. 3–68 = Russian Math. Surveys, 34 (1978), pp. 1–78.

[9] R. M. LEWIS, *Dynamic programming, a functional analytic foundation*, J. Math. Anal. Appl., to appear.

[10] R. B. VINTER, *New global optimality conditions in optimal control theory*, this Journal, this issue, pp. 235–245.

[11] ———, *The equivalence of "strong calmness" and "calmness" in optimal control theory*, J. Math. Anal. Appl., to appear.

[12] R. B. VINTER AND R. M. LEWIS, *A necessary and sufficient condition for optimality of dynamic programming type, making no a priori assumptions on the controls*, this Journal, 16 (1978), pp. 571–583.

[13] J. WARGA, *Optimal Control of Differential and Functional Differential Equations*, Academic Press, New York, 1972.

[14] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2101-0005 $01.25/0

NEW GLOBAL OPTIMALITY CONDITIONS IN OPTIMAL CONTROL THEORY*

R. B. VINTER†

Abstract. We give global optimality conditions expressed in terms of a function ϕ which satisfies conditions related to the Hamilton–Jacobi equation. Thus our results are in the spirit of sufficient conditions for optimality associated with Carathéodory in the calculus of variations, and of the verification theorems of optimal control theory. The novelty here is that ϕ is permitted to be merely Lipschitz continuous. A weakest hypothesis, strong calmness, is provided under which our results apply. Evidence that the strong calmness hypothesis is a reasonable one is presented elsewhere in the literature.

1. Introduction.

We shall be concerned with the optimal control problem:

$$\text{Minimize} \int_{t_0}^{t_1} l(x(t), t, u(t))\, dt \tag{1.1}$$

subject to

$$\frac{dx}{dt}(t) = f(x(t), t, u(t)) \quad \text{a.e. } t \in [t_0, t_1], \tag{1.2}$$

$$x(t_0) = x_0, \qquad x(t_1) = x_1, \tag{1.3}$$

$$(x(t), t) \in A, \tag{1.4}$$

$$u(\cdot) \in \mathcal{U}.$$

Here $l(\cdot,\cdot,\cdot): \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$, $f(\cdot,\cdot,\cdot): \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^n$ are given functions. (x_0, t_0), (x_1, t_1) are given points in $\mathbb{R}^{n+1}$. A is a subset of $\mathbb{R}^n \times \mathbb{R}$. $\mathcal{U}$ is a subset of functions $u(\cdot): \mathbb{R} \to \mathbb{R}^n$ which is "closed under switching". By this we mean $u_1(\cdot), u_2(\cdot) \in \mathcal{U}$, $t \in \mathbb{R}$ imply that $v(\cdot) \in \mathcal{U}$, where $v(s) = u_1(s)$ for $s < t$ and $v(s) = u_2(s)$ for $s \geqq t$.

In the control problem we seek the infimum of the integral functional (1.1) over admissible processes. By an "admissible process" we mean a pair of functions $(x(\cdot), u(\cdot))$, in which $x(\cdot): [t_0, t_1] \to \mathbb{R}$ is absolutely continuous and $u(\cdot) \in \mathcal{U}$, with the following properties: (1.2)–(1.4) are satisfied and $t \to l(x(t), t, u(t))$ is a Lebesgue measurable function on $[t_0, t_1]$. The integral (1.1) is interpreted as taking value $+\infty$ when $t \to \max\{0, l(x(t), t, u(t))\}$ and $t \to \max\{0, -l(x(t), t, u(t))\}$ are both nonintegrable. Sometimes we emphasize the choice of data (x_0, t_0), (x_1, t_1) by referring to a process as "admissible with respect to (x_0, t_0), (x_1, t_1)". We assume that processes admissible with respect to (x_0, t_0), (x_1, t_1) exist, and that the infimum of the functional (1.1) over such processes is finite.

A process admissible with respect to (x_0, t_0), (x_1, t_1) is termed optimal if it minimizes (1.1) over such processes. The infimum of the functional (1.1) over admissible processes is called the infimum cost.

Consider now a well-known sufficient condition for global optimality, which summarizes a methodology generally associated with the name of Carathéodory. (For results in this spirit, see [1], [15], [6] or [8].)

* Received by the editors January 25, 1982.

† Department of Electrical Engineering, Imperial College of Science and Technology, London SW7 2BT England.

PROPOSITION 1.1. *Let* $(x(\cdot), u(\cdot))$ *be an admissible process.*

(a) *Suppose there exists a continuously differentiable function* $\phi(\cdot,\cdot): \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ *such that*

$$\phi_t(x,t) + \phi_x(x,t) f(x,t,v(t)) \leqq l(x,t,v(t)) \quad \text{for all } (x,t) \in A,\ v(\cdot) \in \mathcal{U} \tag{1.5}$$

and, for almost every $t \in [t_0, t_1]$,

$$\phi_t(x(t),t) + \phi_x(x(t),t) f(x(t),t,u(t)) = l(x(t),t,u(t)). \tag{1.6}$$

Then $(x(\cdot), u(\cdot))$ *is optimal.*

(b) *If a continuously differentiable function* ϕ *exists such that* (1.5) *and* (1.6) *are satisfied, then* ϕ *solves the maximization problem*:

Maximize $\psi(x_1,t_1) - \psi(x_0,t_0)$ *over continuously differentiable functions* ψ *subject to*

$$\psi_t(x,t) + \psi_x(x,t) f(x,t,v(t)) \leqq l(x,t,v(t)) \quad \text{for all } (x,t) \in A,\ v(\cdot) \in \mathcal{U}.$$

Part (a) of Proposition 1.1 provides the sufficient condition in terms of a function ϕ. Part (b) identifies the function ϕ as the solution of a certain infinite dimensional linear programming problem.

Proof of Proposition 1.1 is elementary.

While this approach has been employed in the solution of some specific optimal control problems (see, e.g., [6], [9] and [15]), no conditions appear to be known, of an unrestrictive nature, assuring the existence of a continuously differentiable function ϕ characterizing optimal controls as in Proposition 1.1. In fact optimal control problems may easily be devised, of a nature such that we should not like to exclude them from consideration (see, e.g., [14, Example 5.2]) which illustrate how Proposition 1.1 can fail to apply. These problems suggest that the essential limitation in Proposition 1.1 is the smoothness required of the function ϕ.

The difficulty may be circumvented by giving a modified optimality condition in terms of a sequence of continuously differentiable functions [13] (such a condition is implicit too in the duality results of Klötzler [7] and Levin and Milyutin [10]) or, at the price of destroying the local nature of condition (1.5), by expressing the condition in terms of a continuous function [5]. Such results lack precision however, and the question arises whether there exists some class of functions $\mathcal{C}$ larger than the class of continuously differentiable ones, with the following properties:

Firstly, an optimality condition having the flavor of Proposition 1.1 can be given in terms of some $\phi \in \mathcal{C}$.

Secondly, the modified optimality condition applies under reasonable hypotheses.

Our purpose in the present paper is to show that the class of Lipschitz continuous functions meets the first requirement. We derive optimality conditions involving a Lipschitz continuous function ϕ in which either the constraints (1.5) are required merely to hold on the dense subset of the domain of ϕ on which ϕ is differentiable, or (1.5) is replaced by an analogous constraint expressed in terms of the generalized directional derivatives of ϕ.

A weakest hypothesis is presented under which the new optimality conditions apply. The hypothesis, a kind of regularity condition on the infimum cost with respect to data perturbations, we term "strong calmness".

Is strong calmness a reasonable hypothesis? Evidence is given in [12] that strong calmness is indeed reasonable and, therefore, that the class of Lipschitz continuous functions meets the second requirement on the class $\mathcal{C}$ demanded above. Under mild assumptions, concerning Lipschitz continuity of the data in their x-dependence and boundedness of the underlying domains, strong calmness is equivalent (in a sense

made precise in [12]) to calmness. "Calmness", a notion introduced by Clarke (see, e.g., [3]), appears to be the weakest available condition assuring normality of the multipliers arising from application of the Pontryagin maximum principle to our optimal control problem. Thus we can expect our optimality condition to apply whenever the Pontryagin maximum principle yields nontrivial information.

We conjecture also that our results will be significant as regards numerical computation of the infimum cost. We identify the Lipschitz continuous function appearing in our optimality conditions as the solution of an infinite dimensional linear programming problem similar to that in part (b) of Proposition 1.1. This problem lends itself to solution by finite element methods along the lines of those reported in [4] for free-endpoint problems. The fact that the solution is a Lipschitz continuous function would lead us to expect order one convergence for linear elements.

The question concerning weakest conditions under which a result similar to Proposition 1.1 may be given in terms of a Lipschitz continuous function was answered in a general context in [11] (this issue, pp. 215–235). The main result in [11], as it relates to the problem of interest here, is stated in § 4. This paper may be seen as a justification of the developments in [11] in that we show how the results in [11] specialize to give refinements of familiar optimality conditions in optimal control theory.

2. The value function. We define the value function $\eta(\cdot,\cdot)\colon \mathbb{R}^{n+1}\times\mathbb{R}^{n+1}\to \mathbb{R}\cup\{+\infty\}\cup\{-\infty\}$ as

$$\eta((\xi_0,\tau_0),(\xi_1,\tau_1)) = \inf\left\{\int_{\tau_0}^{\tau_1} l(x(s),s,u(s))\,ds\right\},$$

where the infimum is taken over processes admissible with respect to (ξ_0,τ_0), (ξ_1,τ_1) when such exist. Otherwise the value of η is interpreted as $+\infty$.

It is a simple matter to check the following properties of η:

PROPOSITION 2.1.

(i) $\eta(y,y)=0$ *for all* $y\in\mathbb{R}^{n+1}$.

(ii) $\eta(y_1,y_2)+\eta(y_2,y_3)\geqq\eta(y_1,y_3)$ *for all* $y_1,y_2,y_3\in\mathbb{R}^{n+1}$.

(iii) $\eta((x_0,t_0),(x_1,t_1))<\infty$.

3. Strong calmness. The following definition, which enters into the hypotheses imposed in the main results to follow, was first introduced in [11]:

DEFINITION 3.1. The problem (P) is *strongly calm* if there exists a real number c such that for any collection of points in $\mathbb{R}^{n+1}\times\mathbb{R}^{n+1}$

$$\{((\xi_1^-,\tau_1^-),(\xi_1^+,\tau_1^+)),\cdots,((\xi_k^-,\tau_k^-),(\xi_k^+,\tau_k^+))\}$$

we have

$$\begin{aligned}&\sum_{i=1}^{k}\eta((\xi_i^-,\tau_i^-),(\xi_i^+,\tau_i^+))-\eta((x_0,t_0),(x_1,t_1))\\ &\geqq -c\left[\sum_{i=1}^{k-1}|(\xi_{i+1}^-,\tau_{i+1}^-)-(\xi_i^+,\tau_i^+)|+|(x_0,t_0)-(\xi_1^-,\tau_1^-)|+|(\xi_k^+,\tau_k^+)-(x_1,t_1)|\right]\end{aligned}\tag{3.1}$$

and

$$\begin{aligned}&\sum_{i=1}^{k}\eta((\xi_i^-,\tau_i^-),(\xi_i^+,\tau_i^+))\\ &\geqq -c\left[\sum_{i=1}^{k-1}|(\xi_{i+1}^-,\tau_{i+1}^-)-(\xi_i^+,\tau_i^+)|+|(\xi_1^-,\tau_1^-)-(\xi_k^+,\tau_k^+)|\right].\end{aligned}\tag{3.2}$$

The strong calmness condition is discussed, and related to conditions which have arisen elsewhere in the literature in [11]. As an alternative viewpoint on strong calmness, we give the following control theoretic interpretation:

Consider an ordered collection of processes $\{(x_i(\cdot), u_i(\cdot))\}_{i=1}^k$ such that $(x_i(\cdot), u_i(\cdot))$ is admissible with respect to $(\xi_i^-, \tau_i^-), (\xi_i^+, \tau_i^+)$ for $i = 1, \cdots, k$. We may view the whole collection as a single "generalized" process γ in which the trajectory is permitted to jump in $(x, t)-$space. The jumps are from (x_0, t_0) to (ξ_1^-, τ_1^-), from (ξ_1^+, τ_1^+) to $(\xi_2^-, \tau_2^-), \cdots$, and from (ξ_k^+, τ_k^+) to (x_1, t_1). We evaluated the cost of γ as

$$q(\gamma) = \sum_{i=1}^{k} \int_{\tau_i^-}^{\tau_i^+} l(x_i(t), t, u_i(t))\, dt$$

(see Fig. 1).

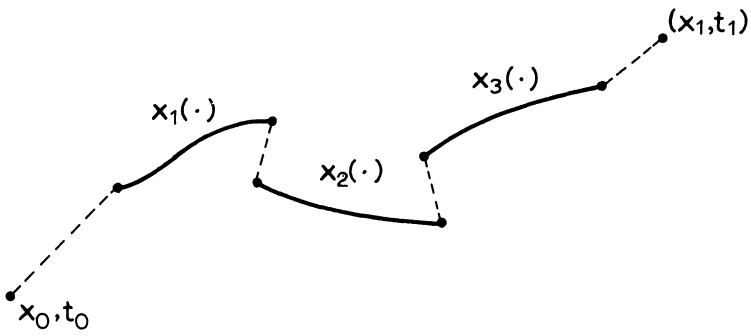


FIG. 1. *The trajectories* $\{x_i(\cdot)\}$ *corresponding to a generalized process.*

Inequality (3.1) in the definition of strong calmness imposes a restriction on the amount by which the infimum cost, inf {P}, can be decreased by admission of generalized processes. Specifically, it is required that there exist a number c such that for any generalized process γ

$$q(\gamma) - \inf\{\mathrm{P}\} > -cd(\gamma).$$

Here $q(\gamma)$ is the cost of γ, as before, and $d(\gamma)$ is the sum of the lengths of the jumps of γ.

The second inequality, (3.2), is usually not significant; we may arrange that it is satisfied under the mild conditions, say, that the function l is bounded below.

As we mentioned in the introduction, the strong calmness condition reduces essentially to Clarke's calmness condition [3] assuring normality of the Pontryagin multipliers, under mild, directly verifiable hypotheses on the data (see [12]).

4. A preliminary result. We summarize in the following theorem the main result in [11] as it bears on our optimal control problem.

Here, and subsequently, Lip (S), the space of Lipschitz continuous functions on a subset $S \subset \mathbb{R}^{n+1}$ is defined to be

$$\mathrm{Lip}\,(S) = \left\{\phi : S \to \mathbb{R} : \sup_{y \neq y'} \left(\frac{|\phi(y) - \phi(y')|}{|y - y'|} \right) < +\infty \right\}.$$

THEOREM 4.1.

$$\eta((x_0, t_0), (x_1, t_1)) = \max\{\phi(x_1, t_1) - \phi(x_0, t_0)\} \tag{4.1}$$

if and only if (P) *is strongly calm. In* (4.1) *the maximum is taken over* $\phi \in \mathrm{Lip}\,(A)$ *such that*

$$(4.2) \qquad \eta((\xi_0, \tau_0), (\xi_1, \tau_1)) \geqq \phi(\xi_1, \tau_1) - \phi(\xi_0, \tau_0) \quad \textit{for all } (\xi_0, \tau_0), (\xi_1, \tau_1) \in A.$$

(*The notation* "max" *in* (4.1) *indicates that the maximum is achieved.*)

The theorem is a special case of [11, Thm 4.1], which applies for any function η having the three properties listed in Proposition 2.1.

Theorem 4.1 expresses the cost as the "value" of a maximization problem in which the underlying elements, $\phi \in \mathrm{Lip}\,(A)$, satisfy the constraint (4.2). It is unsatisfactory as a characterization of the infimum cost insofar as direct verification of (4.2) requires knowledge of η, the function whose value at $((x_0, t_0), (x_1, t_1))$ we are trying to find. The theorem is nonetheless significant as providing the starting point for derivation of related results in which (4.2) is replaced by simple "local" constraints on the ϕ's expressed directly in terms of the problem data. These developments compel us to impose extra assumptions on the problem data. Our purpose in not imposing the extra assumptions from the beginning has been to emphasize the generality of the results in [11]; Theorem 4.1 applies under more or less the weakest assumptions assuring that the optimal control problem makes sense.

5. The main results. We consider the following additional hypotheses: There exists some subset Ω of $\mathbb{R}^n$ such that:

$$(5.1) \qquad \mathcal{U} = \{\text{measurable functions } s \to u(s) : u(t) \in \Omega \text{ a.e.}\}.$$

(5.2) A is an open set.

(5.3) For each $u \in \Omega$, $l(\cdot, \cdot, u) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$, $f(\cdot, \cdot, u) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ are continuous.

THEOREM 5.1-A. *Suppose that* (5.1)–(5.3) *are true. Then*

$$(5.4) \qquad \eta((x_0, t_0), (x_1, t_1)) = \max\{\phi(x_1, t_1) - \phi(x_0, t_0)\}$$

if and only if (P) *is strongly calm. The maximum in* (5.4) *is taken over* $\phi \in \mathrm{Lip}\,(A)$ *such that*

$$\phi_t(x, t) + \phi_x(x, t) f(x, t, u) \leqq l(x, t, u)$$

for all $u \in \Omega$ *and all* $(x, t) \in A$ *at which* ϕ *is differentiable.*

An alternative statement of this result may be given in terms of the "generalized directional derivative" $D^0\phi(y; h)$ of $\phi \in \mathrm{Lip}\,(A)$ at the point $y \in A$ in the direction h, introduced by Clarke in [2]:

$$D^0\phi(y; h) = \limsup_{\substack{y_i \to y \\ \varepsilon_i \downarrow 0}} \{\varepsilon_i^{-1}[\phi(y_i + \varepsilon_i h) - \phi(y_i)]\}.$$

This expression is interpreted as the lowest upper bound over all sequences $\{y_i\}, \{\varepsilon_i\}$ with $y_i \to y$, $\varepsilon_i \downarrow 0$ of the limit superior of the bracketed expression.

THEOREM 5.1-B. *Suppose that* (5.1), (5.2) *and* (5.3) *are true. Then*

$$(5.5) \qquad \eta((x_0, t_0), (x_1, t_1)) = \max\{\phi(x_1, t_1) - \phi(x_0, t_0)\}$$

if and only if (P) *is strongly calm. The maximum in* (5.5) *is taken over* $\phi \in \mathrm{Lip}\,(A)$ *such that*

$$D^0\phi((x, t); (f(x, t, u), 1)) - l(x, t, u) \leqq 0$$

for all $(x, t, u) \in A \times \Omega$.

The preceding results characterize the infimum cost and apply even when this infimum is not attained. The final two results concern optimal processes.

THEOREM 5.2. *Suppose that* (5.1), (5.2) *and* (5.3) *are true. Let* $(x(\cdot), u(\cdot))$ *be an admissible process.*

(i) *If* (P) *is strongly calm then* $(x(\cdot), u(\cdot))$ *is optimal if and only if there exists* $\phi \in \text{Lip}(A)$ *such that*

$$\frac{d}{dt}\phi(x(t), t) = l(x(t), t, u(t)) \quad \text{a.e. } t \in [t_0, t_1], \tag{5.6}$$

$$D^0\phi((x, t); (f(x, t, u), 1)) - l(x, t, u) \leqq 0 \quad \textit{for all } (x, t, u) \in A \times \Omega. \tag{5.7}$$

(ii) *If* (P) *is not strongly calm, then no* $\phi \in \text{Lip}(A)$ *exists satisfying* (5.6) *and* (5.7).

Of course ϕ satisfying (5.6) and (5.7), if it exists, is given by any ϕ achieving the maximum in (5.4) or (5.5).

Finally we state a variant on Theorem 5.2 in which a greater unity between conditions (5.6) and (5.7) is achieved by replacing the derivative of the composite function $t \to \phi(x(t), t)$ in (5.6) by an expression involving the generalized directional derivative of ϕ; the price we pay for this refinement is that we retain only necessity of the optimality condition.

THEOREM 5.3. *Suppose that* (5.1) *and* (5.2) *and* (5.3) *are true and that* (P) *is strongly calm. Then there exists* $\phi \in \text{Lip}(A)$ *with the following property*:

$$D^0\phi((x, t); (f(x, t, u), 1)) - l(x, t, u) \leqq 0$$

for all $(x, t, u) \in A \times \Omega$ *and, if* $(x(\cdot), u(\cdot))$ *is an optimal process, then*

$$D^0\phi((x(t), t); f(x(t), t, u(t)), 1) - l(x(t), t, u(t)) = 0 \quad \text{a.e. } t \in [t_0, t_1].$$

The function ϕ *may be taken as any* ϕ *achieving the maximum in* (5.4) *or* (5.5).

6. Connections with the maximum principle. Let $(x^*(\cdot), u^*(\cdot))$ be an optimal process.

For the purpose of illustrating the relationship between our optimality conditions and the Pontryagin maximum principle let us suppose that the problem considered is such that the conclusions of Theorem 5.2 apply with ϕ a twice continuously differentiable function.

We readily deduce from (5.6) and (5.7), and the assumption that ϕ is twice continuously differentiable, that

$$\phi_t(x^*(t), t) + \phi_x(x^*(t), t) f(x^*(t), t, u^*(t)) - l(x^*(t), t, u^*(t)) = 0 \quad \text{a.e. } t \in [t_0, t_1]$$

and

$$\phi_t(x, t) + \phi_x(x, t) f(x, t, u) - l(x, t, u) \leqq 0 \quad \text{all } (x, t) \in A, u \in \Omega.$$

These properties imply that, for almost every $t \in [t_0, t_1]$,

$$\begin{aligned} &x \to \phi_t(x, t) + \phi_x(x, t) f(x, t, u^*(t)) - l(x, t, u^*(t)) \\ &\text{is maximized over } \{x : (x, t) \in A\} \text{ at } x^*(t) \end{aligned} \tag{6.1}$$

and

$$\begin{aligned} &u \to \phi_x(x^*(t), t) f(x^*(t), t, u) - l(x^*(t), t, u) \\ &\text{is maximized over } \Omega \text{ at } u^*(t). \end{aligned} \tag{6.2}$$

Now suppose that, for almost every $t \in [t_0, t_1]$, $x^*(t)$ is interior to $\{x : (x, t) \in A\}$ and that $x \to l(x, t, u^*(t))$, $x \to f(x, t, u^*(t))$ are differentiable functions at $x^*(t)$.

We deduce from (6.1) that

$$\phi_{tx}(x^*(t), t)+\phi_{xx}(x^*(t), t)f(x^*(t), t, u^*(t)) + \phi_x(x^*(t), t)f_x(x^*(t), t, u^*(t))-l_x(x^*(t), t, u^*(t))=0 \quad \text{a.e. } t\in[t_0, t_1].$$

However

$$\frac{d}{dt}\phi_x(x^*(t), t)=\phi_{xx}(x^*(t), t)f(x^*(t), t, u^*(t))+\phi_{tx}(x^*(t), t) \quad \text{a.e. } t\in[t_0, t_1],$$

whence

$$\frac{dp(t)}{dt}+p(t)f_x(x^*(t), t, u^*(t))-l_x(x^*(t), t, u^*(t))=0 \quad \text{a.e. } t\in[t_0, t_1], \tag{6.3}$$

where the absolutely continuous function $p(\cdot)$ is given by

$$p(t)=\phi_x(x^*(t), t).$$

We note that (6.2) may be written

$$\begin{aligned} &u\to p(t)f_x(x^*(t), t, u)-l(x^*(t), t, u)\\ &\text{is maximized over } \Omega \text{ at } u^*(t), \quad \text{a.e. } t\in[t_0, t_1]. \end{aligned} \tag{6.4}$$

Equations (6.3) and (6.4) will be recognized as a statement of the Pontryagin maximum principle. We may loosely interpret our optimality conditions then as variants on the Pontryagin maximum principle in which a function ϕ replaces the costate variable $p(\cdot)$ and a maximization property with respect to the x-variable (6.1) replaces the costate differential equation (6.3). We caution the reader though against pushing this interpretation too far. The Pontryagin maximum principle, which is a necessary condition for "local" optimality, and Theorem 5.2, which (under the strong calmness hypothesis) is a necessary and sufficient condition for "global" optimality, have very different characters. We have suggested connections only under conditions, namely existence of twice continuously differentiable function ϕ with suitable properties, which are highly restrictive and difficult to test.

7. Proof of the main results.

LEMMA 7.1. *Suppose that hypotheses* (5.1)–(5.3) *are true. Let* (a, b) *be an open interval containing the number* t *and let* x *be a point in* $\mathbb{R}^n$. *Let* $(x(\cdot), u(\cdot))$ *be a process admissible with respect to* $(x(a), a), (x(b), b)$ *and such that* $x(t)=x$. *Let* ϕ *be a Lipschitz continuous function defined on a neighborhood of the graph of* $x(\cdot)$. *Suppose that* t *is a Lebesgue point of* $s\to f(x(s), s, u(s))$. *Then the limit, as* $\alpha\downarrow 0$, *of*

$$\alpha^{-1}[\phi(x(t+\alpha), t+\alpha)-\phi(x, t)] \tag{7.1}$$

exists if and only if the limit, as $\alpha\downarrow 0$, *of*

$$\alpha^{-1}[\phi(x+\alpha f(x(t), t, u(t)), t+\alpha)-\phi(x, t)] \tag{7.2}$$

exists. The limits, if they exist, are the same.

Proof. Write $d(\alpha)$ for the difference of (7.1) and (7.2). Then

$$|d(\alpha)|\leqq K\left|\alpha^{-1}\int_t^{t+\alpha} f(x(s), s, u(s))\,ds-f(x, t, u(t))\right|$$

for α sufficiently small. Here K is the local Lipschitz constant. Since t is a Lebesgue point of $s\to f(x(s), s, u(s))$, the limit of the right-hand side is zero, as $\alpha\downarrow 0$. The assertions of the lemma follow. □

LEMMA 7.2. *Suppose that hypotheses* (5.1)–(5.3) *are true. Let* $\phi \in \text{Lip}(A)$. *Then the properties* (i), (ii) *and* (iii) *below are equivalent.*

(i) $\phi(y_1)-\phi(y_0) \leqq \eta(y_0, y_1)$ *for all* $y_0, y_1 \in A$.

(ii) $\phi_t(x, t)+\phi_x(x, t)f(x, t, u) \leqq l(x, t, u)$ *for all* $u \in \Omega$ *and* $(x, t) \in A$ *at which* ϕ *is differentiable.*

(iii) $D^0\phi((x, t); (f(x, t, u), 1)) \leqq l(x, t, u)$ *for all* $(x, t) \in A$, $u \in \Omega$.

Proof. (i) *implies* (ii): Assume that (i) is satisfied. Let $(x, t) \in A$ be a point at which ϕ is differentiable, and let $u \in \Omega$ be given. By the standard existence theory governing solutions to ordinary differential equations, there exist some positive number ε and a process $(x(\cdot), u)$ admissible with respect to $(x(t-\varepsilon), t-\varepsilon)$, $(x(t+\varepsilon), t+\varepsilon)$ such that $x(t) = x$. Since $s \to f(x(s), s, u)$ is continuous, t is certainly a Lebesgue point of $s \to f(x(s), s, u)$. Property (i) implies that, for $\alpha > 0$ sufficiently small,

$$\alpha^{-1}[\phi(x(t+\alpha), t+\alpha)-\phi(x, t)] \leqq \alpha^{-1}\int_t^{t+\alpha} l(x(s), s, u)\, ds. \tag{7.3}$$

The right-hand side has limit $l(x, t, u)$, as $\alpha \downarrow 0$, since $s \to l(x(s), s, u)$ is continuous. Now

$$\lim_{\alpha \downarrow 0} \alpha^{-1}[\phi(x+f(x, t, u)\alpha, t+\alpha)-\phi(x, t)] = \phi_x(x, t)f(x, t, u)+\phi_t(x, t)$$

since, by assumption, ϕ is differentiable at (x, t). The limit of the left-hand side of (7.3) also exists and takes this value, by Lemma 7.1. We conclude that

$$\phi_t(x, t)+\phi_x(x, t)f(x, t, u) \leqq l(x, t, u)$$

as required.

(ii) *implies* (iii): Fix $u \in \Omega$. It will be convenient temporarily to set $y = (x, t)$, and for us to suppress the u-dependence in the notation and write $\tilde{f}(y)$, $\tilde{l}(y)$ for $f(x, t, u)$, $l(x, t, u)$. As is known [2], for any $y \in A$

$$D^0\phi(y; \tilde{f}(y)) = \limsup_{y_i \to y} \{\phi_y(y_i)\tilde{f}(y)\}.$$

The right-hand side is interpreted as the supremum of all accumulation points of the bracketed term, over sequences of points $\{y_i\}$ at which ϕ is differentiable, converging to y.

But then

$$D^0\phi(y; \tilde{f}(y)) = \limsup_{y_i \to y} \{\phi_y(y_i)\tilde{f}(y_i)\} \tag{7.4}$$

since, for any sequence $\{y_i\}$ as above, the difference of the bracketed terms is $\phi_y(y_i) \cdot (\tilde{f}(y)-\tilde{f}(y_i))$; this difference has limit zero by the continuity of $\tilde{f}$ and the boundedness of $\{\phi_y(y_i)\}$. Now suppose (ii). Condition (ii), (7.4) and the continuity of $\tilde{l}$ imply

$$D^0\phi(y; f(y)) \leqq \limsup_{y_i \to y} \{\tilde{l}(y_i)\} = \tilde{l}(y).$$

Thus (iii) is true.

(iii) *implies* (i): Suppose (iii) and, in contradiction of our claim, that for some pair $(x(\cdot), u(\cdot))$ admissible with respect to (ξ_0, τ_0), (ξ_1, τ_1) we have

$$\phi(\xi_1, \tau_1)-\phi(\xi_0, \tau_0) > \int_{\tau_0}^{\tau_1} l(x(s), s, u))\, ds.$$

The function $t \to \phi(x(t), t)$ is absolutely continuous since, by assumption, ϕ is Lipschitz continuous and $x(\cdot)$ is absolutely continuous. It follows that $(d/dt)\phi(x(t), t)$ exists for almost every $t \in [\tau_0, \tau_1]$ and

$$\int_{\tau_0}^{\tau_1} \frac{d}{dt}\phi(x(t), t)\, dt > \int_{\tau_0}^{\tau_1} l(x(s), s, u(s))\, ds. \tag{7.5}$$

We conclude from (7.5) that we can find $t \in (\tau_0, \tau_1)$ at which $d(\phi(x(t), t))/dt$ exists, which is a Lebesgue point for $s \to l(x(s), s, u(s))$ and is such that

$$\frac{d}{dt}\phi(x(t), t) > l(x(t), t, u(t)).$$

Now

$$\frac{d}{dt}\phi(x(t), t) = \lim_{\alpha \downarrow 0} \alpha^{-1}\left[\phi\left(x(t) + \int_t^{t+\alpha} f(x(s), s, u(s))\, ds, t+\alpha\right) - \phi(x, t)\right].$$

Using Lemma 7.1 to evaluate this limit, we obtain

$$\lim_{\alpha \downarrow 0} \alpha^{-1}[\phi(x(t) + \alpha f(x(t), t, u(t)), t+\alpha) - \phi(x(t), t)] > l(x(t), t, u(t)).$$

But then, by definition of the generalized directional derivative,

$$D^0\phi((x(t), t); (f(x(t), t, u(t)), 1)) > l(x(t), t, u(t))$$

this contradicts (iii). Thus (i) must be true. □

LEMMA 7.3. *Suppose that hypotheses* (5.1)–(5.3) *are true. Let* $(x(\cdot), u(\cdot))$ *be admissible with respect to* (x_0, t_0), (x_1, t_1) *and let* $\phi \in \mathrm{Lip}(A)$ *be such that*

$$D^0\phi((x, t); f(x, t, u)) \leqq l(x, t, u) \quad \text{for all } (x, t) \in A, \quad u \in \Omega. \tag{7.6}$$

Consider the assertions

$$\phi(x_1, t_1) - \phi(x_0, t_0) = \int_{t_0}^{t_1} l(s(t), t, u(t))\, dt, \tag{7.7}$$

$$\frac{d}{dt}\phi(x(t), t) = l(x(t), t, u(t)), \quad \text{a.e. } t \in [t_0, t_1], \tag{7.8}$$

$$D^0\phi((x(t), t), f(x(t), t, u(t))) = l(x(t), t, u(t)), \quad \text{a.e. } t \in [t_0, t_1]. \tag{7.9}$$

Then

(i) (7.7) *is true if and only if* (7.8) *is true,*

and

(ii) (7.9) *is true if* (7.8) *is true.*

Proof. Note first of all that $t \to \phi(x(t), t)$ is absolutely continuous. The function is therefore differentiable almost everywhere and satisfies

$$\phi(x_1, t_1) - \phi(x_0, t_0) = \int_{t_0}^{t_1} \frac{d}{dt}\phi(x(t), t)\, dt.$$

We see immediately that (7.8) implies (7.7).

Now suppose that (7.7) is true. Then

$$\int_{t_0}^{t_1} \left[\frac{d}{dt}\phi(x(t), t) - l(x(t), t, u(t))\right] dt = 0. \tag{7.10}$$

We shall show that (7.8) and (7.9) are true.

The subset of points in (t_0, t_1) at which $s \to \phi(x(s), s)$ is differentiable, and which are Lebesgue points for $s \to l(x(s), s, u(s))$ and $s \to f(x(s), s, u(s))$, is of full measure. Choose such a point t.

We may deduce from (7.6), (7.10) and Lemma 7.2 that

$$\begin{aligned} \frac{d}{dt}\phi(x(t), t) &= \lim_{\alpha \downarrow 0} \alpha^{-1}[\phi(x(t+\alpha), t+\alpha) - \phi(x(t), t)] \\ &= \lim_{\alpha \downarrow 0} \alpha^{-1}\left[\int_t^{t+\alpha} l(x(s), s, u(s))\, ds\right] \\ &= l(x(t), t, u(t)). \end{aligned} \tag{7.11}$$

By Lemma 7.1 and the definite of the generalized directional derivative

$$\begin{aligned} &\lim_{\alpha \downarrow 0} \alpha^{-1}[\phi(x(t+\alpha), t+\alpha) - \phi(x(t), t)] \\ &= \lim_{\alpha \downarrow 0} \alpha^{-1}[\phi(x(t) + \alpha f(x(t), t, u(t)), t+\alpha) - \phi(x(t), t)] \\ &\leqq D^0\phi((x(t), t); f(x(t), t, u(t))). \end{aligned} \tag{7.12}$$

From (7.6), (7.11) and (7.12) we conclude that (7.8) and (7.9) are true.

Finally we turn our attention to the theorems of § 5. We see that Theorems 5.1-A and 5.1-B are immediate consequences of Theorem 4.1 and Lemma 7.2. Theorem 5.2 follows from Theorem 5.1-B and part (i) of Lemma 7.3 while Theorem 5.3 follows from Theorem 5.1-B and part (ii) of Lemma 7.3.

REFERENCES

[1] C. CARATHÉODORY, *Calculus of Variations and Partial Differential Equations of the First Order*, Vol. 2, R. B. Dean, trans., Holden-Day, San Francisco, 1967.

[2] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–267.

[3] ———, *The generalized problem of Bolta*, this Journal, 14 (1976), pp. 165–174.

[4] R. L. GONZALEZ, *On the solution of the Hamilton-Jacobi equation of deterministic control*, Thèse de 3e cycle, Univ. Paris IX–Dauphine, 1980.

[5] A. D. IOFFE, *Convex functions occurring in variational problems and the absolute minimum problem*, Mat. Sb., 88 (1972), pp. 194–210 = Math. USSR–Sb. (1972), pp. 191–208.

[6] A. D. IOFFE AND V. M. TIHOMIROV, *Theory of Extremal Problems*, North-Holland Amsterdam, 1979.

[7] R. KLOTZER, *Strong duality in control theory*, technical report, Karl-Marx University, Leipzig.

[8] V. F. KROTOV, *Methods for solving variational problems on the basis of sufficient conditions for an absolute minimum*, I, II, III, Automation and Remote Control (1962), pp. 1473–1484; (1963), pp. 539–553; (1964), pp. 924–933.

[9] V. F. KROTOV AND V. I. GURMAN, *Metody i Zadachi Optimal 'novo Upravlenija*, Nauka, Moscow, 1973. (In Russian.)

[10] V. L. LEVIN AND A. A. MILYUTIN, *The problem of mass transfer with a discontinuous cost function, and a mass statement of the duality theorem for convex extremal problems*, Uspekhi Mat. Nauk, 34, 3 (1979), pp. 3–68 = Russian Math. Surveys, 34, 3 (1978), pp. 1–78.

[11] R. B. VINTER, *Weakest conditions for existence of Lipschitz conditions Krotov functions in optimal control theory*, this Journal, this issue, pp. 235–245.

[12] ———, *The equivalence of "strong calmness" and "calmness" in optimal control theory*, J. Math. Anal. Appl., to appear.

[13] R. B. VINTER AND R. M. LEWIS, *A necessary and sufficient condition for optimality of dynamic programming type, making no a priori assumptions on the control*, this Journal 16, (1978), pp. 571–583.

[14] R. B. VINTER AND R. M. LEWIS, *A verification theorem which provides a necessary and sufficient condition for optimality*, IEEE Trans. Automat. Control, AC-25, 1 (1980), pp. 84–89.

[15] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0006 $01.25/0

A NEW CLASS OF STABILIZING CONTROLLERS FOR UNCERTAIN DYNAMICAL SYSTEMS*

B. R. BARMISH†, M. CORLESS‡ AND G. LEITMANN‡

Abstract. This paper is concerned with the problem of designing a stabilizing controller for a class of uncertain dynamical systems. The vector of uncertain parameters $q(\cdot)$ is time-varying, and its values $q(t)$ lie within a prespecified bounding set Q in R^p. Furthermore, no statistical description of $q(\cdot)$ is assumed, and the controller is shown to render the closed loop system "practically stable" in a so-called *guaranteed sense*; that is, the desired stability properties are assured no matter what admissible uncertainty $q(\cdot)$ is realized. Within the perspective of previous research in this area, this paper contains one salient feature: the class of stabilizing controllers which we characterize is shown to include linear controllers when the nominal system happens to be linear and time-invariant. In contrast, in much of the previous literature (see, for example, [1], [2], [7], and [9]), a linear system is stabilized via nonlinear control. Another feature of this paper is the fact that the methods of analysis and design do not rely on transforming the system into a more convenient canonical form; e.g., see [3]. It is also interesting to note that a linear stabilizing controller can sometimes be constructed even when the system dynamics are nonlinear. This is illustrated with an example.

Key words. stability, uncertain dynamical systems, guaranteed performance

1. Introduction.

During recent years, a number of papers have appeared which deal with the design of stabilizing controllers for uncertain dynamical systems; e.g., see [1]–[7]. In these papers the uncertain quantities are described only in terms of bounds on their possible sizes; that is, no statistical description is assumed. Within this framework, the objective is to find a class of controllers which *guarantee* "stable" operation for *all* possible variations of the uncertain quantities.

Roughly speaking, the results to date fall into two categories. There are those results which might appropriately be termed *structural* in nature; e.g., see [1]–[3], [6]. By this we mean that the uncertainty cannot enter arbitrarily into the state equations; certain preconditions must be met regarding the locations of the uncertainty within the system description. Such conditions are often referred to as *matching assumptions*. We note that in this situation uncertainties can be tolerated with an arbitrarily large prescribed bound. A second body of results might appropriately be termed *nonstructural* in nature; e.g., see [4] and [5]. Instead of imposing matching assumptions on the system, these authors permit more general uncertainties at the expense of "sufficient smallness" assumptions on the allowable sizes of the uncertainties.

This work falls within the class of structural results mentioned above. Our motivation comes from a simple observation. Namely, given a theory which yields stabilizing controllers for a class of uncertain nonlinear systems, it is often desirable for this theory to have the following property: upon specializing the "recipe" for controller construction from nonlinear to linear systems, one of the possible stabilizing control laws should be linear in form. It is of importance to note that existing results do *not* have this property. Upon specialization to the linear case, one typically obtains controllers of the discontinuous "bang-bang" variety; e.g., see [1] and [2]. One can often approximate these controllers using a so-called saturation nonlinearity; e.g., see

* Received by the editors March 4, 1981, and in revised form January 15, 1982.

† Department of Electrical Engineering, University of Rochester, Rochester, New York 14627. The work of this author was supported by the U.S. Department of Energy under contract no. ET-78-S-01-3390.

‡ Department of Mechanical Engineering, University of California at Berkeley, Berkeley, California 94720. The work of these authors was supported by the National Science Foundation under grant ENG 78-13931.

[7]. Such an approach leads to uniform ultimate boundedness of the state to an arbitrarily small neighborhood of the origin; this type of behavior might be termed *practical stability*.[1]

Our desire in this paper is to develop a controller which is linear when the system dynamics are linear. By taking known results (such as in [3]) which were developed exclusively for linear systems, one encounters a fundamental difficulty when attempting to generalize[2] to a class of nonlinear systems; namely, it is no longer possible to transform the system dynamics to a more convenient canonical form. The subsequent analysis is free of such transformations.

2. Systems, assumptions and the concept of practical stability. We consider an uncertain dynamical system described by the state equation

$$\begin{aligned}\dot{x}(t) &= f(x(t), t) + \Delta f(x(t), q(t), t) \\ &\quad + [B(x(t), t) + \Delta B(x(t), q(t), t)]u(t),\end{aligned} \tag{2.1}$$

where $x(t) \in R^n$ is the *state*, $u(t) \in R^m$ is the *control*, $q(t) \in R^p$ is the *uncertainty* and $f(x, t)$, $\Delta f(x, q, t)$, $B(x, t)$ and $\Delta B(x, q, t)$ are matrices of appropriate dimensions which depend on the structure of the system. Furthermore, it is assumed that the uncertainty, $q(\cdot): R \to R^p$, is Lebesgue measurable and its values $q(t)$ lie within a prespecified *bounding set* $Q \subset R^p$ for all $t \in R$. We denote this by writing $q(\cdot) \in M(Q)$.

As mentioned in the introduction, given that "stabilization" is the goal, we must impose additional conditions on the manner in which $q(t)$ enters structurally into the state equations. We refer to such conditions as *matching assumptions*.

Assumption 1. There are mappings

$$h(\cdot): R^n \times R^p \times R \to R^m \quad \text{and} \quad E(\cdot): R^n \times R^p \times R \to R^{m \times m}$$

such that

$$\begin{aligned}&\Delta f(x, q, t) = B(x, t)h(x, q, t), \\ &\Delta B(x, q, t) = B(x, t)E(x, q, t), \\ &\|E(x, q, t)\| < 1\end{aligned}$$

for all $x \in R^n$, $q \in Q$ and $t \in R$.

We note that this assumption can sometimes be weakened. For example, in [9] a certain *measure of mis-match* is introduced and results are obtained under the proviso that this measure does not exceed a certain critical level termed the *mis-match threshold*.

Our second assumption reflects the fact that the uncertainties must be bounded in order to permit one to guarantee stability.

Assumption 2. The set $Q \subset R^p$ is compact.

Our next assumption is introduced to guarantee the existence of solutions of the state equations.

Assumption 3. The mappings $f(\cdot): R^n \times R \to R^n$, $B(\cdot): R^n \times R \to R^{m \times n}$, $h(\cdot)$ and $E(\cdot)$ (see Assumption 1) are continuous.[3]

---

[1] This notion is not to be interpreted in the sense of Lasalle and Lefschetz [12] but as defined subsequently.

[2] That is, one begins with a linear control law for a linear system and generalizes the controller in such a way that is applies to a class of nonlinear systems.

[3] In fact, one can modify the analysis to follow so as to allow mappings which are Carathéodory and satisfy certain integrability conditions. See, for example, Corless and Leitmann [7]. All the results of this paper still hold under this weakening of hypotheses.

In order to satisfy our final assumption, one may need to "precompensate" the so-called nominal system, that is, the system with $\Delta f(x, q, t) \equiv 0$ and $\Delta B(x, q, t) \equiv 0$; e.g., see [2]. Thus, prior to controlling the effects of the uncertainty, it may be necessary to employ a portion of the control to obtain an *uncontrolled nominal system*

$$\text{(UC)} \qquad \dot{x}(t) = f(x(t), t)$$

that has certain stability properties embodied in the next assumption.

Assumption 4. $f(0, t) = 0$ for all $t \in R$ and, moreover, there exist a C^1 function $V(\cdot): R^n \times R \to [0, \infty)$ and strictly increasing continuous functions $\gamma_1(\cdot)$, $\gamma_2(\cdot)$, $\gamma_3(\cdot): [0, \infty) \to [0, \infty)$ satisfying[4] $\gamma_1(0) = \gamma_2(0) = \gamma_3(0) = 0$ and $\lim_{r\to\infty} \gamma_1(r) = \lim_{r\to\infty} \gamma_2(r) = \lim_{r\to\infty} \gamma_3(r) = \infty$, such that for all $(x, t) \in R^n \times R$,

$$\gamma_1(\|x\|) \leqq V(x, t) \leqq \gamma_2(\|x\|). \tag{2.2}$$

Moreover, defining the Lyapunov derivative $\mathscr{L}_0(\cdot): R^n \times R \to R$ by

$$\mathscr{L}_0(x, t) \triangleq \frac{\partial V(x, t)}{\partial t} + \nabla_x' V(x, t) f(x, t), \tag{2.3}$$

where ∇_x' denotes the transpose of the gradient operation, we also require that

$$\mathscr{L}_0(x, t) \leqq -\gamma_3(\|x\|)$$

for all pairs $(x, t) \in R^n \times R$. This assumption, in effect, asserts the existence of a Lyapunov function for the uncontrolled nominal system (UC). Consequently, the origin, $x = 0$, is a uniformly asymptotically stable equilibrium point for the uncontrolled nominal system (UC).

The stability concept employed in this paper differs slightly from the traditional Lyapunov-type stability. To motivate this change of definition, consider the following very simple example of a system satisfying (2.1) and the associated assumptions: $\dot{x}(t) = x(t) + q(t) + u(t)$, with initial condition $x(t_0) = 1$ and uncertainty $q(\cdot)$ such that $|q(t)| \leqq 1$. Furthermore, suppose the control is a linear feedback of the form $u(t) = kx(t)$, with $k < -1$. Then, if a state $x(t) < -1/(1+k)$ is reached, an admissible uncertainty $q(t) \equiv 1$ results in motion of the state away from zero. Hence, although we cannot guarantee uniform asymptotic stability (using a finite gain), we can nevertheless drive the state to an arbitrarily small neighborhood of the origin.[5] The following uniform ultimate boundedness-type definition captures this notion.

DEFINITION 1. The uncertain dynamical system (2.1) is said to be *practically stabilizable* if, given any $\underline{d} > 0$, there is a control law $p_{\underline{d}}(\cdot): R^n \times R \to R^m$ for which, given any admissible uncertainty $q(\cdot) \in M(Q)$, any initial time $t_0 \in R$ and any initial state $x_0 \in R^n$, the following conditions hold:

(i) The closed loop system

$$\begin{aligned} \dot{x}(t) = {} & f(x(t), t) + \Delta f(x(t), q(t), t) \\ & + [B(x(t), t) + \Delta B(x(t), q(t), t)] p_{\underline{d}}(x(t), t) \end{aligned} \tag{2.4}$$

possesses a solution $x(\cdot): [t_0, t_1] \to R^n$, $x(t_0) = x_0$.

---

[4] The limit condition on $\gamma_3(\cdot)$ can in fact be removed at the expense of a somewhat more technical development; e.g., see [7].

[5] This is not to be confused with Lyapunov stability, because the required gain k depends on the size of the neighborhood to which we wish to drive the state.

(ii) Given any $r>0$ and any solution $x(\cdot):[t_0, t_1]\to R^n$, $x(t_0)=x_0$, of (2.4) with $\|x_0\|\leqq r$, there is a constant $d(r)>0$ such that

$$\|x(t)\|\leqq d(r) \quad \text{for all} \quad t\in[t_0, t_1].$$

(iii) Every solution $x(\cdot):[t_0, t_1]\to R^n$ can be continued over $[t_0, \infty)$.

(iv) Given any $\bar{d}\geqq \underline{d}$, any $r>0$ and any solution $x(\cdot):[t_0, \infty)\to R^n$, $x(t_0)=x_0$, of (2.4) with $\|x_0\|\leqq r$, there exists a finite time $T(\bar{d}, r)<\infty$, possibly dependent on r but not on t_0, such that $\|x(t)\|\leqq \bar{d}$ for all $t\geqq t_0+T(\bar{d}, r)$.

(v) Given any $\bar{d}\geqq \underline{d}$ and any solution $x(\cdot):[t_0, \infty)\to R^n$, $x(t_0)=x_0$, of (2.4), there is a constant $\delta(\bar{d})>0$ such that $\|x_0\|\leqq \delta(\bar{d})$ implies that

$$\|x(t)\|\leqq \bar{d} \quad \text{for all} \quad t\geqq t_0.$$

3. Controller construction. We take $\underline{d}>0$ as given and proceed to construct a control law $p_{\underline{d}}(\cdot)$ which will later be shown to satisfy conditions (i)–(v) in the definition of practical stabilizability.

Construction of $p_{\underline{d}}(\cdot)$. The first step is to select functions $\Delta_1(\cdot)$ and $\Delta_2(\cdot):R^n\times R\to R$ satisfying

$$\Delta_1(x, t)\geqq \max_{q\in Q}\|h(x, q, t)\|, \tag{3.1}$$

$$1>\Delta_2(x, t)\geqq \max_{q\in Q}\|E(x, q, t)\|. \tag{3.2}$$

The standing Assumptions 1–4 assure that there is a $\Delta_2(x, t)$ such that

1) $\Delta_2(x, t)<1$ can be satisfied for all $(x, t)\in R^n\times R$;

2) $\Delta_1(\cdot)$ and $\Delta_2(\cdot)$ can be chosen to be continuous; e.g., see [10, p. 116].

Now, one simply selects any continuous function $\gamma(\cdot):R^n\times R\to[0, \infty)$ satisfying

$$\gamma(x, t)\geqq \frac{\Delta_1^2(x, t)}{4[1-\Delta_2(x, t)][C_2-C_1\mathscr{L}_0(x, t)]}, \tag{3.3}$$

where C_1 and C_2 are any (designer chosen) nonnegative constants such that

(3.4)

a) $C_1<1$;

b) either $C_1\neq 0$ or $C_2\neq 0$;

c) $C_2\neq 0$ whenever $\lim_{x\to 0}[\Delta_1^2(x, t)/\mathscr{L}_0(x, t)]$ does not exist;

d) $$\frac{C_2}{1-C_1}<(\gamma_3\circ\gamma_2^{-1}\circ\gamma_1)(\underline{d}).$$

Note that these conditions can indeed be satisfied because of continuity of the $\gamma_i(\cdot)$ and the fact that $\lim_{r\to 0}\gamma_i(r)=0$ for $i=1, 2, 3$.

This construction then enables one to let

$$p_{\underline{d}}(x, t)\triangleq -\gamma(x, t)B'(x, t)\nabla_x V(x, t). \tag{3.5}$$

Remark. In fact, (3.3) and (3.5) describe a class of controllers yielding practical stability. It will be shown in § 5 that this class includes linear controllers when the nominal system happens to be linear and time-invariant.

4. Main result and stability estimates. The theorem below and its proof differ from existing results (see [1], [2] and [6]) in one fundamental way: The control $p_{\underline{d}}(\cdot)$ which leads to the satisfaction of the conditions for practical stabilizability degenerates

into a linear controller whenever the *nominal* system, obtained by setting $\Delta f(x(t), q(t), t) \equiv 0$ and $\Delta B(x(t), q(t), t) \equiv 0$ in (2.1), is linear and time-invariant. This will be demonstrated in the sequel. In fact, even for certain nonlinear nominal systems, the controller turns out to be linear. This phenomenon will be illustrated with an example of a nonlinear pendulum. Central to the proof of the theorem below is one fundamental concept: a system satisfying Assumptions 1–4 admits a control such that the Lyapunov function for the nominal system (UC) is also a Lyapunov function for the uncertain system (2.1).

THEOREM 1. *Subject to Assumptions* 1–4, *the uncertain dynamical system* (2.1) *is practically stabilizable.*

Proof. For a given $\underline{d} > 0$ and a given uncertainty $q(\cdot) \in M(Q)$, the Lyapunov derivative $\mathscr{L}(\cdot): R^n \times R \to R$ for the closed loop system obtained with the feedback control (3.5) is given by

$$\begin{aligned} \mathscr{L}(x, t) \triangleq \mathscr{L}_0(x, t) + \nabla_x' V(x, t)\{&\Delta f(x, q(t), t) \\ &+ [B(x, t) + \Delta B(x, q(t), t)] p_{\underline{d}}(x, t)\}. \end{aligned} \tag{4.1}$$

By using the matching assumptions in conjunction with (3.5), (4.1) becomes

$$\begin{aligned} \mathscr{L}(x, t) = \mathscr{L}_0(x, t) &- \gamma(x, t)\|B'(x, t)\nabla_x V(x, t)\|^2 \\ &+ \nabla_x' V(x, t) B(x, t)[h(x, q(t), t) \\ &- \gamma(x, t) E(x, q(t), t) B'(x, t) \nabla_x V(x, t)]. \end{aligned}$$

Letting $\phi(\cdot): R^n \times R \to R^m$ be given by

$$\phi(x, t) \triangleq B'(x, t)\nabla_x V(x, t),$$

and recalling the definition of $\Delta_1(\cdot)$ and $\Delta_2(\cdot)$, a straightforward computation yields

$$\begin{aligned} \mathscr{L}(x, t) \leqq \mathscr{L}_0(x, t) &- [1 - \Delta_2(x, t)]\gamma(x, t)\|\phi(x, t)\|^2 \\ &+ \Delta_1(x, t)\|\phi(x, t)\|. \end{aligned}$$

Now there are two cases to consider.

Case 1. The pair (x, t) is such that $\Delta_1(x, t) = 0$. It then follows from the preceding inequality that

$$\mathscr{L}(x, t) \leqq \mathscr{L}_0(x, t).$$

Case 2. The pair (x, t) is such that $\Delta_1(x, t) \neq 0$. Then it follows from (3.3) that $\gamma(x, t) > 0$. Moreover, in view of (3.3) and the conditions on the C_i,

$$\begin{aligned} \mathscr{L}(x, t) &\leqq \mathscr{L}_0(x, t) - [1 - \Delta_2(x, t)]\gamma(x, t)\|\phi(x, t)\|^2 + \Delta_1(x, t)\|\phi(x, t)\| \\ &= \mathscr{L}_0(x, t) + \frac{\Delta_1^2(x, t)}{4\gamma(x, t)(1 - \Delta_2(x, t))} \\ &\quad - \frac{(1 - \Delta_2(x, t))\gamma(x, t)}{\Delta_1^2(x, t)}\left[\Delta_1(x, t)\|\phi(x, t)\| - \frac{\Delta_1^2(x, t)}{2(1 - \Delta_2(x, t))\gamma(x, t)}\right]^2 \\ &\leqq \mathscr{L}_0(x, t) + \frac{\Delta_1^2(x, t)}{4\gamma(x, t)(1 - \Delta_2(x, t))} \\ &\leqq (1 - C_1)\mathscr{L}_0(x, t) + C_2. \end{aligned}$$

Combining Cases 1 and 2, and noting that $C_1 < 1$, we conclude (as a consequence of Assumption 4) that

$$\mathscr{L}(x, t) \leqq (1 - C_1)\mathscr{L}_0(x, t) + C_2 \leqq -(1 - C_1)\gamma_3(\|x\|) + C_2 \tag{4.2}$$

for all $(x, t) \in R^n \times R$. Having this inequality available now enables one to apply directly the results of [7]. That is, the closed loop system (2.4) possesses a solution $x(\cdot): [t_0, t_1] \to R^n$, $x(t_0) = x_0$, which is required by condition (i) in the definition of practical stabilizability. Moreover, in accordance with [7], if $\|x_0\| \leqq r$, one can satisfy the uniform boundedness requirement (ii) by selecting

$$d(r) \triangleq \begin{cases} (\gamma_1^{-1} \circ \gamma_2)(R) & \text{if } r \leqq R, \\ (\gamma_1^{-1} \circ \gamma_2)(r) & \text{if } r > R, \end{cases}$$

where

$$R \triangleq \gamma_3^{-1}(C_2/1 - C_1).$$

It now follows that there is no finite escape time so that the solution is continuable over $[t_0, \infty)$ and hence condition (iii) holds. Again for $\bar{d} \geqq \underline{d}$, using the estimates provided in [7], one can define

$$T(\bar{d}, r) \triangleq \begin{cases} 0 & \text{if } r \leqq (\gamma_2^{-1} \circ \gamma_1)(\bar{d}), \\ \dfrac{\gamma_2(r) - (\gamma_1 \circ \gamma_2^{-1} \circ \gamma_1)(\bar{d})}{(1 - C_1)(\gamma_3 \circ \gamma_2^{-1} \circ \gamma_1)(\bar{d}) - C_2} & \text{otherwise,} \end{cases} \tag{4.3}$$

and in accordance with [7], the desired uniform ultimate boundedness condition (iv) holds with the proviso that

$$(1 - C_1)(\gamma_3 \circ \gamma_2^{-1} \circ \gamma_1)(\bar{d}) - C_2 > 0. \tag{4.4}$$

Note that this requirement is implied by the satisfaction of condition (d) of (3.4) which entered into the construction of the controller.

Finally, to complete the proof, it remains to establish the desired uniform stability property. Indeed, let $\bar{d} \geqq \underline{d}$ be specified and notice that if $\delta(\bar{d}) = R$, the following property will hold: Given any solution $x(\cdot): [t_0, \infty) \to R^n$, $x(t_0) = x_0$ of (2.4) with $\|x_0\| \leqq \delta(\bar{d})$, it follows (in view of the uniform boundedness property (ii) and the requirements on the C_i) that $\|x(t)\| \leqq d(R) \leqq \bar{d}$ for all $t \geqq t_0$. □

5. Specialization to linear systems. The objective of this section is to show that the flexibility permitted in choosing $\gamma(x, t)$ (see (3.3)) can be exploited in a "nice way" when the nominal system dynamics happen to be linear and time-invariant; that is, the control $p_{\underline{d}}(\cdot)$ in (3.5) can be selected to be a linear time-invariant feedback of the state. We consider the special case when

$$\begin{aligned} \dot{x}(t) &= [A + \Delta A(q(t))]x(t) + [B + \Delta B(q(t))]u(t) + w(q(t)), \\ x(t_0) &= x_0, \end{aligned} \tag{5.1}$$

where A, $\Delta A(q(t))$, B and $\Delta B(q(t))$ are matrices of appropriate dimensions and $w(q(t))$ is an n-dimensional vector. In light of Assumptions 1 and 3 given in § 2, it follows that for all $q \in Q$

$$\begin{aligned} \Delta A(q) &= BD(q), \\ \Delta B(q) &= BE(q), \\ w(q) &= Bv(q), \\ E(q)\| &< 1 \end{aligned} \tag{5.2}$$

where $D(\cdot)$, $E(\cdot)$ and $v(\cdot)$ have appropriate dimensions and depend continuously on their arguments. In accordance with Assumption 4, the matrix A must be asymptotically stable. To obtain a Lyapunov function for the uncontrolled nominal system, we select simply an $n \times n$ positive-definite symmetric matrix H and solve the equation

$$A'P + PA = -H \tag{5.3}$$

for P which is positive-definite; see [11]. Then we have

$$V(x, t) = x'Px \tag{5.4}$$

and

$$\mathscr{L}_0(x, t) = -x'Hx. \tag{5.5}$$

It is clear from (5.4) and (5.5) that one can take the bounding functions $\gamma_i(\cdot)$ to be

$$\gamma_1(r) \triangleq \lambda_{\min}[P]r^2, \quad \gamma_2(r) \triangleq \lambda_{\max}[P]r^2, \quad \gamma_3(r) \triangleq \lambda_{\min}[H]r^2, \tag{5.6}$$

where $\lambda_{\max(\min)}[\cdot]$ denotes the operation of taking the largest (smallest) eigenvalue.

Construction of the controller. We take $\underline{d} > 0$ as prescribed and construct the controller $p_{\underline{d}}(\cdot)$ given in § 3. Using the notation above, we define first[6]

$$\rho_D \triangleq \max_{q \in Q} \|D(q)\|, \quad \rho_E \triangleq \max_{q \in Q} \|E(q)\| < 1, \quad \rho_v \triangleq \max_{q \in Q} \|v(q)\|. \tag{5.7}$$

Then, in agreement with (3.1) and (3.2), we may take

$$\Delta_1(x, t) = \rho_D \|x\| + \rho_v, \qquad \Delta_2(x, t) = \rho_E. \tag{5.8}$$

Using these choices in (3.3) and the fact that $\mathscr{L}_0(x, t) \leqq -\lambda_{\min}[H]\|x\|^2$, one can select $\gamma(\cdot)$ such that

$$\gamma(x, t) \geqq \frac{(\rho_D \|x\| + \rho_v)^2}{4(1-\rho_E)(C_1 \lambda_{\min}[H]\|x\|^2 + C_2)}, \tag{5.9}$$

with the constants C_1 and C_2 yet to be specified. We shall examine three possible cases and see that in all instances one can take $\gamma(x, t) \equiv$ constant. Of course, this implies that the control $p_{\underline{d}}(\cdot)$ is a linear time-invariant feedback; that is,

$$p_{\underline{d}}(x, t) = -2\gamma_0 B'Px, \tag{5.10}$$

where γ_0 is the constant value of $\gamma(\cdot)$, which will be specified.

Case 1. $\rho_D > 0$, $\rho_v = 0$. In this case, we may select $C_2 = 0$ and $C_1 \in (0, 1)$. Consequently, we can satisfy (5.9) by choosing

$$\gamma(x, t) \equiv \gamma_0 \geqq \frac{\rho_D^2}{4(1-\rho_E)C_1 \lambda_{\min}[H]}. \tag{5.11}$$

Case 2. $\rho_D = 0$, $\rho_v > 0$. Clearly, it suffices to take $C_1 = 0$ and

$$\gamma(x, t) \equiv \gamma_0 \geqq \frac{\rho_v^2}{4(1-\rho_E)C_2}, \tag{5.12}$$

where C_2 is required to satisfy condition d) of (3.4). Using the descriptions of the $\gamma_i(\cdot)$ given in (5.6), this amounts to restricting C_2 by

$$\frac{C_2}{1-C_1} < \frac{\lambda_{\min}[P]}{\lambda_{\max}[P]} \lambda_{\min}[H]\underline{d}^2 \tag{5.13}$$

with $C_1 = 0$ in the above.

[6] One can in fact use overestimates $\tilde{\rho}_D$ and $\tilde{\rho}_E$ for ρ_D and ρ_E as long as the inequality $\tilde{\rho}_E < 1$ is satisfied.

Case 3. $\rho_D > 0$, $\rho_v > 0$. Now, in order to satisfy (5.9), we select $C_1 \in (0, 1)$, C_2 satisfying (5.13) and

$$\gamma(x, t) \equiv \gamma_0 \geqq \max_{r \geqq 0} \left\{ \frac{(\rho_D r + \rho_v)^2}{4(1-\rho_E)[C_1 \lambda_{\min}[H] r^2 + C_2]} \right\}. \tag{5.14}$$

Letting $f(r)$ denote the bracketed quantity in (5.14) above, a straightforward but lengthy differentiation yields

$$\max_{r \geqq 0} f(r) = \frac{1}{4(1-\rho_E)} \left\{ \frac{\rho_D^2}{C_1 \lambda_{\min}[H]} + \frac{\rho_v^2}{C_2} \right\}. \tag{5.15}$$

Hence, any γ_0 equal to or exceeding this maximum value will be appropriate in (5.14).

6. Illustrative example. We consider now the simple pendulum which was analyzed in [7]. However, here it will be shown that the desired practical stability can actually be achieved via a linear control. This may seem somewhat surprising in light of the fact that the nominal system dynamics are nonlinear. A pendulum of length l is subjected to a control moment $u(\cdot)$ (per unit mass). The point of support is subject to an uncertain acceleration $q(\cdot)$, with $|q(t)| \leqq \hat{\rho} l \equiv$ constant. Letting x_1 denote the angle between the pendulum's arm and a vertical reference line, one obtains the state equations

$$\begin{aligned} \dot{x}_1(t) &= x_2(t), \\ \dot{x}_2(t) &= -a \sin x_1(t) + u(t) - \frac{q(t) \cos x_1(t)}{l}, \end{aligned} \tag{6.1}$$

where $a > 0$ is a given constant. In order to satisfy the assumptions of § 2 one must assure a uniformly asymptotically stable equilibrium for (UC), the uncontrolled nominal system. Hence, for a given $\underline{d} > 0$, we propose a controller of the form

$$u(t) = -bx_1(t) - cx_2(t) + p_{\underline{d}}(x(t), t), \tag{6.2}$$

where b and c are positive constants and $p_{\underline{d}}(\cdot)$ will be specified later in accordance with the results of § 3. The linear portion of the controller (6.2) is used to obtain a stable nominal system. Substitution of (6.2) into (6.1) now yields the state equation

$$\dot{x}(t) = f(x(t), t) + B[p_{\underline{d}}(x(t), t) + h(x(t), q(t), t)], \tag{6.3}$$

where

$$\begin{aligned} B &= \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \qquad f(x, t) = \begin{bmatrix} x_2 \\ -bx_1 - cx_2 - a \sin x_1 \end{bmatrix}, \\ h(x, q, t) &= \frac{-q \cos x_1}{l}. \end{aligned} \tag{6.4}$$

A suitable Lyapunov function for the uncontrolled nominal system (with $x = 0$ as equilibrium) is

$$V(x, t) = (b + \tfrac{1}{2}c^2)x_1^2 + cx_1x_2 + x_2^2 + 2a(1 - \cos x_1), \tag{6.5}$$

and, provided b is sufficiently large, the associated $\gamma_i(\cdot)$ are given by

$$\begin{aligned} \gamma_1(r) &= \lambda_1 r^2, \\ \gamma_2(r) &= \begin{cases} \lambda_2 r^2 + 2a(1 - \cos r) & \text{if } r \leqq \pi, \\ \lambda_2 r^2 + 4a & \text{if } r > \pi, \end{cases} \\ \gamma_3(r) &= \lambda_3 r^2, \end{aligned} \tag{6.6}$$

$$\lambda_1 \triangleq \lambda_{\min}[P], \quad \lambda_2 \triangleq \lambda_{\max}[P], \quad \lambda_3 \triangleq \min\{\bar{b}c, c\}, \tag{6.7}$$

$$P = \begin{bmatrix} b + \frac{1}{2}c^2 & \frac{1}{2}c \\ \frac{1}{2}c & 1 \end{bmatrix}, \qquad \bar{b} \triangleq b + a \min_{x_1}\left(\frac{\sin x_1}{x_1}\right) > 0. \tag{6.8}$$

Following the procedure described in § 3 for the construction of the controller $p_{\underline{d}}(\cdot)$, we select first

$$\Delta_1(x, t) = \hat{\rho}|\cos x_1|, \qquad \Delta_2(x, t) = 0. \tag{6.9}$$

Inequality (3.3) can then be assured by requiring

$$\gamma(x, t) \geqq \frac{\hat{\rho}^2 \cos^2 x_1}{4[C_2 + C_1\lambda_3\|x\|^2]}. \tag{6.10}$$

Given our desire for a linear feedback, one can select $C_1 = 0$ and satisfy (6.10) by choosing

$$\gamma(x, t) \equiv \gamma_0 \geqq \frac{\hat{\rho}^2}{4C_2}. \tag{6.11}$$

To complete the design, C_2 must be selected to satisfy condition d) of (3.4). The analysis must account for two cases, depending on the size of the given radius $\underline{d} > 0$.

Case 1. $\lambda_1\underline{d}^2 > \lambda_2\pi^2 + 4a$. The required conditions on C_2 are

$$0 < C_2 < (\gamma_3 \circ \gamma_2^{-1} \circ \gamma_1)(\underline{d}) = \frac{\lambda_3}{\lambda_2}(\lambda_1\underline{d}^2 - 4a). \tag{6.12}$$

Case 2. $\lambda_1\underline{d}^2 \leqq \lambda_2\pi^2 + 4a$. In this case, the constraints imposed by (6.12) are met if $C_2 > 0$ is chosen sufficiently small so that

$$\frac{\lambda_2 C_2}{\lambda_1\lambda_3} + \frac{2a}{\lambda_1}\left(1 - \cos\sqrt{\frac{C_2}{\lambda_3}}\right) < \underline{d}^2. \tag{6.13}$$

Having now selected C_2, the controller is specified by (3.5) in conjunction with (6.11); that is,

$$p_{\underline{d}}(x, t) = -\gamma_0 B'(x, t)\nabla_x V(x, t) = -\gamma_0(cx_1 + 2x_2), \tag{6.14}$$

with the proviso that $\gamma_0 \geqq \hat{\rho}^2/4C_2$. It is interesting to note that there is an obvious tradeoff between the required gain constant γ_0 and the given radius $\underline{d} > 0$. As the radius $\underline{d}$ decreases, C_2 decreases, which in turn implies that γ_0 increases. In contrast, the nonlinear saturation controller of [7] remains bounded by the bound of the uncertainty, and the radius $\underline{d}$ can be decreased by increasing the nonlinear gain; i.e., by approaching a discontinuous control.

7. Conclusion. This paper addresses the so-called problem of practical stabilizability for a class of uncertain dynamical systems. In contrast to previous work on problems of this sort, the main emphasis here is on the structure of the controller. It is shown that by choosing the function $\gamma(\cdot)$ in a special way, the resultant control law can often be realized as a linear time-invariant feedback.

REFERENCES

[1] G. LEITMANN, *Guaranteed asymptotic stability for some linear systems with bounded uncertainties*, J. Dynamic Systems, Meaurement and Control, 101 (1979), pp. 212–216.

[2] S. GUTMAN, *Uncertain dynamical systems, a Lyapunov min–max approach*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 437–443; correction, 25 (1980), p. 613.

[3] J. S. THORP AND B. R. BARMISH, *On Guaranteed Stability of Uncertain Linear Systems via Linear Control*, J. Optim. Theory and Applications, in press.
[4] S. S. L. CHANG AND T. K. C. PENG, *Adaptive Guaranteed Cost Control of Systems with Uncertain Parameters*, IEEE Transactions on Automatic Control, AC-17 (1972), pp. 474–483.
[5] A. VINKLER AND I. J. WOOD, *Multistep Guaranteed Cost Control of Linear Systems with Uncertain Parameters*, J. Guidance and Control, 2, 6 (1980), pp. 449–456.
[6] P. MOLANDER, *Stabilization of Uncertain Systems*, Report LUTFD2/(TRFT-1020)/1-111/(1979), Lund Institute of Technology, August 1979.
[7] M. CORLESS AND G. LEITMANN, *Continuous State Feedback Guaranteeing Uniform Ultimate Boundedness for Uncertain Dynamic Systems*, IEEE Trans. Automat. Control, AC-26, 5 (1981), pp. 1139–1144.
[8] A. VINKLER AND I. J. WOOD, *A Comparison of Several Techniques for Designing Controllers of Uncertain Dynamic Systems*, Proceedings of the IEEE Conference on Decision and Control, San Diego CA, 1979.
[9] B. R. BARMISH AND G. LEITMANN, *On Ultimate Boundedness Control of Uncertain Systems in the Absence of Matching Conditions*, IEEE Transactions on Automatic Control, AC-27 (1982), pp. 153–157.
[10] C. BERGE, *Topological Spaces*, Oliver and Boyd, London, (1963).
[11] N. N. KRASOVSKII, *Stability of Motion*, Stanford University Press, Stanford, CA, (1963).
[12] J. P. LASALLE AND S. LEFSCHETZ, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, (1961).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0007 $01.25/0

GENERAL INSTABILITY RESULTS FOR INTERCONNECTED SYSTEMS*

DAVID J. HILL† AND PETER J. MOYLAN†

Abstract. General results giving conditions for an interconnected system to be input-output and/or Lyapunov unstable are considered. These results are derived in terms of the theory of dissipative systems. This enables a very simple formulation of the requirements for instability. In particular, the restrictions of linearity and unstable subsystems, that appear in previous results, are seen to be unnecessary. Consequently, the relationship between instability and stability conditions is made clearer. A wide variety of useful instability criteria can be easily obtained as special cases.

Key words. instability, stability, large-scale systems, interconnected systems

1. Introduction. In the analysis of general nonlinear systems, useful necessary and sufficient conditions for stability are not available. Thus, specific attention has been given to deriving sufficient conditions for instability as well as for stability. As expected, the main results on stability have instability counterparts. In this paper, new results on the instability of general interconnected systems are presented. The main theorems are of a very general nature and the parallelism with results on stability is seen to be intuitively clearer than in other approaches.

Stability results in the theory of interconnected systems tend to fall into two categories: the state-space approach, which was initiated (in the sense of interest here) by Popov, Kalman and Yakubovich, and the input-output approach which was derived by Sandberg [30] and Zames [48]. The historical details of all this work have been documented in many books and papers. Reference is only made here to books by Desoer and Vidyasagar [9] and Michel and Miller [20] and the survey paper by Willems [46] since these are most useful to later discussions. Willems has given considerable attention to illuminating the relationship between the two approaches, and a summary of this work appears in [46]. The basic results are widely recognized by such names as the circle criterion, Popov criterion, and various more abstract results such as the passivity and small-gain theorems. These results have counterparts in each approach, although the abstract results on passivity and system gain emerged via input-output methods.

The first work on instability of nonlinear feedback systems was due to Brockett and Lee [7], who used Lyapunov methods. Further results on this approach were given by Anderson and Moore [1]. Important input-output instability results have been given by Willems [42] and Takeda and Bergen [34]; these results are summarized in [9]. Further work for both feedback and general interconnected systems by the input-output approach has been reported in [10]–[17], [20], [27], [31], [33], [35]–[41]. For general interconnected or large-scale systems, the approach of constructing a Lyapunov function predicting instability for the composite system as a vector or weighted sum of subsystem functions has also been studied [12], [20]. The instability results given by Brockett and Lee [7] are a natural extension from earlier stability results. The same cannot be said for the subsequent results obtained by input-output methods. Typically, the conditions are considerably more complicated and restrictive than the counterpart stability results. It is generally assumed, for instance, that one

* Received by the editors June 29, 1981, and in revised form March 19, 1982.

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

subsystem is linear and unstable; this is the case in all the above-mentioned input-output results, except [31], [35]. The requirement of an unstable subsystem is clearly unsatisfactory since stable subsystems can surely combine to give an unstable system. Further, whereas Lyapunov versions of the abstract input-output stability results have been established [46], the known approaches to input-output instability results seem to preclude development of Lyapunov versions (for general nonlinear systems, at least).

Recently, the authors have presented a unifying viewpoint for the stability of interconnected systems [15], [24]. This work is based on the general property of dissipativeness, which Willems has introduced as a property for state-space representations [44], [45]. By casting this property in functional analysis terms, a general result is given in [24] for input-output stability. Previously known results can be obtained as special cases. Moreover, by using an improved connection between input-output dissipativeness and properties of a state-space representation, input-ouput and Lyapunov stability concepts are treated side by side in a manner yielding considerable insight into stability theory. Willems has discussed the derivation of Lyapunov instability results within the context of dissipative systems [45], [46]. Thus, the authors were led to consider the development of a more complete theory of instability which is comparable in scope to the stability results in [24]. Some brief descriptions of intermediate progress have already appeared [16], [13], [25]. Such stringent conditions as requiring a linear unstable subsystem are dispensed with. Further, the instability conditions appear to be considerably simpler than those obtained by other approaches.

The present paper and the previous one on stability [24] provide a fairly complete exposition of the dissipative systems viewpoint on the stability theory of interconnected systems whose subsystems are defined on inner product signal spaces. This work reflects the opinion that in general the stability analysis of a *nonlinear* interconnected system is not adequately studied by input-output or state-space methods alone. For instance, an input-output stability test is effectively done for a single initial condition. For linear systems, input-output behavior (usually described by input-output pairs for zero initial state) and state behavior are so closely related for a minimal state-space that it does not really matter which approach is used. For nonlinear systems, the choice of initial condition (or reference state) has a major influence on the input-output qualitative behavior. A theory which allows for variable initial state and input is needed. Dissipative systems results have this facility.

Section 2 of the paper contains a brief treatment of dissipative systems. Since an important aspect of the paper is the close relationship between stability and instability results, a brief review of the former is given in § 3. This section also serves to describe the model for interconnected systems. Sections 4 and 5 present input-output and Lyapunov instability results, respectively. In § 6 it is demonstrated how to translate the general results into frequency domain tests. This is illustrated by deriving an instability version of the multivariable circle criterion. Section 7 offers some conclusions.

2. Dissipative systems. This section provides background information on dissipative systems. This theory has evolved through a number of papers including [44], [46], [17]. The exposition given here draws from the ideas presented in [17].

2.1. Notation and definitions. We adopt a description of a dynamical system as an operator on appropriately defined signal spaces.

DEFINITION 1. Let $\mathcal{S}_e$ be a space associated with spaces S, $\mathcal{S}$, and τ where $\mathcal{S} \subset \mathcal{S}_e$ has elements which are functions $v : \tau \to S$. $\mathcal{S}_e$ is called a *signal space* if $\mathcal{S}$ is a real

Hilbert space and $\mathcal{S}_e$ is an extended space in the following sense: There exists a family of projections $P_T: \mathcal{S}_e \to \mathcal{S}$ such that for every $v \in \mathcal{S}_e$ and all $T \in \tau$, we have $P_T v \in \mathcal{S}$.

In general, $\mathcal{S}_e$ is not an inner product space. It is often the case that $\tau = \mathbb{R}_+ \triangleq [t_0, \infty)$ and P_T is an operator which truncates a signal at time T. (Note the facility not to restrict τ to being on the real line.) $\mathcal{S}_e$ is the basic signal space and $\mathcal{S}$ is referred to as the *small signal space*. The value of $v \in \mathcal{S}_e$ at $t \in \tau$ is denoted $v(t)$. P_T is called the causal truncation operator, and it is convenient to use the notation $v_T \triangleq P_T v$. It is assumed that the inner product $\langle \cdot, \cdot \rangle$ on $\mathcal{S}$ fulfills the following properties:

(P1) $$\langle u_T, v \rangle = \langle u, v_T \rangle = \langle u_T, v_T \rangle \triangleq \langle u, v \rangle_T \quad \forall \langle u, v \rangle \in \mathcal{S} \text{ and } \forall T \in \tau,$$

(P2) $$\langle v_T, v_T \rangle^{1/2} \triangleq \|v_T\| \leqq \|v\| \quad \forall v \in \mathcal{S} \text{ and } \forall T \in \tau.$$

We can now introduce the definition of a dynamical system in input-output form. Suppose that $\mathcal{U}_e$ and $\mathcal{Y}_e$ are two signal spaces defined relative to τ with projection operators P_T^u and P_T^y respectively. $\mathcal{U}_e$ is called the *input signal space* and $\mathcal{Y}_e$ is called the *output signal space.*

DEFINITION 2. A *dynamical system input-output representation* is an operator $G: \mathcal{U}_e \to \mathcal{Y}_e$.

DEFINITION 3. G is *causal* if and only if $P_T^y G P_T^u = P_T^y G$.

For the next definition, it is convenient to introduce the anticausal truncation operator $Q_T \triangleq I - P_T$.

DEFINITION 4. G is *anticausal* if and only if $Q_T^y G Q_T^u = Q_T^y G$.

If G is both causal and anticausal, then we call it *memoryless*. Note that the definition of a dynamical system above does not include causality of G.

In discussing stability, it is useful to refer to the set

$$\mathcal{K}(G) \triangleq \{u \in \mathcal{U}: y = Gu \in \mathcal{Y}\}.$$

Note that $\mathcal{K}(G)$ is a subspace of $\mathcal{U}$ if G is linear.

An alternative description of a dynamical system is provided by a state-space representation. We introduce the so-called *state-space* X which is just an abstract set with a zero element. For simplicity, we consider only the case of $\tau \subset \mathbb{R}$ and a time-invariant system. The system description is via the state transition mapping $\psi: \tau^2 \times X \times \mathcal{U}_e \to X$ and the readout mapping $r: \tau \times X \times U \to Y$ satisfying the usual axioms [8]. These mappings describe the time evolution of the state and output according to $x(t) = \psi(t, t_0, x(t_0), u)$ and $y(t) = r(t, x(t), u(t))$ where these are causally dependent on u.

The state-space representation has been presented as more restrictive than the input-output representation. Causality is an inherent feature of the usual concept of system state [8]. The assumption of time-invariance is inessential to the ideas presented later, but allows a convenient simplification of details. Further, note that the space τ becomes one-dimensional so t can be interpreted as time.

For dynamical system G with a state-space representation, we should be clear about the following. The input-output mapping depends on the initial state $x(t_0) = x_0 \in X$. That is, we write $y = G(x_0)u$ where $G(x_0): \mathcal{U}_e \to \mathcal{Y}_e$ is an operator depending on x_0. (The precise relation to mappings ψ and r is given by $y(t) = r(t, \psi(t, t_0, x_0, u), u(t))$.) However, the input-output theory of nonlinear systems [9] is based on properties of the nonlinear map G and can be developed with no mention of an underlying state-space. When dealing with input-output properties only, we drop reference to dependence of G on x_0. If the state-space becomes relevant, then it is convenient to realize that the input-output results refer to $G(x_0)$ where x_0 is some fixed initial state.

DEFINITION 5. A state $x_1 \in X$ is *reachable* if there exist finite $t_1 \geqq t_0$ and $u \in \mathcal{U}_e$ such that $\psi(t_1, t_0, 0, u) = x_1$.

DEFINITION 6. A state $x_0 \in X$ is *controllable* if there exists finite $t_1 \geqq t_0$ and $u \in \mathcal{U}_e$ such that $\psi(t_1, t_0, x_0, u) = 0$.

It is convenient to introduce the following notation. Let X_r denote the set of all reachable points and X_c denote the set of all controllable points.

DEFINITION 7. A region $X_d \subseteq X$ is *zero-state-detectable* if $\forall x_0 \in X_d$ $G(x_0)0 = 0$ implies $x_0 = 0$.

Zero-state-detectability of X_d may be thought of as a weak form of observability. It implies that, with zero input, the outputs resulting from nonzero initial states in X_d are distinguishable from the output resulting from the zero initial state.

We can now give the definitions of dissipativeness for the dynamical system G. Introduce now the *energy supply functional* $E: \mathcal{U}_e \times \mathcal{Y}_e \times \tau \to \mathbb{R}$ defined by

$$E(u, y, T) \triangleq \langle y, Qy \rangle_T + 2\langle y, Su \rangle_T + \langle u, Ru \rangle_T \tag{1}$$

where operators $Q: \mathcal{Y}_e \to \mathcal{Y}_e$, $S: \mathcal{U}_e \to \mathcal{Y}_e$ and $R: \mathcal{U}_e \to \mathcal{U}_e$ are memoryless continuous and linear with both Q and R also selfadjoint. (Since $\mathcal{U}_e$, $\mathcal{Y}_e$ are not inner product spaces in general, continuity refers to the operators $Q|_{\mathcal{Y}}: \mathcal{Y} \to \mathcal{Y}$ etc.) Actually the basic theory of dissipative systems carries over to a more general class of functionals than given by (1) [44]. We give $E(u, Gu, T)$ the interpretation of energy input to the system over time interval T when driven by signal u_T. We also are interested in the energy input caused by the complete signal u when $Gu \in \mathcal{Y}$. This is denoted by $E_c(u, Gu)$ where

$$E_c(u, y) \triangleq \langle y, Qy \rangle + 2\langle y, Su \rangle + \langle u, Ru \rangle.$$

Comment. If it is assumed that the function $T \to \|v_T\|$ is monotone increasing and $\lim_{T\to\infty} \|v_T\| = \|v\|$, then, with E having the quadratic form in (1), it follows that

$$\lim_{T\to\infty} E(u, y, T) = E_c(u, y).$$

This property of the inner product is usually assumed in input-output analysis [9], but is not required for stability or instability results. Hence it is not assumed here.

DEFINITION 8. Dynamical system G with energy supply E is *weakly dissipative* if and only if there exists a constant β such that

$$E(u, Gu, T) + \beta \geqq 0 \tag{2}$$

for all $u \in \mathcal{U}_e$ and for all $T \in \tau$. With $\beta = 0$, we call G *dissipative*.

DEFINITION 9. Dynamical system G with energy supply E_c is *weakly ultimately virtual-dissipative* if and only if there exists a constant β such that

$$E_c(u, Gu) + \beta \geqq 0 \tag{3}$$

for all $u \in \mathcal{K}(G)$. With $\beta = 0$, we call G *ultimately virtual-dissipative*. If in addition $\mathcal{K}(G) = \mathcal{U}$, the system is called (weakly) *ultimately dissipative*.

DEFINITION 10. Assume dynamical system G has a state-space representation and energy supply E. $G(x_0)$ is *cyclo-dissipative* if and only if

$$E(u, Gu, T) \geqq 0$$

for all $u \in \mathcal{U}_e$ and for all $T \in \tau$ such that $x(T) = x_0$.

Comments. 1) Dissipativeness as defined in Definition 8 has been studied by the authors in [17], [18]. When E has the quadratic form given by (1), we refer to (Q, S, R)

dissipativeness. Similar nomenclature is used for the other properties. Dissipativeness contains as special cases the well-known properties used in earlier work in input-output stability analysis [30], [48], [9], [42]. Table 1 summarizes the special cases which are of relevance to this paper along with their usual names. It by no means exhausts the useful ones. Note that the term "strong passivity" in [24] has been abandoned in favor of "strict passivity". The term "pseudo passivity" has been used by Vidyasagar [38].

2) Definitions 9 and 10 describe less restrictive dissipativeness concepts. Basically, the system is required to behave like a dissipative system for a special subset of input signals. They contain as special cases various properties introduced in [34], [45], [17]. Ultimate virtual-dissipativeness was introduced in [16]. A causal ultimately virtual-dissipative system is called virtual-dissipative.

3) Definitions 8 and 9 do not depend on the existence of a state-space representation. If one is available, β should be regarded as dependent on x_0 [18].

TABLE 1
Special dissipativeness properties

Q, S, R	Type of dissipativeness
$-\delta I, \frac{1}{2}I, -\varepsilon I$	pseudo strict passivity
$0, \frac{1}{2}I, 0$	passivity
$-I, 0, k^2 I$	finite gain
$I, 0, -l^2 I$	lower bound on gain
$-I, \frac{1}{2}(a+b)I, -abI$	inside sector $[a, b]$
$I, -\frac{1}{2}(a+b)I, abI$	outside sector $[a, b]$

I denotes the identity operator, δ, ε, k, l, a, b denote scalar parameters.

4) If G is linear, it is easy to show that the distinction between the weak and nonweak versions of the dissipativeness concepts disappears. It turns out that the obvious definition of weak cyclo-dissipativeness is equivalent to cyclo-dissipativeness for general dynamical systems.

5) The study of all the relationships between the various forms of dissipativeness will not be undertaken here. Let us note, however, that cyclo-dissipativeness is a weaker property than weak virtual-dissipativeness (assuming existence of a state representation). By imposing certain smoothness and observability restrictions, cyclo-dissipativeness and virtual-dissipativeness of $G(x_0)$ become equivalent [17].

The basic stability and instability results are derived from the quadratic form for E and E_c with sign definiteness properties on the operator Q. These properties are defined for elements of the Banach algebra of continuous linear operators on a Hilbert space $\mathcal{S}$, which is denoted by $\mathcal{B}(\mathcal{S})$.

DEFINITION 11. A selfadjoint operator $A \in \mathcal{B}(\mathcal{S})$ is said to be *positive semidefinite* if $\langle v, Av \rangle \geqq 0$ for all $v \in \mathcal{S}$. If $A - \mu I$ is positive semidefinite for some scalar $\mu > 0$, then A is *strictly positive definite*. Similarly, negative semidefinite and strictly negative operators are defined by reversing the inequalities.

To be precise, when referring to sign properties of Q, it is meant that the restricted operator $Q|_{\mathcal{S}}$ takes one of the properties in Definition 11.

The study of instability in the sequel relies on the following connection between ultimate dissipativeness and dissipativeness. It is proved in [17].

LEMMA 1. *Suppose that the system G is causal and (weakly) (Q, S, R) ultimately dissipative with Q negative semidefinite. Then G is (weakly) (Q, S, R) dissipative.*

Comment. This result generalizes the well-known relationship between passivity, ultimate passivity and causality [9].

2.2. Implications of dissipativeness on state-space representation. An important property of a dissipative system is that it possesses a scalar-valued energy-like function, which, under certain circumstances, can act as a Lyapunov function. The following results formalize this fundamental connection between input-output and state-space representations of G.

THEOREM 1. *Dynamical system $G(0)$ with energy supply E is dissipative if and only if there exists a function $\phi: X_r \to \mathbb{R}$ such that $\phi(x) \geqq 0$ for all $x \in X_r$, $\phi(0) = 0$ and*

$$\phi(x_1) + E(u, G(x_1)u, T) \geqq \phi(x_2) \tag{4}$$

for all $x_1 \in X_r$, for all $u \in \mathcal{U}_e$ and for all $T \geqq t_0$ where $x_2 = \psi(T, t_0, x_1, u)$.

THEOREM 2. *Dynamical system $G(0)$ with energy supply E is cyclo-dissipative if and only if there exists a function $\phi: X_r \cap X_c \to \mathbb{R}$ such that $\phi(0) = 0$ and (4) is satisfied for all $x_1 \in X_r \cap X_c$, for all $u \in \mathcal{U}_e$ and for all $T \geqq t_0$ where $x_2 = \psi(T, t_0, x_1, u)$.*

Comments. Theorems 1 and 2 are proved in [17]. These results are the culmination of a sequence of generalizations of the well-known Kalman–Yakubovich–Popov Lemma. In [18], a slightly more general version of Theorem 1 is given which allows ϕ to vanish on a set.

2) Similar results can be stated for every other dissipativeness property, but only these two will be needed in the sequel. For weak dissipativeness, the only change to the properties of ϕ in Theorem 1 is that the condition $\phi(0) = 0$ need not hold [18].

3. Review of stability. The present section gives a quick review of stability results obtained from the dissipative systems approach [15], [24]. This serves to introduce the basic interconnected systems setup and provides for a detailed comparison to the instability results in the next section.

DEFINITION 12. The dynamical system G is *input-output stable* if and only if $\mathcal{H}(G) = \mathcal{U}$.

A stronger form of stability is the following property.

DEFINITION 13. The dynamical system G is *weakly finite-gain stable* if and only if there exists constants k and β such that

$$\|Gu\|_T \leqq k\|u\|_T + \beta$$

for all $u \in \mathcal{U}_e$ and for all $T \in \tau$. If $\beta = 0$, we call G finite-gain stable.

Note that finite-gain stability is a special case of dissipativeness. The basic lemma for deriving results on finite-gain stability is as follows.

LEMMA 2 [24]. *Suppose the dynamical system G is (weakly) (Q, S, R) dissipative. If Q is strictly negative definite, then G is (weakly) finite-gain stable.*

Comment. This result becomes an equivalence between dissipativeness and finite-gain stability when $\beta = 0$.

At this point, the interconnected system will be described. Suppose we have N subsystems represented by operators $G_i: \mathcal{U}_{ei} \to \mathcal{Y}_{ei}$, $i = 1, \cdots, N$. Let the subsystems be interconnected via

$$u_i = u_{ei} - \sum_{j=1}^{N} H_{ij} y_j, \qquad i = 1, \cdots, N, \tag{5}$$

where u_i and $y_i = G_i u_i$ are the input and output of G_i, the u_{ei} are external inputs, and

the $H_{ij}: \mathcal{Y}_{ei} \to \mathcal{U}_{ei}$ are memoryless continuous linear operators. Further, suppose that the subsystem G_i is weakly (Q_i, S_i, R_i) dissipative. Then, we have

$$\langle y_i, Q_i y_i \rangle_T + 2\langle y_i, S_i u_i \rangle_T + \langle u_i, R_i u_i \rangle_T + \beta_i \geqq 0 \tag{6}$$

for all $u_i \in \mathcal{U}_{ei}$ and for all $T \in \tau$. The outputs of the overall system are taken to be the y_i.

Now clearly the input and output signal spaces for the interconnected system G are product spaces $\hat{\mathcal{U}}_e = \prod_{i=1}^N \mathcal{U}_{ei}$ and $\hat{\mathcal{Y}}_e = \prod_{i=1}^N \mathcal{Y}_{ei}$ respectively. The elements of $\hat{\mathcal{U}}_e$ and $\hat{\mathcal{Y}}_e$ are N-tuples $u = (u_1, u_2, \cdots, u_N)$ and $y = (y_1, y_2, \cdots, y_N)$. Inner products in these spaces are derived by summing inner products in the component spaces. For instance, let $u, v \in \hat{\mathcal{U}}_e$ and we define

$$\langle u, v \rangle \triangleq \sum_{i=1}^N \langle u_i, v_i \rangle.$$

Also, let $P_T u = (P_T u_1, P_T u_2, \cdots, P_T u_N)$. We require that the system satisfies certain restrictions to ensure it is well-posed; that is, the operator $G: \hat{\mathcal{U}} \to \hat{\mathcal{Y}}$ must be well defined [39]. At this point, it is convenient to adopt a notational device which enables representation of operators on $\hat{\mathcal{U}}$, $\hat{\mathcal{Y}}$ as arrays of operators mapping the components. The array representing an operator A will be denoted by $\mathbf{A}$. With operator H defined as an array of the operators H_{ij}, (5) can be written more compactly as

$$u = u_e - \mathbf{H}y. \tag{7}$$

Let $\mathbf{Q} = \operatorname{diag}\{Q_1, Q_2, \cdots, Q_N\}$, $\mathbf{S} = \operatorname{diag}\{S_1, S_2, \cdots, S_N\}$, and $\mathbf{R} = \operatorname{diag}\{R_1, R_2, \cdots, R_N\}$. Adding the inequalities (6) and substituting (7), it is straightforward to show that the overall system is $(\hat{Q}, \hat{S}, \hat{R})$ dissipative with

$$\hat{\mathbf{Q}} = \mathbf{Q} + \mathbf{H}^*\mathbf{R}\mathbf{H} - \mathbf{S}\mathbf{H} - \mathbf{H}^*\mathbf{S}^* \tag{8}$$

and

$$\hat{\beta} = \sum_{i=1}^N \beta_i$$

(where the * denotes an adjoint)[1]. The forms of $\hat{S}$ and $\hat{R}$ are of no interest here. In the special case of the feedback system shown in Fig. 1, we have

$$\hat{\mathbf{Q}} = \begin{bmatrix} Q_1 + pR_2 & -S_1 + pS_2^* \\ -S_1^* + pS_2 & R_1 + pQ_2 \end{bmatrix} \tag{9}$$

where p is any positive scalar. (This scalar is introduced since the inequalities (6) are unchanged if multiplied by positive scalars.)

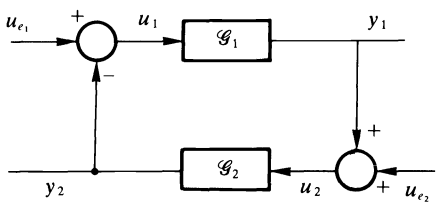


FIG. 1. *Feedback configuration.*

[1] The existence of H^* and S^* as mappings on $\hat{\mathcal{Y}}_e$ and $\hat{\mathcal{U}}_e$ respectively follows only because H and S are memoryless.

THEOREM 3. *The interconnected systems, with subsystems H_i which are (weakly) (Q_i, S_i, R_i) dissipative, $i = 1, \cdots, N$, and interconnected according to (5), is (weakly) finite-gain stable if the operator $\hat{\mathbf{Q}}$ is strictly negative definite.*

Comments. 1) This result has been given by the authors in [24], [25]. In [15], [24], a Lyapunov version was also presented (with [15] considering only the feedback configuration). This, of course, requires the imposition of the extra assumptions associated with existence of state-space representations. Further discussion of Lyapunov stability results is left to § 5.

2) Consideration of finite-gain stability when $\hat{Q}$ is negative semidefinite (but not necessarily strictly negative definite) has been given in [23]. Stability can still hold if $\hat{Q}$ satisfies some extra rank conditions.

3) In [24], Theorem 3 and its Lyapunov version are shown to include most of the known results on interconnected system stability as special cases. Moreover, these results are almost trivial to extract and there is the flexibility to consider a variety of new situations. Only one result on passivity is given here for illustration. This and the general result will be adequate for later comparison of stability and instability results. Here, and in later results, the scalar parameters, ε, δ, etc., are those in the definitions given by Table 1.

COROLLARY 1. *Suppose that the two subsystems of a feedback system are pseudo strictly passive. Then the system is finite-gain stable if*

$$\varepsilon_1 + \delta_2 > 0, \qquad \varepsilon_2 + \delta_1 > 0.$$

Proof. Setting $p = 1$, $\hat{\mathbf{Q}}$ becomes

$$\hat{\mathbf{Q}} = \begin{bmatrix} -(\varepsilon_2 + \delta_1)I & 0 \\ 0 & -(\varepsilon_1 + \delta_2)I \end{bmatrix}$$

where I is the identity operator on the common signal space for the subsystems. The result is then immediate from Theorem 1. □

Comment. The result was given by Vidyasagar [37] as a generalization of earlier passivity stability results—see the references given in [37]. A similar result for Lyapunov stability was given in [15].

4. Input-ouput instability.

4.1. General results. In this section, the main results of the paper are presented. Firstly, two basic instability lemmas will be proved.

LEMMA 3. *Suppose that the dynamical system G is causal and ultimately (Q, S, R) virtual-dissipative, but not (Q, S, R) dissipative. If Q is negative semidefinite, then G is input-output unstable.*

Proof. We argue via a contradiction. Suppose that for all $u \in \mathcal{U}$ it holds that $y \in \mathcal{Y}$; that is, we have $\mathcal{K}(G) = \mathcal{U}$. From Lemma 1, it is immediate that G is dissipative. The contradiction is thus established. □

LEMMA 4. *Suppose that the dynamical system G is causal and weakly ultimately (Q, S, R) virtual-dissipative with constant β, but not weakly (Q, S, R) dissipative for some constant $\beta_1 \geqq \beta$. If Q is negative semidefinite, then G is input-output unstable.*

Proof. Follows by a minor extension of that for Lemma 3. □

Comments. 1) These results, although almost trivial to prove, given Lemma 1, play the same role for instability results as Lemma 2 does for stability results.

2) A further possibility for producing instability results could appear to be mixing of "weak" and "nonweak" versions of dissipativeness properties. In this discussion, we assume G is causal and Q is negative semidefinite. Clearly, if G is ultimately

virtual-dissipative and not weakly dissipative, then Lemma 4 can be applied. This leaves in question the situation where G is weakly ultimately virtual-dissipative and not dissipative. Such systems can be input-output stable [18].

3) It is interesting to note that Lemmas 3 and 4 offer the possibility of constructing a destabilizing control. In Lemma 3, since G is not dissipative, there exists a control $u \in \mathcal{U}_e$ and time $T \geqq 0$ such that

$$\langle y, Qy\rangle_T + 2\langle y, Su\rangle_T + \langle u, Ru\rangle_T < 0. \tag{10}$$

Now let $u_d = P_T u$. Then $u_d \in \mathcal{U} - \mathcal{K}(G)$. To see this, we assume that $y_d = Gu_d \in \mathcal{Y}$. It is easy to show that (10) implies

$$\langle y_d, Qy_d\rangle + 2\langle y_d, Su_d\rangle + \langle u_d, Ru_d\rangle < 0. \tag{11}$$

This contradicts the assumption that G is ultimate virtual-dissipative. A similar construction holds for Lemma 4.

The above lemmas lead to two very general results on the instability of interconnected systems. They provide counterparts to the stability result in Theorem 3.

THEOREM 4. *Assume that the interconnected system is causal. Suppose the subsystems G_i are ultimately (Q_i, S_i, R_i) virtual-dissipative for $i = 1, \cdots, N$, but not (Q_i, S_i, R_i) dissipative for at least one i. Then the interconnected system is unstable if $\hat{Q}$ is negative semidefinite and one or both of the following conditions holds*:

(i) *at least one of the subsystems which is not dissipative is linear*;

(ii) *for each subsystem (except possibly one of the nondissipative ones), G_i is unbiased and/or Q_i is negative semidefinite.*

Proof. It is easy to see that the interconnected system G is $(\hat{Q}, \hat{S}, \hat{R})$ ultimately virtual-dissipative. We have

$$\mathcal{K}(G) = \left\{ u_e \in \hat{\mathcal{U}} : u = u_e - \mathbf{H}y \in \prod_{i=1}^{N} \mathcal{K}(G_i) \right\}.$$

Let E_i be the energy supply functions for the subsystems G_i, $i = 1, \cdots, N$. We identify G_k as one of the nondissipative subsystems. Then it follows that there exists an input $u_{k_0} \in \mathcal{U}_{ek}$ and time $T_0 \in \tau$ such that $E_k(u_{k_0}, G_k u_{k_0}, T_0) < 0$.

Note for the interconnected system

$$E(u_e, Gu_e, T) = \sum_{i=1}^{N} E_i(u_i, G_i u_i, T),$$

subject to equations (5).

Case (i). Consider a set of subsystem inputs u_{i_0}, $i = 1 \cdots, N$. Now modify the u_e to achieve

$$u_i = \begin{cases} \lambda u_{k_0}, & i = k, \\ u_{i_0} & \text{otherwise.} \end{cases}$$

(This relies on the well-posedness assumption for the interconnected system G.) Then, since G_k is linear, we have

$$E(u_e, Gu_e, T_0) = \lambda^2 E_k(u_{k_0}, G_k u_{k_0}, T_0) + \sum_{\substack{i=1 \\ i \neq k}}^{N} E_i(u_{i_0}, G_i u_{i_0}, T_0).$$

Obviously, taking λ large enough will ensure that an input u_e can be found such that $E(u_e, Gu_e, T_0) < 0$; that is, the system G is not dissipative. The result follows from Lemma 3.

Case (ii). Choose external input u_e such that

$$u_i = \begin{cases} u_{k_0}, & i = k, \\ 0 & \text{otherwise.} \end{cases}$$

Then, we have

$$E(u_e, Gu_e, T_0) = E_k(u_{k_0}, G_k u_{k_0}, T_0) + \sum_{\substack{i=1 \\ i \neq k}}^{N} \langle G_i 0, Q_i G_i 0 \rangle_{T_0}.$$

If G_i is unbiased, then $G_i 0 \equiv 0$. If Q_i is negative semidefinite for $i \neq k$, then $\sum_{i=1, i\neq k}^{N} \langle G_i 0, Q_i G_i 0 \rangle_{T_0} \leqq 0$. Thus under condition (ii), the system G is again not dissipative and Lemma 3 can be applied. □

THEOREM 5. *Assume that the interconnected system is causal. Suppose the subsystems* G_i *are weakly ultimately* (Q_i, S_i, R_i) *virtual-dissipative for* $i = 1, \cdots, N$, *but not* (Q_i, S_i, R_i) *weakly dissipative for any constant* β_{1i} *for at least one* i. *Then the interconnected system is unstable if* $\hat{Q}$ *satisfies the conditions of Theorem* 4.

Proof. Follows by essentially the same argument as for Theorem 4. □

Comments. 1) Theorems 4 and 5 are considerably more general in scope than previous results on input-output instability and appear to emanate from simpler technical arguments. Most of the previous results have been derived for rather restrictive situations which include the undesirable assumption that some subsystems are linear and unstable [20], [27], [33], [36]–[38], [40]. The results in [31], [35] for instance manage to avoid this to some extent, but do not appear to be closely related to those presented here. Neither theorem requires subsystem instability, and linearity only enters as part of Theorems 4 and 5.

2) Theorem 4 sharpens results which were previously presented in [16], [13], [25]. Theorem 5 is new. In [16] the basic ideas were developed and the need to explicitly assume unstable subsystems was dispensed with. In [25], it was observed that the linearity of one subsystem is not needed when the remaining subsystems are unbiased.

3) It is of interest to compare Theorems 4 and 5 on instability with Theorem 3 on stability. Given the framework that all subsystems are ultimately virtual-dissipative, loosely speaking we have (a) stability if all subsystems are dissipative and $\hat{Q}$ is strictly negative definite and (b) instability if one subsystem is not dissipative, the overall system is causal, and $\hat{Q}$ is negative semidefinite. The interesting difference is the requirement of causality in the instability result. It is well known [9], [42] that instability and noncausality are intimately related.

4) Condition (ii) of Theorem 4 could be weakened further to apply only to the dissipative subsystems if on choosing u_{i_0} and T_{i_0} such that $E(u_{i_0}, G_i u_{i_0}, T_{i_0}) < 0$ for each nondissipative G_i, there is no loss of generality in taking all the T_{i_0} to be equal. In general, though, such an assumption would be hard to justify.

5) The condition on β_{1k} in Theorem 5 can be relaxed to require G_k not be weakly dissipative for some β_{1i}, satisfying

$$\beta_{1i} \geqq \sum_{i=1}^{N} \beta_i.$$

6) In view of comment 4) following Definitions 8–10, we see that Theorems 4 and 5 are equivalent for linear systems.

4.2. Special cases of general results.

4.2.1. Feedback systems. The following result summarizes general instability criteria for feedback systems.

THEOREM 6. *Assume that the feedback system is causal. Suppose subsystems* G_i *are* (Q_i, S_i, R_i) *ultimately virtual-dissipative for* $i = 1, 2$ *and* $\hat{Q}$ *given by* (9) *is negative semidefinite. Then the feedback system is unstable if one or more of the following conditions holds*:

(i) G_i *is linear and not* (Q_1, S_1, R_1) *dissipative*;

(ii) G_1 *is not* (Q_1, S_1, R_1) *dissipative and* G_2 *is unbiased and/or* Q_2 *is negative semidefinite*;

(iii)–(iv) *interchange* G_1 *and* G_2 *in* (i) *and* (ii).

Comments. 1) Theorem 6 is a special case of Theorem 4. Results based on weak dissipativeness properties which correspond to Theorem 6 and subsequent Corollaries 2 to 6 are easily obtained from Theorem 5. The details will not be presented.

2) If Q_2 is strictly negative definite and $0 \in \mathcal{K}(G_2)$, then G_2 is unbiased. This gives some connection between the alternative conditions in (ii).

Now we state a general passivity instability result.

COROLLARY 2. *Assume that the feedback system is causal. Suppose that the subsystems* G_i *are ultimately virtual pseudo strict passive for* $i = 1, 2$. *Then the feedback system is unstable if*

$$\delta_1 + \varepsilon_2 \geqq 0, \qquad \delta_2 + \varepsilon_1 \geqq 0$$

and one or more of the following conditions holds:

(i) G_1 *is linear and not pseudo strict passive*;

(ii) G_1 *is not pseudo strict passive and* G_2 *is unbiased and/or* $\delta_2 \geqq 0$;

(iii), (iv) *interchange* G_1 *and* G_2 *in* (i) *and* (ii).

Comments. 1) Corollaries 1 and 2 provide a complete set of conditions for stability and instability of the feedback connection of passive subsystems.

2) Note that a further special case of Corollary 2 is the following very simple result: If the feedback system is causal, G_1 and G_2 are virtual-passive and G_1 *or* G_2 is not passive, then the system is unstable. This should be compared with the complex set of conditions presented in other passivity instability results for feedback systems [27], [33], [34].

The next result deals with a feedback interconnection of finite-gain systems.

COROLLARY 3. *Assume that the feedback system is causal. Suppose that the subsystems* G_i *are ultimately virtual finite-gain with gain bounds of* k_i *for* $i = 1, 2$, *but* G_1 *is not finite-gain. Then the feedback system is unstable if* $k_1 k_2 \geqq 1$.

Proof. We have

$$\hat{\mathbf{Q}} = \begin{bmatrix} (pk_2^2 - 1)I & 0 \\ 0 & (k_1^2 - p)I \end{bmatrix}.$$

Clearly $\hat{\mathbf{Q}}$ is negative semidefinite if $k_1 k_2 \leqq 1$. Since $Q_2 = -I$, the result follows from Theorem 6(ii). □

Comment. This result is clearly superior to previously published versions of a small gain instability theorem [34], [16].

Considering subsystems with a lower bound on gain leads to the following result.

COROLLARY 4. *Assume that the feedback system is causal. Suppose that the subsystems* G_i *are ultimately virtual lower bounded with gain bounds* ℓ_i *for* $i = 1, 2$.

Then the feedback is unstable if $\ell_1\ell_2 \geqq 1$ *and one or more of the following conditions holds*:

(i) G_1 *is linear and not lower bounded*;

(ii) G_1 *is not lower bounded and* G_2 *is unbiased*;

(iii), (iv) *interchange roles of* G_1 *and* G_2 *in* (i) *and* (ii).

Proof. Again, this follows by simple use of Theorem 6. Clearly, condition (ii) with Q_2 negative semidefinite does not apply. □

Besides passivity and finite-gain results, other known input-output stability results are based on the property of conicity.

COROLLARY 5. *Assume that the feedback system is causal. Suppose that* G_2 *is ultimately virtual inside the sector* $[a+\Delta, b-\Delta]$, *where* $b>0$, *and* G_1 *is* $(ab, \frac{1}{2}(a+b), (1-b\delta)(1+a\delta))$ *ultimately virtual-dissipative. Then the feedback system is unstable if both constants* Δ *and* δ *are nonnegative and one or more of the following conditions holds*:

If $a \leqq 0$,

(i) G_1 *is not* $(ab, \frac{1}{2}(a+b), (1-b\delta)(1+a\delta))$ *dissipative*;

(ii) G_2 *is not inside the sector* $[a+\Delta, b-\Delta]$.

If $a>0$,

(i) G_2 *is linear and not inside the sector* $[a+\Delta, b-\Delta]$;

(ii) G_2 *is not inside the sector* $[a+\Delta, b-\Delta]$ *and* G_1 *is unbiased.*

Proof. Note that G_2 is $(-I, \frac{1}{2}(a+b)I, -(a+\Delta)(b-\Delta)I)$ ultimately virtual-dissipative. The details are more tedious for this case, but are basically similar to those used for the corresponding stability result [15]. □

Comments. Corollaries 2, 3 and 5 provide instability counterparts to the well-known positive operator, small gain, and conic operator theorems [48], [9]. Corollary 4 is less familiar, but demonstrates the flexibility of the dissipativeness framework. Of course, other situations can be considered, as for special cases of the stability result [15].

4.2.2. General interconnected systems. In the interest of brevity, we will only prove results for passive and finite-gain systems. At this stage, it should be clear that numerous other results could be derived.

COROLLARY 6. *Assume that the interconnected system is causal. Suppose that the subsystems* G_i *are ultimately virtual pseudo strict passive for* $i = 1, \cdots, N$, *but not pseudo strict passive for at least one i. Then the interconnected system is unstable if the matrix*

$$\mathbf{M} = \mathbf{D} + \mathbf{H}^*\mathbf{E}\mathbf{H} + \tfrac{1}{2}(\mathbf{H} + \mathbf{H}^*)$$

is positive semidefinite and one or more of the following conditions holds:

(i) *at least one of the nondissipative subsystems is linear*;

(ii) *for each subsystem (except possibly one of the nondissipative ones),* G_i *is unbiased and/or* $\delta_i \geqq 0$.

Proof. Define

$$\mathbf{E} = \operatorname{diag}\{\varepsilon_i I_{ni}\}, \qquad \mathbf{D} = \operatorname{diag}\{\delta_i I_{n_i}\}$$

where I_{n_i} are appropriate identity operators. We have

$$\hat{\mathbf{Q}} = -\mathbf{D} - \mathbf{H}^*\mathbf{E}\mathbf{H} - \tfrac{1}{2}(\mathbf{H} + \mathbf{H}^*).$$

The result follows immediately from Theorem 4. □

Comments. 1) It usually holds that the interconnections are passive; that is, the symmetric part of $\mathbf{H}$ is positive semidefinite. In this case, we merely require array $\mathbf{D} + \mathbf{H}^*\mathbf{E}\mathbf{H}$ to be positive semidefinite.

2) This result is considerably simpler than that given by Vidyasagar [37]. A more detailed comparison is reserved for the next section.

From Theorem 4, we get immediately the following result on interconnections of finite-gain systems.

COROLLARY 7. *Assume that the interconnected system is causal and the subsystems have a scalar output and input. Suppose that the subsystems G_i are ultimately virtual finite-gain with gain bounds k_i for $i = 1, \cdots, N$, but not finite-gain for at least one i. Then the interconnected system is unstable if there exists a diagonal positive definite matrix $\mathbf{P}$ such that $\mathbf{P} - \mathbf{A}^T\mathbf{P}\mathbf{A}$ is nonnegative definite where $\mathbf{A} = \mathbf{K}\mathbf{H}$ and $\mathbf{K} = \operatorname{diag}\{k_i\}$.*

Proof. The ith subsystem is $(-p_i, 0, p_i k_i^2)$ ultimately virtual-dissipative for any $p_i > 0$. This leads to

$$\hat{\mathbf{Q}} = -\mathbf{P} + \mathbf{A}^T\mathbf{P}\mathbf{A},$$

and the result follows from Theorem 4. □

The study of positive definiteness of $\mathbf{P} - \mathbf{A}^T\mathbf{P}\mathbf{A}$ has been important for other results in stability theory [4], [20], [24]. For computational simplicity, it is desirable to replace the difficult task of calculating $\mathbf{P}$ by a more direct matrix test on $\mathbf{A}$. (This more direct test will be simpler to apply, but is possibly more conservative.) For the present situation, we can use the following matrix instability test. The relevant matrix definitions and theory have been summarized in Appendix A.

COROLLARY 8. *Under the conditions of Corollary 7, the interconnected system is unstable if $\mathbf{I} - |\mathbf{A}|$ is an M-matrix or $\mathbf{A}$ is strongly irreducible and $\mathbf{I} - |\mathbf{A}|$ is a semi-M-matrix.*

Proof. Follows immediately from Theorems A.1 and A.2. □

Comments. 1) An equivalent test on $\mathbf{A}$ is to require $\rho(\mathbf{A}) < 1$ or $\rho(\mathbf{A}) = 1$ and $\mathbf{A}$ strongly irreducible, where $\rho(\mathbf{A})$ denotes the spectral radius of $\mathbf{A}$. (This follows from Fact A.4.)

2) Vidyasagar [36] has given closely related results for systems with some subsystems assumed inter alia to be linear and unstable. The instability condition is either $\rho(\mathbf{A}) < 1$ or $\rho(\mathbf{A}) \leqq 1$, and at least one of the unstable subsystems is connected to every stable subsystem. A further related result is presented by Michel and Miller [20], where only one linear unstable system is assumed. In the case where $\rho(\mathbf{A}) = 1$, they require $\mathbf{A} > \mathbf{0}$, which certainly implies strong irreducibility.

3) Using the graph theoretic interpretation of irreducibility given in Fact A.5, the strong irreducibility requirement on $\mathbf{H}$ in Corollary 8 can be related to system structure: the digraphs $\mathcal{G}(\mathbf{H})$ and $\mathcal{G}(|\mathbf{H}|^T|\mathbf{H}|)$ must be strongly connected. $\mathcal{G}(\mathbf{H})$ is a direct representation of system structure with vertices and edges corresponding to subsystems and interconnections respectively. Let $|\mathbf{H}| = [\mathbf{h}_1 \cdots \mathbf{h}_N]$ where $\mathbf{h}_i$ are vectors $i = 1, \cdots, N$. Then the (i, j)th element of $|\mathbf{H}|^T|\mathbf{H}|$ is $\mathbf{h}_i^T\mathbf{h}_j$. Clearly $\mathbf{h}_i^T\mathbf{h}_j = 0$ if and only if outputs y_i and y_j are not connected to a common input. So if $\mathcal{G}(|\mathbf{H}|^T|\mathbf{H}|)$ is not strongly connected, the subsystem outputs can be divided into two groups which have feedback interconnections to disjoint groups of inputs.

Example. Figure 2(a) shows an interconnected system with three subsystems. Figures 2(b), (c) show the associated digraphs. Ignoring the dotted interconnections, we see that $\mathcal{G}(\mathbf{H})$ is strongly connected, but $\mathcal{G}(\mathbf{H}^T\mathbf{H})$ is not even connected. On adding the two extra connections, both digraphs become strongly connected and $\mathbf{H}$ is strongly irreducible.

4.3. Allowing for strongly unstable subsystems. A major portion of the previously published results on interconnected system instability has assumed that some

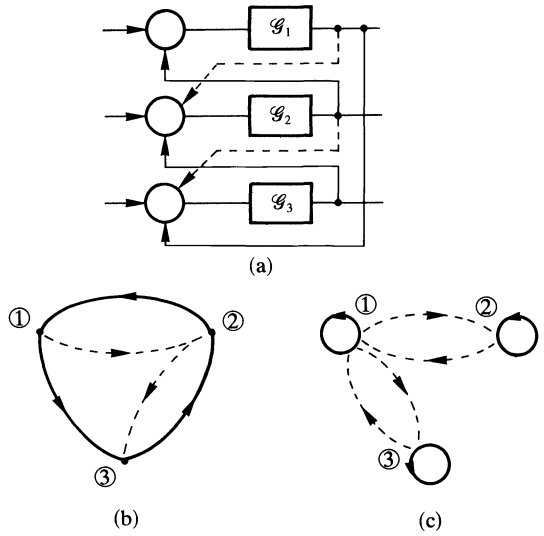


FIG. 2. (A) *Interconnected system.* (b) *Digraph* $G(H)$. (c) *Digraph* $G(H^TH)$.

subsystems satisfy a special strong instability assumption [20], [34], [36]–[38], [40]. In this section, we establish some connection with these results.

DEFINITION 14 [36]. The operator $G: \mathcal{U}_e \to \mathcal{Y}_e$ is said to belong to Class U if

(i) G is linear;

(ii) $\mathcal{K}(G)$ is a proper subset of $\mathcal{U}$;

(iii) There is a finite constant γ_c such that

$$\|Gu\| \leqq \gamma_c \|u\| \quad \forall\, u \in \mathcal{K}(G);$$

(iv) there is a family of constants $\alpha(T)$ such that

$$\|(Gu)_T\| \leqq \alpha(T)\|u_T\| \quad \forall T \geqq 0,\ \forall u \in \mathcal{U}.$$

Comments. 1) It is shown in [34] that conditions (i)–(iv) imply that G is a strongly unstable system in the sense that the orthogonal complement of $\mathcal{K}(G)$, denoted $\mathcal{K}^{\perp}(G)$, contains nonzero elements.

2) Condition (iii) merely says that G is ultimately virtual finite-gain with gain bound γ_c. Conditions (i) and (iv) imply G is causal. Condition (iv) is a well-posedness restriction on G. $\gamma_c = \sup_{u \in \mathcal{K}(G)} (\|Gu\|/\|u\|)$ is called the conditional gain of G in [40].

It is convenient to note the following result.

LEMMA 5. *Suppose the system G is causal and (Q, S, R) ultimately virtual-dissipative with Q negative definite. Then G is not (Q, S, R) dissipative if and only if $\mathcal{K}(G) \neq \mathcal{U}$.*

Proof. Only if. Follows immediately from Lemma 3.

If. Suppose G is (Q, S, R) dissipative. Then Lemma 2 implies G is finite-gain stable, which is a contradiction. □

It is now evident that a class U operator represents a linear, well-posed, causal, ultimately virtual finite-gain, but not finite-gain system. The well-posedness condition ensures strong instability, and this is essential when it is required to exhibit destabilizing controls. This feature is of no importance in the results presented in this paper, but is required in carrying over L_2-instability criteria to L_∞-instability criteria [41].

We now suppose that each subsystem G_i is (Q_i, S_i, R_i) ultimately virtual-dissipative and some are also in class U. Each class U subsystem is $(Q_i - \alpha I, S_i, R_i + \alpha\gamma_{c_i}^2 I)$

ultimately virtual-dissipative for any $\alpha \geqq 0$. The original matrix $\hat{\mathbf{Q}}$ is now augmented to

$$\hat{\mathbf{Q}}_a = \hat{\mathbf{Q}} + \alpha((\mathbf{\Gamma}_c\mathbf{H})^T\mathbf{D}(\mathbf{\Gamma}_c\mathbf{H}) - \mathbf{D})$$

where $\mathbf{\Gamma}_c = \text{diag}\{\gamma_{c_i}\mathbf{I}_{n_i}\}$ and $\mathbf{D} = \text{diag}\{d_i\mathbf{I}_{n_i}\}$ where

$$d_i = \begin{cases} 1 & \text{if } G_i \text{ is class } U, \\ 0 & \text{otherwise.} \end{cases}$$

Obviously, the general instability results in Theorems 4–6 can be rewritten here with $\hat{\mathbf{Q}}$ replaced by $\hat{\mathbf{Q}}_a$. However, it is useful to state a more specific result.

THEOREM 7. *Assume that the interconnected system is causal. Suppose the subsystems G_i are (Q_i, S_i, R_i) ultimately virtual-dissipative for $i = 1, \cdots, N$ and some are class U. Then the interconnected system is unstable if*:

A. *At least one of the Class U systems has Q_i negative definite and $\hat{Q}$ is negative semidefinite, or*

B. *At least one of class U systems has Q_i negative semidefinite and one of the following holds*:

(i) *$\hat{Q}$ is negative definite*;

(ii) *$\hat{Q}$ is negative semidefinite and*

$$\text{rank}[\hat{\mathbf{Q}} \quad \mathbf{D}] = r, \qquad \text{rank}\begin{bmatrix} \hat{\mathbf{Q}} \\ \mathbf{DH} \end{bmatrix} = \text{rank}\,\hat{\mathbf{Q}}, \tag{12}$$

where r is the total number of outputs.

Proof. Case A. Suppose G_k is Class U and Q_k is negative definite. Then G_k is linear, and from Lemma 5 we see that it is not (Q_k, S_k, R_k) dissipative. The result follows from Theorem 4 Part (i).

Case B. Since $Q_k - \alpha I$ is negative definite for $\alpha > 0$, the result follows from Case A if α can be chosen to ensure $\hat{\mathbf{Q}}_a$ is negative semidefinite. This is clearly possible if $\hat{Q}$ is negative definite. It is shown in [23] that the conditions in (ii) imply the existence of an appropriate α. □

Comments. 1) The special case of Theorem 7 Part A where all subsystems are ultimately virtual finite-gain is clearly a special case of Corollary 7.

2) In Theorem 7 it is possible to replace the Class U constraint on some subsystems by the requirement $\gamma_c < \infty$ and $\mathcal{K}(G_i) \neq \mathcal{U}_i$. The result holds if appropriate extra assumptions of unbiasedness and/or Q_i negative semidefinite are made according to Theorem 4. This version of the result represents a sharpening of a result given in [40].

3) Suppose we group all the Class U systems together as the first m numbered subsystems and partition $\mathbf{H}$ in the obvious way as

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_a \\ \mathbf{H}_b \end{bmatrix}.$$

Then the rank condition (12) can be written as

$$\text{rank}\begin{bmatrix} \hat{\mathbf{Q}} \\ \mathbf{H}_a \end{bmatrix} = \text{rank}\,\hat{\mathbf{Q}}.$$

4) If there is no Class U subsystem for which Q_k is negative semidefinite, the instability conditions must be stated as the existence of some $\alpha > 0$ such that $Q_k - \alpha I$ is negative definite for some Class U subsystem and $\hat{\mathbf{Q}}_a$ is negative semidefinite. In this case, α will not be arbitrarily small.

To illustrate one special case of Theorem 7, we consider the following passivity result for feedback systems.

COROLLARY 9. *Assume that the feedback system is causal. Suppose that the subsystems G_i are ultimately virtual pseudo strict passive for $i = 1, 2$ and G_1 is Class U. Then the interconnected system is unstable if conditions*

a) $\delta_1 \geqq 0$,

b) $\delta_1 + \varepsilon_2 \geqq 0$,

c) $\delta_2 + \varepsilon_1 \geqq 0$

are satisfied and either inequality a) *or* c) *is strict.*

Proof. The cases where a) and c) are strict follow from Theorem 7 Parts A and B(ii) respectively. □

Comments. 1) The result of Takeda and Bergen [34] has the conditions of this corollary with $\delta_1 = \varepsilon_2 = 0$ and predicts instability if $\varepsilon_1 + \delta_2 > 0$ and $G_2 x = 0$ implies $x = 0$. The present result does not require this last condition on G_2. A closer look at the proof in [34] reveals that in deriving the result, the assumption $u_1 = 0$ is made. In driving the system unstable from one input, one expects more stringent conditions, although this issue needs more attention.

2) It is straightforward to derive a large-scale version of Corollary 9 and to show it relates very closely to the results of Vidyasagar [38], [40] which are large-scale counterparts of the feedback result in [34].

5. Lyapunov instability. We have seen in Theorems 1 and 2 that dissipativeness properties of a dynamical system with a state-space representation imply the existence of energy-like functions of the state. In this section, we briefly look at how these energy functions can be used to deduce internal instability properties of interconnected systems.

Firstly, it is convenient to review the technique for producing stability results. Willems's original formulation of dissipative systems favored a state-space outlook and included the basic ideas for studying Lyapunov stability [44], [45]. Some more precise results have been presented in [15], [24].

Let x_i be the state for subsystem G_i; then $x = (x_1, x_2, \cdots, x_N)$ is the state for the interconnected system under certain well-posedness conditions [26]. For simplicity, we now assume $L_{2e}(\mathbb{R}^n)$ signal spaces and $\mathbf{x}_i \in \mathbb{R}^n$. Suppose that each subsystem $G_i(0)$ is (Q_i, S_i, R_i) dissipative and has a storage function $\phi_i : X_{ir} \to \mathbb{R}_+$ such that $\phi_i(0) = 0$. In [17], it is shown that if the subsystem also satisfies a local controllability condition in X_{ir}, then $\phi_i(\cdot)$ is continuous. For convenience, a system which is controllable reachable and locally controllable is called *strongly controllable*. If $\phi_i(\cdot)$ is continuous, the time derivative along system trajectories can be defined by

$$D^+\phi_i(\mathbf{x}_i(t)) = \overline{\lim_{h \to 0_+}} \frac{1}{h}\{\phi_i[\mathbf{x}_i(t+h)] - \phi_i[\mathbf{x}_i(t)]\}.$$

The following result is a restatement of Theorems 1 and 2 in the present context.

LEMMA 6 [17]. *Suppose that the dynamical system $G_i(0)$ is strongly controllable. Then the system is cyclo-dissipative (dissipative) if and only if there exists a continuous function $\phi_i : \mathbb{R}^{n_i} \to \mathbb{R}$ satisfying $\phi_i(0) = 0$ ($\phi_i(0) = 0$, $\phi_i(\mathbf{x}_i) \geqq 0$ for all x_i) and*

$$D^+\phi_i(\mathbf{x}_i(t)) \leqq w_i(\mathbf{u}_i(t), \mathbf{r}_i(\mathbf{x}_i(t), \mathbf{u}_i(t))) \tag{13}$$

for almost all $t \geqq t_0$ along the system trajectories, where

$$w_i(\mathbf{u}_i, \mathbf{y}_i) = \mathbf{y}_i^T \mathbf{Q}_i \mathbf{y}_i + 2\mathbf{y}_i^T \mathbf{S}_i \mathbf{u}_i + \mathbf{u}_i^T \mathbf{R}_i \mathbf{u}_i.$$

Now consider the *total storage function*

$$\phi(\mathbf{x}) = \sum_{i=1}^{N} \phi_i(\mathbf{x}_i)$$

under the assumption that all $G_i(0)$ states are strongly controllable. It is straightforward to show that

$$D^+\phi(\mathbf{x}(t)) \leqq \mathbf{y}^T \hat{\mathbf{Q}} \mathbf{y} \tag{14}$$

for $\mathbf{u}_e = 0$. Now suppose that all subsystems are zero-state-detectable. It follows under some mild restrictions on the $w_i(\cdot, \cdot)$ that $\phi_i(\mathbf{x}_i) > 0$ for all $\mathbf{x}_i \neq 0$ [17]. The total storage function ϕ is then positive definite.

Let the origin $\mathbf{x} = 0$ be an equilibrium point for the autonomous interconnected system. Then under the above-mentioned assumptions, there are two main stability criteria [15], [24]:

1. If $\hat{\mathbf{Q}}$ is negative semidefinite, the origin is Lyapunov stable.
2. If $\hat{\mathbf{Q}}$ is negative definite, the origin is locally asymptotically stable. Global asymptotic stability follows if a stronger form of zero-state-detectability is imposed on the subsystems.

Comments. 1) If $\hat{\mathbf{Q}}$ is negative semidefinite and certain subsystem combinations are zero-state-detectable, the interconnected system can be shown to be locally asymptotically stable [15].

2) The second form of result and Theorem 3 indicate a close relationship between finite-gain input-output stability and local asymptotic stability. This has been studied in [18]. We note that if the subsystems are only weakly dissipative, then the system need not be locally asymptotically stable or even stable. The discussion in [18] allows for stability of a subset of X rather than just a point.

3) The first result has no counterpart in the input-output stability results.

It is important to realize that all the above-mentioned results, although different in detail, are based on the input-output conditions for stability. That is, by imposing minimality assumptions, relaxing definiteness of $\hat{\mathbf{Q}}$, etc., the input-output stability conditions become conditions for some form of state stability. In fact, we readily see that results in all four categories of input-output stability, input-output instability, Lyapunov stability and Lyapunov instability are variants to a basic set of dissipativeness requirements for stability [13].

Now consider instability of the origin for the autonomous interconnected system. The details for the following results follow from standard Lyapunov theory [47] using the total storage function. The above minimality assumptions still hold.

THEOREM 8. *Suppose that the subsystems $G_i(0)$ are (Q_i, S_i, R_i) cyclo-dissipative for $i = 1, \cdots, N$, but not dissipative for at least one i. Then the origin is not globally asymptotically stable if $\hat{\mathbf{Q}}$ is negative semidefinite.*

Comment. As in the results to follow, the input-output property of "ultimate virtual-dissipativeness" has been replaced by the slightly weaker property "cyclo-dissipativeness" in the state-space instability results. Recall Comment 5) following Definition 10.

THEOREM 9. *Suppose that the subsystems $G_i(0)$ are (Q_i, S_i, R_i) cyclo-dissipative for $i = 1, \cdots, N$, but not dissipative for at least one i such that $\phi_i(\cdot)$ takes negative values arbitrarily close to the origin. Then the origin is not locally asymptotically stable if $\hat{\mathbf{Q}}$ is negative semidefinite and Lyapunov unstable if $\hat{\mathbf{Q}}$ is negative definite.*

Comments. 1) Suppose one of the nondissipative subsystems in Theorem 9 is linear and finite-dimensional. Then we have $\phi_i(\mathbf{x}_i) = \frac{1}{2}\mathbf{x}_i^T \mathbf{P}_i \mathbf{x}_i$, where $\mathbf{P}_i$ is found as the

solution of a set of algebraic equations [2]. Moreover, $\mathbf{P}_i$ is nonsingular under the restrictions which ensure $\phi_i(\cdot)$ to be positive definite when $G_i(0)$ is dissipative. It is easy to see that for $G_i(0)$ to be cyclo-dissipative, but not dissipative, $\mathbf{P}_i$ must have a negative eigenvalue. Hence, $\phi_i(\cdot)$ takes negative values arbitrarily close to the origin.

2) Willems has presented some general instability results for dissipative systems which are closely related to Theorems 8 and 9 [45]. However, the present results reveal more clearly an essential unity with input-output instability results.

3) In Theorem 9, if $\phi_i(\mathbf{x}_i) < 0$ in a neighborhood of the origin for some subsystem, then the origin is completely unstable.

4) As demonstrated by Willems [45], it is almost trivial to generate the known Lyapunov instability criteria [7] for feedback systems from general results such as these, once certain frequency domain connections are made. A general discussion of this is left to the next section. It is not so easy to compare the results on large-scale system Lyapunov instability [12], [20]. They start by assuming the general Lyapunov theory conditions for instability on some subsystems and combining them to demonstrate instability of the overall system.

6. Frequency domain tests for instability. The previous results have given general conditions for instability of nonlinear interconnected systems in terms of dissipativeness properties. The classical absolute stability problem deals with the stability of systems consisting of linear dynamical and nonlinear memoryless subsystems. The stability conditions on these subsystems are expressed as frequency domain and sector constraints respectively. In this section, we briefly indicate how such results for instability are easily derived from the general results.

Consider a linear, time-invariant system G with proper transfer function matrix description $\mathbf{G}(s)$. If $\mathbf{G}(s)$ has a state-space realization $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ then

$$\mathbf{G}(s) = \mathbf{D} + \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}.$$

We assume any such realization is minimal. We then have the following frequency domain tests for dissipativeness [44], [14]. The proofs are simple consequences of results concerning a certain frequency domain inequality in linear optimal control theory [43], [21].

THEOREM 10. *Let $w(\cdot, \cdot)$ have the property that for all y there exists u such that $w(u, y) \leqq 0$. Then system G is (Q, S, R) dissipative if and only if*

$$\mathbf{M}(s) = \mathbf{R} + \mathbf{S}^T\mathbf{G}(s) + \mathbf{G}^H(s)\mathbf{S} + \mathbf{G}^H(s)\mathbf{Q}\mathbf{G}(s) \tag{15}$$

is positive semidefinite for all s such that $\operatorname{Re}[s] \geqq 0$ *where $\mathbf{G}^H$ denotes the usual Hermitian transpose of $\mathbf{G}$.*

The function $w(\cdot, \cdot)$ was defined in Lemma 6.

THEOREM 11. *System G is $(\mathbf{Q}, \mathbf{S}, \mathbf{R})$ cyclo-dissipative if and only if $\mathbf{M}(jw)$, defined in* (15), *is nonnegative definite for all real w.*

For the next result we introduce the notation G_K for the system G with linear feedback $\mathbf{u} = \mathbf{u}_e - \mathbf{K}\mathbf{y}$.

THEOREM 12. *Suppose system G is $(\mathbf{Q}, \mathbf{S}, \mathbf{R})$ cyclo-dissipative and let $\mathbf{K}$ be any matrix such that $\mathbf{I} + \mathbf{K}\mathbf{G}(\infty)$ is nonsingular and*

$$\mathbf{Q}_K = \mathbf{Q} - \mathbf{S}\mathbf{K} - \mathbf{K}^T\mathbf{S}^T + \mathbf{K}^T\mathbf{R}\mathbf{K} \tag{16}$$

is negative semidefinite. Then G is $(\mathbf{Q}, \mathbf{S}, \mathbf{R})$ dissipative if the linear system G_K is asymptotically stable.

Comments. 1) As special cases of Theorems 10 and 11 we see that (cyclo-passivity) passivity corresponds to (generalized) positive real matrices and finite-gain with unity gain bound corresponds to bounded real matrices [2].

2) Theorems 11 and 12 imply that if $\mathbf{M}(jw)$ is positive semidefinite and a suitable stabilizing control exists, then the system is dissipative. This often provides a more convenient test for dissipativeness than Theorem 10.

3) Note that if $\mathbf{Q}_K$ in Theorem 12 is in fact negative definite, then Lemma 2 gives the converse relationship that dissipativeness implies $\mathbf{K}$ is a stabilizing control. (The implied equivalence is also a consequence of Lemma 5). Further, it can be shown that $\mathbf{P}_i$ in Comment 1) following Theorem 9 is nonsingular since the state-space is completely observable.

Using Theorems 10–12 to substitute frequency domain conditions for the dissipativeness properties, we can obtain frequency domain instability tests. We will now illustrate this straightforward process by deriving an instability version of the multivariable circle criterion.

Suppose G_2 is a nonlinear memoryless system described by $\mathbf{y}_2 = \boldsymbol{\psi}(\mathbf{u}_2)$, and G_1 is a linear time-invariant dynamical system with transfer function matrix $\mathbf{G}_1(s)$. Let G_2 be inside the sector $[a+\Delta, b-\Delta]$ where $b \geqq a > 0$ and $\Delta \geqq 0$. Associate G_1 with the dissipativeness triple $(ab, \frac{1}{2}(a+b), (1-b\delta)(1+a\delta))$, without being specific on the type of dissipativeness at this stage. Applying the input-output conicity result of Zames [48], its Lyapunov version [15], Corollary 5 and Theorem 9, we find the following stability conditions for the feedback system:

A. Suppose G_1 is dissipative:

1) δ, Δ both nonnegative $\Rightarrow$ Lyapunov stability;

2) δ, Δ both nonnegative and either is positive $\Rightarrow$ finite-gain stability and asymptotic stability.

B. Suppose G_1 is ultimately virtual-dissipative ($\equiv$ cyclo-dissipative in this case), but not dissipative:

1) δ, Δ both nonnegative $\Rightarrow$ input-output unstable and not asymptotically stable;

2) δ, Δ both nonnegative and either is positive $\Rightarrow$ input-output unstable and Lyapunov unstable.

To save carrying these details, in the sequel we will assume both δ, Δ are nonnegative and refer to "stability" or "instability" according to whether G_1 is dissipative or cyclo-dissipative and not dissipative.

Note that if G_1 is dissipative, then it is outside the sector $[-1/a - \delta, -1/b + \delta]$. It is convenient to introduce two parameters

$$c = -\frac{1}{2}\left(\frac{a+b}{ab}\right), \qquad r = \frac{1}{2}\left(\frac{b-a}{ab}\right) + \delta. \tag{17}$$

Then $(\mathbf{Q}_1, \mathbf{S}_1, \mathbf{R}_1) = (\mathbf{I}, -cI, (c^2 - r^2)\mathbf{I})$. Now consider linear feedback $u = u_e + y/c$; that is, $\mathbf{K} = -\mathbf{I}/c$. Then we have that $\mathbf{Q}_K = -r^2\mathbf{I}/c^2$ is negative definite. So if $\det(cI - \mathbf{G}(\infty)) \neq 0$, we can use Theorems 10, 12 to obtain frequency domain tests. The matrix $\mathbf{M}(jw)$ in (15) is given by

$$\mathbf{M}(jw) = (\mathbf{G}_1(-jw) - c\mathbf{I})^T(\mathbf{G}_1(jw) - c\mathbf{I}) - r^2\mathbf{I}.$$

The nonlinear feedback system is (stable) unstable if $\mathbf{M}(jw)$ is positive semidefinite and G_K is (asymptotically stable) unstable. Now G_K is a linear system for which stability can be assessed by well-known techniques [9], [19]. In the simple case where

$\mathbf{G}_1(jw)$ is a normal matrix, it is easy to show that testing whether $\mathbf{M}(jw)$ is positive semidefinite reduces to the scalar tests

$$|\lambda_i(\mathbf{G}_1(jw)-c|\geqq r$$

where $\lambda_i(\mathbf{G}_1(jw))$ are eigenvalues of $\mathbf{G}_1(jw)$, $i=1,\cdots,m$; such a simple test on these eigenvalues for the general case does not exist [29]. The required positivity of $\mathbf{M}(jw)$ condition is therefore left as $\lambda_{\min}(\mathbf{M}(jw))\geqq 0$ where $\lambda_{\min}$ refers to the minimum eigenvalue. An equivalent way to express this is

$$\sigma_{\max}((\mathbf{G}_1(jw)-cI)^{-1})\leqq\frac{1}{r}$$

where $\sigma_{\max}(\mathbf{F})$ denotes the largest singular value of $\mathbf{F}$ [11]. Collecting these facts together, we can state the following result.

COROLLARY 10. *Suppose that the nonlinear feedback* $\psi(\cdot)$ *is inside the sector* $[a+\Delta, b-\Delta]$ *where* $b\geqq a>0$, $\det(\mathbf{G}_1(\infty)-cI\neq 0$, *and*

$$\sigma_{\max}((\mathbf{G}_1(jw)-cI)^{-1}\leqq\frac{1}{r}$$

where r, c are given by (17). *Let* $\mathbf{G}_1(s)=\mathbf{N}(s)\mathbf{D}^{-1}(s)$ *where* $\mathbf{N}(s)$, $\mathbf{D}(s)$ *are right coprime polynomial matrices. Then the feedback system is* (*stable*) *unstable if the polynomial* $\det(c\mathbf{D}(s)-\mathbf{N}(s))$ *has* (*all its zeros with negative real parts*) *one zero with positive real part.*

Comments. 1) Recall that the exact type of stability or instability which is implied for the feedback system depends on the positivity properties of δ and Δ.

2) The test for stability of $\mathbf{G}_K(s)$ is taken from [9]. The generalized Nyquist criterion due to MacFarlane and Postlethwaite [19] could also have been used. In this case, stability or instability follows according to the number of encirclements of the "critical circle" of centre $0+jc$ and radius r in the complex plane.

3) This result provides a version of the multivariable circle criterion with its instability counterpart. The stability result was first given by Sandberg [30] in an input-output setting (with $\psi(\cdot)$ assumed to be decoupled). A scalar Lyapunov instability version of the circle criterion is due to Brockett and Lee [7] and is a special case of Corollary 10. An input-output version of the result in [7] appears in [9].

7. Conclusions. This paper gives a theory for the instability of interconnected systems. Many of the constraints used in previous results are seen to be unnecessary. Along with an earlier paper on stability [24], the paper provides an essentially complete stability theory by the so-called dissipative systems approach. An important feature is the treatment of both input-output and Lyapunov stability concepts within a unified setting.

It has been suggested by Sandell et al. [32] that input-output methods are superior to Lyapunov methods for large-scale systems. The approach taken here offers a perspective which shows that each method is basically similar in scope. However, each method needs to complement the other in order for one to fully study the stability of nonlinear systems.

Appendix A. The following notation will be used. If $\mathbf{A}$ is a square real $n\times n$ matrix, its elements will be denoted by a_{ij}, and its transpose and inverse, by $\mathbf{A}^T$ and $\mathbf{A}^{-1}$ respectively. If $\mathbf{A}$ or the vector $\mathbf{x}$ have all elements nonnegative (positive), this is represented by $\mathbf{A}\geqq\mathbf{0}$, $\mathbf{x}\geqq\mathbf{0}$ ($\mathbf{A}>0$, $\mathbf{x}>\mathbf{0}$). The matrix $|\mathbf{A}|$ will be the matrix obtained

from **A** by replacing every element by its absolute value. Matrices with nonnegative (positive) elements are generally called nonnegative (positive). By a permutation of **A**, we mean a permutation of the rows of **A** combined with the same permutation of the columns. The spectral radius of **A** is denoted $\rho(\mathbf{A})$.

DEFINITION A.1. Let **A** satisfy $a_{ii} \geqq 0$ for each i and $a_{ij} \leqq 0$ whenever $i \neq j$. Then **A** is called a *semi-M-matrix* (*M-matrix*) if all principal minors of **A** are nonnegative (positive).

DEFINITION A.2. The matrix **A** is called *semiquasidominant* (*quasidominant*) if there exists a vector $\mathbf{d} > 0$ such that

$$d_i a_{ii} \geqq \sum_{j \neq i} d_j |a_{ij}| \quad \text{for all } i \qquad \left(d_i a_{ii} > \sum_{j \neq i} d_j |a_{ij}| \quad \text{for all } i \right).$$

DEFINITION A.3. A matrix **A** is called *reducible* if there is a permutation that puts it into the form

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$$

where $\mathbf{A}_{11}$ and $\mathbf{A}_{22}$ are square matrices. Otherwise **A** is called irreducible. If both **A** and $|\mathbf{A}|^T|\mathbf{A}|$ are irreducible, we call **A** strongly irreducible.

Comments. 1) Much has been written about the theory of *M*-matrices emanating from their utility in stating results on competitive equilibrium in economics, numerical analysis, and the stability of large-scale systems [3], [6], [10], [22], [28]. For a more complete list of references, compare those cited in the surveys [6], [28]. The term semi-*M*-matrix is borrowed from Araki [3]. These matrices have also been called singular *M*-matrices [28]. Similarly, dominance properties for matrices are well known. Definition A.2 extends one in [22] to include semistrict dominance.

2) The property of irreducibility plays an important role in the theory of semi-*M*-matrices [6], [10], [28]. The property of strong irreducibility is introduced here for convenience in presenting later results.

3) Obviously, reducibility of **A** is equivalent to reducibility of $|\mathbf{A}|$ and has nothing to do with magnitudes of a_{ij}. However, examples show that the reducibility properties of **A**, $\mathbf{A}^T\mathbf{A}$, and $|\mathbf{A}|^T|\mathbf{A}|$ are independent in general.

In the remainder of this Appendix we derive a result for semi-*M*-matrices which is not available elsewhere. We are concerned with conditions on matrix **A** to ensure that matrix $\mathbf{P} - \mathbf{A}^T\mathbf{P}\mathbf{A}$ is nonnegative definite for positive definite diagonal **P**. The following facts will be useful.

FACT A.1 [10]. *Let* **A** *satisfy* $a_{ii} \geqq 0$, *for each* i, $a_{ij} \leqq 0$ *whenever* $i \neq j$, *and be irreducible. Then* **A** *is a semi-M-matrix if and only if there exists* $\mathbf{x} > \mathbf{0}$ *such that* $\mathbf{A}\mathbf{x} \geqq \mathbf{0}$.

FACT A.2. *An irreducible semi-M-matrix is semiquasidominant.*

Proof. Immediate from Fact A.1. □

FACT A.3. *The real part of each eigenvalue of a semiquasidominant matrix is nonnegative.*

Proof. Let λ be an eigenvalue of semiquasidominant matrix **A**. The matrix $\lambda\mathbf{I} - \mathbf{A}$ has diagonal elements $\lambda - a_{ii}$. Suppose that $\operatorname{Re}\lambda < 0$. Then

$$|\lambda - a_{ii}| \geqq |\operatorname{Re}\lambda - a_{ii}| > |a_{ii}|.$$

So $\lambda\mathbf{I} - \mathbf{A}$ is quasidominant. It follows that this matrix is nonsingular [22] and λ cannot be an eigenvalue of **A**. This contradiction implies that for λ to be an eigenvalue it must satisfy $\operatorname{Re}\lambda \geqq 0$. □

FACT A.4. *Let* $\mathbf{A}$ *be a nonnegative matrix. Then* $\mathbf{I}-\mathbf{A}$ *is a semi-M-matrix (M-matrix) if and only if* $\rho(A)\leqq 1$ $(\rho(A)<1)$.

Proof. Follows easily from results in [6]. □

As an intermediate step to obtaining the main result, we establish the following:

LEMMA A.1.[2] *Let* $\mathbf{A}$ *be an irreducible nonnegative matrix. Then* $\mathbf{I}-\mathbf{A}$ *is a semi-M-matrix if and only if there exists a positive definite diagonal matrix* $\mathbf{P}$ *such that* $\mathbf{P}-\mathbf{A}^T\mathbf{P}\mathbf{A}$ *is nonnegative definite.*

Proof. The steps are similar to those used by Araki [3] to relate the M-matrix property of $\mathbf{I}-\mathbf{A}$ to existence of $\mathbf{P}$ satisfying the strict version of inequality (A.1). The key tool is Fact A.1. □

THEOREM A.1. *Let* $\mathbf{A}$ *be a strongly irreducible matrix. If* $\mathbf{I}-|\mathbf{A}|$ *is a semi-M-matrix, there exists a positive definite diagonal matrix* $\mathbf{P}$ *such that* $\mathbf{P}-\mathbf{A}^T\mathbf{P}\mathbf{A}$ *is nonnegative definite.*

Proof. From Lemma A.1, it follows that there exists a positive definite diagonal matrix $\mathbf{P}$ such that $\mathbf{P}-|\mathbf{A}|^T\mathbf{P}|\mathbf{A}|$ is a semi-M-matrix. Irreducibility of $|\mathbf{A}|^T|\mathbf{A}|$ gives that $\mathbf{P}-|\mathbf{A}|^T\mathbf{P}|\mathbf{A}|$ is semiquasidominant from Fact A.2. It is then easy to see that $\mathbf{P}-\mathbf{A}^T\mathbf{P}\mathbf{A}$ must be semiquasidominant for the same $\mathbf{P}$. It then follows from Fact A.3 that $\mathbf{P}-\mathbf{A}^T\mathbf{P}\mathbf{A}$ is nonnegative definite. □

This result is a counterpart to the following one due to Moylan [22] on M-matrices which is an extension of Araki's counterpart of Lemma A.1.

THEOREM A.2 [22]. *If* $\mathbf{I}-|\mathbf{A}|$ *is an M-matrix, there exists a positive definite diagonal matrix* $\mathbf{P}$ *such that* $\mathbf{P}-\mathbf{A}^T\mathbf{P}\mathbf{A}$ *is positive definite.*

A convenient characterization of irreducibility is in terms of digraphs.

DEFINITION A.4 [6]. The *associated digraph* $\mathcal{G}(\mathbf{A})$ of matrix $\mathbf{A}$ consists of vertices labelled $1,\cdots,n$ where an edge leads from vertex i to vertex j if and only if $a_{ij}\neq 0$.

DEFINITION A.5 [6]. A digraph $\mathcal{G}$ is *strongly connected* if for any ordered pair (i,j) of vertices in $\mathcal{G}$, there exists a path which leads from vertex i to vertex j.

FACT A.5 [6]. *A matrix* $\mathbf{A}$ *is irreducible if and only if* $\mathcal{G}(\mathbf{A})$ *is strongly connected.*

Acknowledgment. The authors wish to thank Professor M. Vidyasagar of the University of Waterloo for his helpful comments on an earlier draft of the paper.

REFERENCES

[1] B. D. ANDERSON AND J. B. MOORE, *Unstable system Lyapunov function generation*, Notes on System Science 1, Dept. Electrical Engineering, University of Newcastle, 1967, pp. 39–48.

[2] B. D. ANDERSON AND S. VONGPANITLERD, *Network Analysis and Synthesis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[3] M. ARAKI, *M-matrices (matrices with non-positive off-diagonal elements and positive principal minors)*, Publication 74/19, Dept. Computing and Control, Imperial College, London, 1974.

[4] ———, *Application of M-matrices to the stability problems of composite dynamical systems*, J. Math. Anal. Appl., 52 (1975), pp. 309–321.

[5] ———, *M-matrices and their applications* IV, Systems and Control, 21 (1977), pp. 214–222 (in Japanese).

[6] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

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

[8] C. A. DESOER, *Notes for a Second Course on Linear Systems*, Van Nostrand Reinhold, New York, 1980.

[9] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input-Output Properties*, Academic Press, New York, 1975.

[2] M. Araki has pointed out that Lemma A.1 appears in [5].

[10] M. FIEDLER AND V. PTÁK, *On matrices with non-positive off-diagonal elements and positive principal minors*, Czechoslovak Math. J., 12 (1962), pp. 382–400.
[11] G. FORSYTHE AND C. B. MOLER, *Computer Solution of Linear Algebraic Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[12] LJ. T. GRUJIĆ AND D. D. ŠILJAK, *Asymptotic stability and instability of large-scale systems*, IEEE Trans. Automat. Control, 18 (1973), pp. 636–645.
[13] D. J. HILL, *On the relation between input-output and Lyapunov stability results for interconnected systems*, 21st Midwest Symposium on Circuits and Systems, Ames, Iowa, 1978, pp. 505–507.
[14] D. J. HILL AND P. J. MOYLAN, *Cyclo-dissipativeness, dissipativeness and losslessness for nonlinear dynamical systems*, Tech. Rep. EE7526, Dept. Electrical Engineering, University of Newcastle, 1975.
[15] ———, *Stability results for nonlinear feedback systems*, Automatica, 13 (1977), pp. 377–382.
[16] ———, *A general result on the instability of feedback systems*, IEEE International Symposium on Circuits and Systems, New York, 1978, pp. 652–657.
[17] ———, *Dissipative dynamical systems: Basic input-output and state properties*, J. Franklin Inst. 309 (1980), pp. 327–357.
[18] ———, *Connections between finite gain and asymptotic stability*, IEEE Trans. Automat. Control, 25 (1980), pp. 931–936.
[19] A. G. J. MACFARLANE AND I. POSTLETHWAITE, *The generalized Nyquist stability criterion and multivariable root loci*, Internat. J. Control, 25 (1977), pp. 81–127.
[20] A. N. MICHEL AND R. K. MILLER, *Qualitative Analysis of Large Scale Dynamical Systems*, Academic Press, New York, 1977.
[21] P. J. MOYLAN, *On a frequency-domain condition in linear optimal control theory*, IEEE Trans. Automat. Control, 20 (1975), pp. 806.
[22] ———, *Matrices with positive principal minors*, Linear Algebra and Appl., 17 (1977), pp. 53–58.
[23] ———, *A relaxed dissipativeness test for the stability of large-scale systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 545–547.
[24] P. J. MOYLAN AND D. J. HILL, *Stability criteria for large-scale systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 143–149.
[25] ———, *Tests for stability and instability of interconnected systems*, IEEE Trans. Automat. Control, 24 (1979), pp. 574–579.
[26] P. J. MOYLAN, A. VANNELLI AND M. VIDYASAGAR, *On the stability and well-posedness of interconnected nonlinear dynamical systems*, IEEE Trans. Circuits and Systems, 27 (1980), pp. 1097–1101.
[27] E. NOLDUS, *Criteria for unbounded motion by positive operator methods*, Internat. J. Control, 18 (1973), pp. 289–296.
[28] G. POOLE AND T. BOULLION, *A survey on M-matrices*, SIAM Rev., 16 (1974), pp. 419–427.
[29] H. H. ROSENBROCK AND P. A. COOK, *Stability and the eigenvalues of $G(s)$*, Int. J. Control, 21 (1975), pp. 99–104.
[30] I. W. SANDBERG, *On the L_2-boundedness of solutions of nonlinear functional equations*, Bell System Tech. J., 43 (1964), pp. 1581–1599.
[31] I. W. SANDBERG, *On the stability of interconnected systems*, Bell System Tech. J., 57 (1978), pp. 3031–3046.
[32] N. R. SANDELL, JR., P. VARAIYA, M. ATHANS AND M. G. SAFONOV, *Survey of decentralized control methods for large scale systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 108–128.
[33] R. A. SKOOG, *Positivity conditions and instability criteria for feedback systems*, this Journal, 12 (1974), pp. 83–98.
[34] S. TAKEDA AND A. R. BERGEN, *Instability of feedback systems by orthogonal decomposition of L_2*, IEEE Trans. Automat. Control, 18 (1973), pp. 631–636.
[35] Y. V. VENKATESH, *Converse Schwarz inequality and an instability-related property of feedback systems*, Internat. J. Systems Sci., 3 (1977), pp. 339–352.
[36] M. VIDYASAGAR, *L^2-instability criteria for interconnected systems*, this Journal, 15 (1977), pp. 312–328.
[37] ———, *L_2-stability of interconnected systems using a reformulation of the passivity theorem*, IEEE Trans. Circuits and Systems, 24 (1977), pp. 637–645.
[38] ———, *New passivity-type criteria for large-scale interconnected systems*, IEEE Trans. Automat. Control, 24 (1979), pp. 575–579.
[39] ———, *On the well-posedness of large-scale interconnected systems*, IEEE Trans. Automat. Control, 25 (1980), pp. 413–421.

[40] M. VIDYASAGAR, *New L^2-instability criteria for large-scale interconnected systems*, IEEE Trans. Circuits and Systems, 27 (1980), pp. 970–973.

[41] ———, *L_∞-instability of large-scale interconnected systems using orthogonal decomposition and exponential weighting*, IEEE Trans. Circuits and Systems, 27 (1980), pp. 973–976.

[42] J. C. WILLEMS, *Stability, instability, invertibility and casuality*, this Journal, 7 (1969), pp. 645–671.

[43] ———, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621–634.

[44] ———, *Dissipative dynamical systems, Part* I: *General theory*; *Part* II: *Linear systems with quadratic supply rates*, Arch. Rational Mech. Anal., 45 (1972), pp. 321–393.

[45] ———, *Qualitative behaviour of interconnected systems*, Ann. Systems Res., 3 (1973), pp. 61–80.

[46] ———, *Mechanisms for the stability and instability in feedback systems*, Proc. IEEE, 64 (1976), pp. 24–35.

[47] ———, *Stability Theory of Dynamical Systems*, Nelson, London, 1970.

[48] G. ZAMES, *On the input-output stability of time-varying nonlinear feedback systems. Part* I: *Conditions derived using concepts of loop gain, conicity and positivity*, IEEE Trans. Automat. Control, 11 (1966), pp. 228–238.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0008 $01.25/0

THE MATRIX RICCATI EQUATION AND THE NONCONTROLLABLE LINEAR-QUADRATIC PROBLEM WITH TERMINAL CONSTRAINTS*

PAVOL BRUNOVSKÝ† AND JOZEF KOMORNÍK‡

Abstract. It is proved that each positive semidefinite symmetric solution of the matrix Riccati equation corresponds to an optimal control problem with suitable terminal cost and constraints. The approximation scheme for the computation and characterization of the optimal cost and optimal controls of the problem with terminal constraints is extended to the noncontrollable case.

Key words. matrix Riccati, linear-quadratic, terminal constraints

Introduction. Consider the linear-quadratic optimal control problem on the interval $[s, T]$, $t_0 \leqq S \leqq T$, given by the equation

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

($x \in R^n$, $u \in R^r$), the initial state

$$x(s) = y, \tag{2}$$

the cost function

$$C_s^T(y, u) = \int_s^T c(t, x, u)\, dt + x'(T)Rx(T) \tag{3}$$

with $c(t, x, u) = x'Q(t)x + u'M(t)u$ and the terminal constraint

$$Dx(T) = 0, \tag{4}$$

D being $q \times n$, $q \leqq n$, with full rank, A, B, Q, M being continuous, Q, M symmetric, $Q \geqq 0$, $M > 0$ on $[t_0, T]$, $R \geqq 0$ symmetric.

Under the condition that the system with output $\xi = Dx$ is output controllable on $[s, T]$ for each $t_0 \leqq s < T$, we have shown in [1] that the minimal cost for this problem can be expressed by a solution of the corresponding matrix Riccati equation

$$\dot{W} + A'W + WA + Q - W'BM^{-1}B'W = 0 \tag{5}$$

(cf. also [2]) on $[t_0, T)$ that blows up for $t \nearrow T$. We have characterized this solution as a limit for $m \to \infty$ of solutions of (5) expressing the optimal cost of the corresponding unconstrained problem with cost

$$C_{s,m}^T(y, u) = C_s^T(y, u) + m\|Dx(T)\|^2 \tag{6}$$

containing a term penalizing the deviation of the response of u from the terminal subspace. Also, we have shown that the optimal control and optimal trajectory for the problem (1)–(4) are limits for $m \to \infty$ of those for the problems (1)–(3), (6).

This result can be put into an interesting context with the ideas of [3]. By associating with (5) a flow on the Grassmann manifold $GR(n)$ of n-dimensional subspaces of R^{2n}, we can prove an inverse theorem on the solutions of (5) (§ 3) and extend our results from [1] to noncontrollable problems and problems with constraints at several points (§ 4). In § 5 we show that the techniques of § 4 can be used to deal with the infinite interval problem in case the finiteness of cost is not assumed for all

* Received by the editors March 11, 1981, and in revised form October 7, 1981.
† Institute of Applied Mathematics, Comenius University, 84215 Bratislava, Czechoslovakia.
‡ Department of Probability and Statistics, Comenius University, 84215 Bratislava, Czechoslovakia.

points. Section 2 contains a summary of the items of [3] that are important for this paper.

All the necessary material about the Riccati matrix equation and the unconstrained linear-quadratic problem is summarized in [1].

2. The associated flow on $GR(n)$. Denote

$$J = \begin{pmatrix} 0 & E_n \\ -E_n & 0 \end{pmatrix}$$

where E_n is the $n \times n$ unity matrix. For $z_i = (x_i, p_i) \in R^n \times R^n$, $i = 1, 2$, denote

$$\omega(z_1, z_2) = z_1' J z_2 = x_1' p_2 - x_2' p_1:$$

ω is a skew symmetric nondegenerate form on R^{2n}. An n-dimensional linear subspace L of R^{2n} is called Lagrangian, if the restriction of ω to L vanishes, i.e. $\omega(z_1, z_2) = 0$ as soon as $z_1, z_2 \in L$. We denote the set of Lagrangian subspaces of R^{2n} by $\mathscr{L}$.

A linear differential equation in R^{2n}

$$\dot{z} = H(t)z \tag{7}$$

is called Hamiltonian if ω is its integral, i.e. ω is constant along its solutions. This is equivalent to

$$H'J + JH = 0. \tag{8}$$

By π_x we denote the natural projection of $R^{2n} = R^n \times R^n$ onto its first factor, and we denote

$$\mathscr{L}_0 = \{L \in \mathscr{L} | \pi_x(L) = R^n\}.$$

We have $L \in \mathscr{L}_0$ if and only if there exists a symmetric $n \times n$ matrix W such that

$$L = \{(x, Wx) | x \in R^n\}.$$

The (time-dependent) flow of the equation (7) carries linear subspaces into linear subspaces of the same dimension and thus generates an associated (time-dependent) flow Φ on the Grassmann manifold $GR(n)$ of the subspaces of R^{2n} of dimension n. More precisely, if $L \in GR(n)$ and we denote by $\Phi_{t,s}(L)$ the linear subspace filled by the values at t of the solutions of (7) with values in L at time s, then there is a differential equation on $GR(n)$ such that $\Phi_{t,s}(L)$ is the value at time t of its solution having L as its value at time s. Since $GR(n)$ is compact, the solutions of this equation are defined for all $t \in R$. Since ω is an integral of (7), it is invariant under Φ, i.e. $L \in \mathscr{L}$ implies $\Phi_{t,s}(L) \in \mathscr{L}$ for all $t, s \in R$.

Consider now the flow Φ on $GR(n)$ associated with the differential equation

$$\dot{x} = Ax - BM^{-1}B'p, \qquad \dot{p} = -Qx - A'p \tag{9}$$

with A, B, M, Q coming from (1), (3). The matrix

$$H = \begin{pmatrix} A, & -BM^{-1}B'p \\ -Q & -A \end{pmatrix}$$

obviously satisfies (8), which means that (9) is Hamiltonian. If $L \in \mathscr{L}$ and $L(t) = \Phi_{t,s}(L) \in \mathscr{L}_0$ for all $t \in I = (t_1, t_2)$ and some $s \in I$, then there exists a matrix function $W(t)$, $t \in (t_1, t_2)$, such that $L(t) = \{(x, W(t)x) | x \in R^n\}$. This matrix satisfies (5).

Note that although $\lim_{t \to t_0} W(t)$ may not exist for $t_0 = t_1$ or $t_0 = t_2$, $L(t) = \Phi_{t,s}(L)$ can always be extended beyond I to all R.

3. The inverse theorem. Let A, B, Q, M be as in (1) (3).

THEOREM 1. *Let $W(t)$ be a positive semidefinite solution of the matrix Riccati equation* (5) *on* $[t_0, T)$. *Then there exist a $q \leqq n$, a $q \times n$ matrix D and a positive semidefinite symmetric matrix R such that $y'W(s)y$ is the optimal cost for the problem* (1)–(4) *for* $t_0 \leqq s < T$.

Proof. Denote $L(t) = \{(x, W(t)x) | x \in R^n\}$. If $\lim_{t \to T-} W(t)$ exists then we take $D = 0$, $R = \lim_{t \to T-} W(t)$; the statement of the theorem in this case is standard [4].

If $\lim_{t \to T-} W(t)$ does not exist, then $L(T) = \lim_{t \to T-} L(t) \notin \mathscr{L}_0$. Let $q = \operatorname{codim} \pi_x(L(T)) > 0$.

There exist $n \times n$ matrices S_1, S_2 such that rank

$$(S_1, S_2) = n \quad \text{and} \quad L(T) = \{(x, p) | S_1 x + S_2 p = 0\}.$$

If $x \in \pi_x(L(T))$ then there exists a p such that $S_2 p = -S_1 x$, i.e. $S_1 x \in \operatorname{Range} S_2$. The condition rank $(S_1, S_2) = n$ means

$$\operatorname{Range} S_1 + \operatorname{Range} S_2 = R^n, \tag{10}$$

from which it follows codim Range $S_2 = q$. Consequently, there exists a $q \times n$ matrix with full rank N such that $y \in \operatorname{Range} S_2$ if and only if $Ny = 0$. From (10) it also follows that if we denote $D = NS_1$, then rank D = rank $N = q$. Also, $x \in \pi_x(L(T))$ if and only if $Dx = 0$, i.e., $x \in \operatorname{Ker} D$.

Let K be any $n \times n$ matrix, the restriction of which to Range S_2 is a right inverse of S_2, i.e. we have $S_2 K S_2 = S_2$. Then, $(x, p) \in L(T)$ if and only if $x \in \operatorname{Ker} D$ and

$$p + KS_1 x \in \operatorname{Ker} S_2. \tag{11}$$

Denote $R_0 = -KS_1$.

Since $L(T) \in \mathscr{L}$, for any $p_1, p_2 \in \operatorname{Ker} S_2$, $x_1, x_2 \in \operatorname{Ker} D$ we have

$$(R_0 x_1 + p_1)' x_2 - (R_0 x_2 + p_2)' x_1 = 0. \tag{12}$$

Choosing $x_1 = 0$ and using (12) we obtain $p'x = 0$ for any $p \in \operatorname{Ker} S_2$, $x \in \operatorname{Ker} D$. However, $p'x = 0$ for all $x \in \operatorname{Ker} D$ is equivalent to $p \in \operatorname{Range} D'$, so $\operatorname{Ker} S_2 \subset \operatorname{Range} D'$. Since rank $D = q$ = codim Range S_2 = dim Ker S_2, we have

$$\operatorname{Ker} S_2 = \operatorname{Range} D'. \tag{13}$$

Choosing $p_1 = p_2 = 0$ in (12) we have $x_1' R_0' x_2 = x_1' R_0 x_2$ for all $x_1, x_2 \in \operatorname{Ker} D$. Also, if we take any $x \in \operatorname{Ker} D$, then $(x, R_0 x) \in L(T)$. Since $L(T) = \lim_{t \to T-} L(t)$ (in $GR(n)$), there exists a sequence of points $t_i \nearrow T$, $(x_i, p_i) = (x_i, W(t_i)x_i) \in L(t_i)$, $(x_i, p_i) \to (x, R_0 x)$. Since $W(t)$ is positive semidefinite, for each $t < T$, we have $x' R_0 x = \lim_{i \to \infty} x_i' W(t_i) x_i \geqq 0$.

Denote $R_0^1 = PR_0$, $R_0^2 = (E - P)R_0$, where P is the orthogonal projection of R^n onto Ker D. For any $x_1, x_2 \in \operatorname{Ker} D$ we have

$$x_1' R_0 x_2 = x_1' R_0^1 x_2$$

and, consequently,

$$x_1' R_0^1 x_1 \geqq 0, \qquad x_1' R_0^1 x_2 = x_2' R_0^1 x_1.$$

Thus, the restriction of R_0^1 to Ker D is symmetric and positive semidefinite. Obviously, we can find an R symmetric and positive semidefinite on all R^n such that

$$R|_{\operatorname{Ker} D} = R_0^1|_{\operatorname{Ker} D}. \tag{14}$$

By (13), for $x \in \operatorname{Ker} D$, (11) is equivalent to

$$p - R_0 x = p - R_0^1 x - R_0^2 x \in \operatorname{Range} D'.$$

Since $R_0^2 x$ is orthogonal to $\operatorname{Ker} D$, we have $R_0^2 x \in \operatorname{Range} D'$, which means that (11) is equivalent to

$$p - Rx \in \operatorname{Range} D' \tag{15}$$

for all $x \in \operatorname{Ker} D$. By [5], [2], (14) is the transversality condition for the solution of the adjoint equation of the problem (1)–(4). Since $R \geqq 0$, $M(t) > 0$ and $Q(t) \geqq 0$ for $t \in [t_0, T]$, if $(x(t), p(t))$ is a solution of (9) with $Dx(T) = 0$ and $p(T)$ satisfying (15), then $x(t)$, $t \in [s, T]$ is an optimal trajectory for the problem (1), (3), (4) with initial state $x(s)$, the corresponding optimal control being generated by the feedback law

$$u(t) = -M^{-1}(t)B(t)p(t) = -M^{-1}(t)B(t)W(t)x(t) \tag{16}$$

for $s \leqq t < T$. Since $\pi_x(L(S)) = R^n$, the points $x(s)$ obtained in this way for all possible choices of $x(T)$ and $p(T)$ fill up all R^n.

We have

$$\begin{aligned} x'(s)W(s)x(s) = p(s)x(s) &= -\int_s^T \frac{d}{dt}(p(t)x(t))\,dt + p'(T)x(T) \\ &= x'(T)Rx(T) - \int_s^T [\dot{p}'(t)x(t) + p'(t)\dot{x}(t)]\,dt \\ &= x'(T)Rx(T) + \int_s^T [x'(t)Q(t)x(t) + u'(t)M(t)u(t)]\,dt. \end{aligned}$$

Since $x(t)$, $u(t)$ are the optimal trajectory and control, respectively, this completes the proof.

4. The noncontrollable problem. In this section we consider the problem (1)–(4), but unlike in [1], [2], we shall not assume that the system (1) with output $\xi = Dx$ is output controllable. It is obvious that the set of points that can be controlled to the terminal set $Dx(T) = 0$ on $[s, T]$ is a linear subspace of R^n, but for a nonautonomous problem it is moving with s in general, and it is not entirely obvious how to characterize it.

The following theorem gives two characterizations of this subspace—one in terms of the flow on $GR(n)$, the other in terms of the approximation scheme of [1]. Also, it shows that for this approximation scheme to work, the output controllability assumption is not essential.

As in [1], we denote by $\mathcal{U}_s^T(y)$ the set of controls steering the system from the point y to the terminal set (4) on $[s, T]$ and by W_m the solution of (5) satisfying the terminal condition $W_m(T) = R + mD'D$. Note that $y'W_m(s)y$ is the minimal value of the cost for the unconstrained problem (1)–(3), (6). The optimal control $u_m(t)$ for this problem is given by the optimal feedback law

$$u = -M^{-1}(t)B'(t)W_m(t)x, \tag{17}$$

i.e., we have $u_m(t) = -M^{-1}(t)B'(t)W_m(t)x_m(t)$, where $x_m(t)$ is the solution of the equation

$$\dot{x} = (A - BM^{-1}B'W_m)x$$

with $x_m(s) = y$.

Denote

$$U(s)=\{y|\mathcal{U}_s^T(y)\neq\varnothing\},$$

$$V(s)=\{y|\limsup_{m\to\infty} y'W_m(s)y<\infty\},$$

$$L(s)=\Phi_{s,T}(\{(x,p)|Dx=0,\ p-Rx\in\text{Range } D'\}).$$

THEOREM 2. *For all* $s\in[t_0,T)$,

$$U(s)=V(s)=\pi_x(L(s)). \tag{18}$$

For $y\in U(s)$, *the optimal control* $u_0(t)$ *for the problem* (1)–(4) *is given by*

$$u_0(t)=\lim_{m\to\infty}u_m(t)=-\lim_{m\to\infty}M^{-1}(t)B'(t)W_m(t)x_m(t),$$

and the optimal value of the cost is given by

$$\min_{u\in\mathcal{U}_s^T(t)} C_s^T(y,u)=C_s^T(y,u_0)=\lim_{m\to\infty}y'W_m(s)y. \tag{19}$$

Proof. First, we prove $V(s)\subset U(s)$. From (9) we obtain by simple calculation for any k, m, s fixed, $y=x_i(s)$ and $p_i(t)=W_i(t)x_i(t)$, $i=k, m$,

$$\begin{aligned}
&\int_s^T[(x_m(t)-x_k(t))'Q(t)(x_m(t)-x_k(t))+(u_m(t)-u_k(t))'M(t)(u_m(t)-u_k(t))]\,dt\\
&\quad=-\int_s^T\frac{d}{dt}[(p_m(t)-p_k(t))'(x_m(t)-x_k(t))]\,dt\\
&\quad=-(p_m(T)-p_k(T))'(x_m(T)-x_k(T))\\
&\quad=-(x_m(T)-x_k(T))'(W_m(T)x_m(T)-W_k(T)x_k(T))\\
&\quad=-(x_m(T)-x_k(T))'(R+kD'D)(x_m(T)-x_k(T))\\
&\qquad-(x_m(T)-x_k(T))'(m-k)D'Dx_m(T)\\
&\quad=-(x_m(T)-x_k(T))'(R+kD'D)(x_m(T)-x_k(T))\\
&\qquad-(m-k)x_m'(T)D'Dx_m(T)+x_k'(T)(W_m(T)-W_k(T))x_m(T).
\end{aligned} \tag{20}$$

Using the invariance of ω, we have

$$\begin{aligned}
x_k'(T)(W_m(T)-W_k(T))x_m(T)&=p_m'(T)x_k(T)-p_k'(T)x_m(T)\\
&=p_m'(s)x_k(s)-p_k'(s)x_m(s)\\
&=y'(W_m(s)-W_k(s))y.
\end{aligned}$$

Denote $\delta_{k,m}(s)=y'(W_m(s)-W_k(s))y$.

$$\bar{\delta}_k(s)=\lim_{m\to\infty}\delta_{k,m}(s);$$

$$0\leqq\delta_{k,m}(s)\leqq\bar{\delta}_k(s)<\infty,\quad\lim_{k\to\infty}\bar{\delta}_k(s)=0\quad\text{for } k<m,\ y\in V(s). \tag{21}$$

From (20) it follows that

$$\begin{aligned}\delta_{k,m}(s) = \int_s^T & [(x_m(t)-x_k(t))'Q(t)(x_m(t)-x_k(t)) \\ & +(u_m(t)-u_k(t))'M(t)(u_m(t)-u_k(t))]\,dt \\ & +(x_m(T)-x_k(T))'(R+kD'D)(x_m(T)-x_k(T)) \\ & +(m-k)x_m'(T)D'Dx_m(T).\end{aligned} \tag{22}$$

Since all the right-hand side terms are nonnegative, we have

$$\begin{aligned}&(m-k)\|Dx_m(T)\|^2 = (m-k)x_m'(T)D'Dx_m(T) = \delta_{k,m}(s) \leqq \bar{\delta}_k(s), \\ &0 \leqq \|Dx_m(T)\|^2 \leqq \frac{1}{m-k}\bar{\delta}_k(s),\end{aligned} \tag{23}$$

and, by (21),

$$\lim_{m\to\infty} \|Dx_m(T)\| = 0.$$

Also, from (22) it follows that

$$\sup_m \int_s^T (u_m(t)-u_k(t))'M(t)(u_m(t)-u_k(t))\,dt \leqq \bar{\delta}_k(s).$$

Since $M(t)$ is continuous and positive definite on $[s, T]$, it is uniformly positive definite on $[s, T]$. From this and (21) it follows that $\{u_m\}$ is a Cauchy sequence in $L_2(s, T)$ and therefore has a limit $u_0(t)$ in $L_2(s, T)$. From the representation of $x_m(t)$ by the variation of constant formula it follows immediately that $\{x_m\}$ converges uniformly to the response $x_0(t)$ of $u_0(t)$ satisfying $x_0(s) = y$.

By (23), we have

$$Dx_0(T) = 0.$$

This proves $V(s) \subset U(s)$ and also the second equality of (19). To prove the first equality (having as its consequence the optimality of u_0) we note that for each $u \in \mathcal{U}_s^T(y)$ we have

$$C_s^T(y, u) = C_{s,m}^T(y, u) \geqq \min C_{s,m}^T(y, u) = y'W_m(s)y.$$

This also proves $U(s) \subset V(s)$. To complete the proof of the theorem it remains to prove the second equality of (18).

If $y \in \pi_x(L(s))$ then there exists a solution $(x(t), p(t))$ of (9) with $Dx(T) = 0$ such that $x(s) = y$. The function $x(t)$ is a response of the control $u(t) = -M^{-1}(t)B'(t)p(t)$ which means $u \in \mathcal{U}_s^T(y)$. Consequently, $\mathcal{U}_s^T(y) \neq \varnothing$ and $y \in U(s)$.

On the other hand, if $y \in U(s)$, then by [5], there exists an optimal control u_0 in $\mathcal{U}_s^T(y)$, the response $x_0(t)$ of which, together with a suitable function $p(t)$, satisfies (9). In addition, $p(T)$ satisfies the transversality condition (15). This proves $y \in \pi_x(L(s))$.

Remark 1. Since $u_m \to u_0$ in $L_2(s, T)$ we have

$$\lim_{m\to\infty} C_s^T(y, u_m) = C_s^T(y, u_0). \tag{24}$$

On the other hand, we have

$$\begin{aligned}C_s^T(y, u_0) &= \lim_{m\to\infty} y'W_m(s)y = \lim_{m\to\infty} C_{s,m}^T(y, u_m) \\ &= \lim_{m\to\infty} C_s^T(y, u_m) + m\|Dx_m(T)\|^2.\end{aligned} \tag{25}$$

From (24), (25) we obtain

$$\lim_{m\to\infty} m\|Dx_m(T)\|^2 = 0,$$

or

$$\|Dx_m(T)\|^2 = o(m^{-1/2}).$$

This gives an estimate for the deviation of the endpoint of the optimal trajectory of the approximate unconstrained problem from the terminal set.

Remark 2. From $\pi_x(L(s)) = U(s)$ it follows that the dimension of $\pi_x(L(s))$ cannot decrease with s decreasing. From [6] it follows that for A, B analytic it is constant for $s < T$ and equal to the dimension of the space $\operatorname{Ker} D + C$, where $C = \operatorname{span}\{b_i(T), (\mathscr{A}b_i)(T), \cdots, (\mathscr{A}^{n-1}b_i)(T) | i = 1, \cdots, n\}$, where b_i are the column vectors of B and $\mathscr{A}f(t) = f(t) - Af(t)$ for a differentiable function f on $[t_0, T]$.

Theorem 2 allows us to deal with the problem (1)–(3) with additional constraints and costs at intermediate points of the interval. We shall restrict ourselves to the case of one intermediate point, the extension to the case of a higher number of points being straightforward.

Let $T_1 \in (t_0, T)$, $q_1 \leqq n$ and let $R_1 \geqq 0$, D_1 be $n \times n$ symmetric and $q_1 \times n$ with full rank, respectively. Consider the problem given by the system (1), the initial point (2), the cost function

$$\tilde{C}_s^T(y, u) = C_s^T(y, u) + x'(T_1)R_1x(T_1), \tag{26}$$

the constraints (4) and

$$D_1x(T_1) = 0. \tag{27}$$

Of course, for $s \in (T_1, T]$ the problem coincides with the problem (1)–(4).

Let $U(t)$, $W_m(t)$ be defined as in Theorem 2. It is obvious that the optimal control for the problem (1), (2), (26), (4), (27) for $s = T_1$ will be a concatenation of the optimal control on $[s, T_1]$ for the problem (1), (2), the cost function

$$\int_s^{T_1} c(t, x, u)\, dt + x'(T_1)R_1x(T_1) + \lim_{M\to\infty} x'(T_1)W_m(T_1)x(T_1)$$

and the linear constraint

$$x(T_1) \in U(T_1) \cap \operatorname{Ker} D_1,$$

and the optimal control for the problem (1), (3), (4), with initial point $x(t_1)$ on $[t_1, T]$.

5. The infinite interval. Consider the unconstrained problem (1), (3) with $R = 0$ and denote W^T the corresponding solution of (5), which is the solution satisfying $W^T(T) = 0$. For fixed s, y, denote u^T, x^T the optimal control and trajectory respectively. In [1], we have shown that $\lim_{T\to\infty} W^T(s)$ exists and represents the optimal cost for the infinite interval problem, provided for each s, y there exists a u such that $C_s^\infty(y, u) = \lim_{T\to\infty} C_s^T(y, u) < \infty$. Like Theorem 2, the following theorem deals with problems not satisfying this condition.

By $U^\infty(s)$ we denote the set of those $y \in R^n$ for which there is a control u on $[s, \infty)$ such that $C_s^T(y, u) < \infty$. Further, we denote

$$V^\infty(s) = \left\{ y \,\middle|\, \lim_{T\to\infty} y'W^T(s)y < \infty \right\}.$$

By $L^2_M(s, \infty)$ we denote the space of functions $u:[s, \infty) \to R^r$ which are square integrable with weight $M(t)$, i.e., $\int_s^\infty u'(t)M(t)u(t)\,dt < \infty$. $L^2_M(s, \infty)$ is a Banach space.

THEOREM 3. *We have $U^\infty(s) = V^\infty(s)$ for every $s \geqq t_0$. For $y \in U^\infty(s)$ we have*

$$\min_u C_s^\infty(y, u) = \lim_{T\to\infty} y'W^T(s)y.$$

The optimal control $u^\infty(t)$ and trajectory $x^\infty(t)$ are given by

$$u^\infty(t) = \lim_{T\to\infty} u^T(t) \qquad (\text{in } L^2_M(s, \infty)), \tag{28}$$

$$x^\infty(t) = \lim_{T\to\infty} x^T(t) \tag{29}$$

(uniformly on each finite interval).

Let us note that in (28), (29) we understand u^T, x^T to be extended to $[s, \infty)$ by having value 0 for $t > T$.

Proof. Let $y \in V^\infty(s)$, $T_2 = T_1 \geqq s$. Denote $W^{T_i} = W_i$, $x^{T_i} = x_i$, $u^{T_i} = u_i$, $i = 1, 2$. By computations similar to those leading to (20) we obtain

$$\begin{aligned} y'(W_1(s) - W_2(s))y &= \int_s^T [(x_1(t) - x_2(t))'Q(t)(x_1(t) - x_2(t)) \\ &\qquad + (u_1(t) - u_2(t))'M(t)(u_1(t) - u_2(t))]\,dt \\ &\quad + (x_1(T_1) - x_2(T_1))W_1(T_1)(x_1(T_1) - x_2(T_1)) \\ &\quad + x_2(T_1)(W_2(T_1) - W_1(T_1))x_2(T_1) \\ &= \int_s^{T_1} [(x_1(t) - x_2(t))'Q(t)(x_1(t) - x_2(t)) + (u_1(t) \\ &\qquad - u_2(t))'M(t)(u_1(t) - u_2(t))]\,dt \\ &\quad + \int_{T_1}^{T_2} [x_2(t)'Q(t)x_2(t) + u_2'(t)M(t)u_2(2)]\,dt \\ &\geqq \int_s^{T_1} (u_1(t) - u_2(t))'M(t)(u_1(t) - u_2(t))\,dt. \end{aligned} \tag{30}$$

From the estimate (30) it follows that the family of functions $\{u_T | T \geqq s\}$ is a Cauchy family in $L^2_M(s, \infty)$. Since $L^2_M(s, \infty)$ is complete, it has a limit $u \in L^2_M(s, \infty)$. From the variation of constants formula it follows immediately that the response x^∞ of u^∞ is a pointwise limit of the functions x^T, the convergence being uniform on each finite subinterval of $[s, \infty)$.

For every fixed $T_0 \geqq s$ we have

$$C_s^{T_0}(y, u^\infty) = \lim_{T\to\infty} C_s^{T_0}(y, u^T) = \lim_{T\to\infty} C_s^T(y, u^T) = \lim_{T\to\infty} y'W^T(s)y,$$

from which it follows that $C_s^\infty(y, u^\infty)$ is finite and, thus, that $V^\infty(s) \subset U^\infty(s)$. On the other hand, we have for any control u,

$$C_s^{T_0}(y, u) \geqq y'W^{T_0}(s)y. \tag{31}$$

From (30), (31) it follows that

$$C_s^\infty(y, u) \geqq C_s^\infty(y, u^\infty) = \lim_{T\to\infty} y'W^T(s)y,$$

which implies that u^∞ is optimal.

The inclusion $U^{\infty}(s) \subset V^{\infty}(s)$ follows immediately from (31).

Note added in proof. There is an overlap of our § 4 and the paper of G. Chen and W. Mills, *Finite elements and terminal penalization for quadratic cost optimal control problems governed by ordinary differential equations*, this Journal, 19 (1981), pp. 744–764. In particular, the essential part of Theorem 3 or our paper is contained in Theorem 2.2 of the quoted paper.

REFERENCES

[1] P. BRUNOVSKÝ AND J. KOMORNÍK, *The Riccati equation solution of the linear-quadratic problem with constrained terminal state*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 398–402.

[2] B. FRIEDLAND, *On solutions of the Riccati equation in optimization problems*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 303–304.

[3] R. HERMANN, *Cartanian Geometry, Nonlinear Waves and Control Theory*, Part A, Math. Science Press, Brookline, MA, 1979.

[4] V. KUČERA, *A review of the matrix Riccati equation*, Kybernetika (Prague), 9 (1973), pp. 42–61.

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

[6] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of non-linear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0009 $01.25/0

GENERALIZED HERMITE MATRICES AND COMPLETE INVARIANTS OF STRICT SYSTEM EQUIVALENCE*

D. HINRICHSEN† AND D. PRÄTZEL-WOLTERS†

Abstract. A complete list of invariants for reachable system matrices

$$\Sigma(s)=\begin{bmatrix} P(s) & -Q(s) \\ V(s) & W(s) \end{bmatrix}$$

with respect to strict system equivalence (s.s.e.) is determined by polynomial methods. The polynomial input-output pairs (u, y) for which there exists a polynomial vector z such that $Pz = Qu$ and $y = Vz + Wu$ form a $K[s]$-module $\mu(\Sigma)$. It is shown that the unique basis matrix of $\mu(\Sigma)$ in Hermite form yields a complete set of discrete ("Hermite indices") (resp. continuous) invariants of s.s.e. The Hermite invariants are characterized in state space terms, and a realization of $\Sigma(s)$ in Hermite canonical form is presented. Nice orders and generalized Hermite forms are introduced in order to develop a framework that encompasses Hermite invariants and Kronecker invariants. Hermite's theorem is generalized to these matrices. Finally, nice orders are used to single out unique representatives among all minimal bases of a given full submodule $M \subset \mathbb{K}[s]^m$ and Forney's echelon form is characterized in this framework.

Key words. invariants, canonical forms, system matrices, strict system equivalence, Hermite form, canonical realization, Kronecker invariants, Hermite's theorem, minimal basis, echelon form

1. Introduction.

For linear systems in differential operator representation

$$P\left(\frac{d}{dt}\right)z = Q\left(\frac{d}{dt}\right)u, \tag{1.1a}$$

$$y = V\left(\frac{d}{dt}\right)z + W\left(\frac{d}{dt}\right)u, \tag{1.1b}$$

the concept of *strict system equivalence* (s.s.e.) as defined by Rosenbrock (1970) and characterized by Fuhrmann (1977) has proven to be an adequate generalization of the similarity relation between state space systems. In this paper we study polynomial methods to determine complete lists of invariants for s.s.e., polynomial canonical forms and canonical state space realizations for reachable systems of type (1.1).

To be more precise, let us recall the following terminology: If S is a set and $\sim$ an equivalence relation on S then a family $(f_i)_{i\in I}$ of functions $f_i : S \to \mathbb{R}$ is called a *complete list of (numerical) invariants for* $\sim$ if

$$s \sim \hat{s} \Leftrightarrow f_i(s) = f_i(\hat{s}) \quad \text{for all } i \in I. \tag{1.2}$$

A mapping $c : s \mapsto c(s)$ from S into itself is called a *canonical form* on S with respect to $\sim$ if $c(s) \sim s$ and for $s, \hat{s} \in S$:

$$s \sim \hat{s} \Leftrightarrow c(s) = c(\hat{s}). \tag{1.3}$$

For reachable linear *state space systems* various canonical forms (with respect to similarity) have been proposed in the literature (e.g. Popov (1972), Mayne (1972), Weinert and Anton (1972), Denham (1974), Rissanen (1974)). In particular it was observed that most of the so-called standard or canonical forms presented in earlier publications (Brunovsky (1966), Luenberger (1967), Rosenbrock (1970)) do not satisfy

* Received by the editors December 3, 1980, and in revised form October 30, 1981.

† Forschungsschwerpunkt Dynamische Systeme, Universität Bremen, Bibliothekstraße, Postfach 330440, 2800 Bremen 33, West Germany.

the requirements of the above definition (cf. Popov (1972), Denham (1974)). From a practical point of view the interest in canonical forms was mainly motivated by problems of identification. The usefulness of canonical forms for identification is discussed—partly in comparison with other parametrization methods—by Mayne (1972), Denham (1974) and Glover and Willems (1974). The theoretically very interesting problem to find a complete list of invariants for general linear systems (without the reachability assumption) still appears to be unsolved.

For reachable state space systems, two basic approaches can be discerned in the literature. Some authors (e.g. Denham (1974)) reduce the set $\mathcal{R}$ of all reachable pairs $(A, B) \in \mathbb{R}^{n\times(n+m)}$ to the subset $\mathcal{R}^\gamma$ of all pairs for which the determinant of an arbitrarily fixed "nice selection" $R_\gamma(A, B)$ of n column vectors of the reachability matrix

$$R(A,B)=[B \quad AB \quad A^2B \quad \cdots \quad A^{n-1}B]$$

is nonzero. It is easily verified that the mapping $(A, B) \to (R_\gamma(A, B)^{-1}AR_\gamma(A, B), R_\gamma(A, B)^{-1}B)$ defines a canonical form *on* $\mathcal{R}^\gamma$ with respect to similarity. However, different subsets $\mathcal{R}^\gamma$ corresponding to different nice selections γ may overlap since γ is not uniquely determined for a given pair $(A, B) \in \mathcal{R}$. Therefore the set of parameters obtained by this method does not constitute a complete list of similarity invariants. Rather, it should be viewed as a *local* coordinate system or *chart* on the quotient space $\mathcal{R}/\sim$ of similarity classes. Hazewinkel and Kalman (1976) and Hazewinkel (1977) used these charts in order to define a manifold structure on $\mathcal{R}/\sim$. Analyzing the manifold thus defined, they were able to show that there does not exist a *continuous* global canonical form on $\mathcal{R}$.

However, this result should not prejudice the investigation of global canonical forms for linear systems. As the Jordan normal form for single matrices illustrates, global canonical forms may be very useful without being continuous. This leads us to a second approach to our problem which has been adopted in the papers of Popov, Mayne and Weinert and Anton. The Jordan normal form is determined by a combination of discrete parameters specifying the sizes of the Jordan blocks and of continuous parameters specifying the corresponding eigenvalues. Analogously, we shall describe the classes of strictly system equivalent systems of type (1.1) by a family of discrete indices (integers) and a set of continuous parameters whose number depends on the list of indices of the system.

We proceed as follows. In § 1 we introduce the module of return to zero $\mu(P, Q)$ associated with any system of the form (1.1) and characterize the relation of s.s.e. for reachable systems via this module. This leads us to the problem of finding a parametrization procedure for all full submodules of the free module $\mathbb{K}[s]^m$ of vector polynomials over a field $\mathbb{K}$. In § 2 the Hermite basis and Hermite indices of $\mu(P, Q)$ are defined from which a complete list of s.s.e. invariants can be derived. For any given list of Hermite indices the associated number of "continuous" parameters is determined. Subsequently we describe the Hermite basis in terms of a state space representation of the system.

In § 3 we present a general method for the generation of canonical forms with respect to s.s.e. We define the concept of *nice order* and associate with it a *generalized Hermite form* of which the Hermite form is a special case. For these generalized Hermite forms, an existence and uniqueness theorem is proved which extends the classical theorem of Hermite. It is shown that basis matrices of $\mu(P, Q)$ of generalized Hermite form yield again complete lists of invariants for s.s.e. (under the assumption of reachability).

In § 4 we consider the problem of singling out uniquely determined representatives among all *minimal bases* of a given full submodule $M \subset \mathbb{K}[s]^m$ (cf. Forney (1975, p. 499)). We characterize those generalized Hermite matrices which are column proper and determine those nice orders for which *all* generalized Hermite matrices are column proper. Finally, we characterize Forney's echelon form in terms of generalized Hermite matrices and derive from it an analogue of Forney's existence and uniqueness theorem for basis matrices of full submodules $M \subset \mathbb{K}[s]^m$ in echelon form.

2. Strict system equivalence and the module of return to 0. We consider time-invariant linear finite-dimensional control systems described by system-matrices

$$\Sigma(s) = \begin{bmatrix} P(s) & -Q(s) \\ V(s) & W(s) \end{bmatrix} \in \mathbb{K}[s]^{(r+p)\times(r+m)} \tag{2.1}$$

where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. Throughout this paper we assume $P(s)$ to be nonsingular, $\Sigma(s)$ to be reachable (i.e., $P(s)$, $Q(s)$ left coprime) and the transfer function

$$G(s) = V(s)P(s)^{-1}Q(s) + W(s) \tag{2.2}$$

of $\Sigma(s)$ to be proper rational. $\Sigma(s)$ determines the following system equations:

$$\Sigma(D)\begin{pmatrix} z \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix},$$

i.e.,

$$P(D)z = Q(D)u, \tag{2.3a}$$

$$y = V(D)z + W(D)u, \tag{2.3b}$$

where u is the control function, y the output function and z a vector of internal variables of the system. In the continuous time case, u, z, y are functions on the time domain $\mathbb{R}$ with values in $U = \mathbb{K}^m$, $Z = \mathbb{K}^r$ and $Y = \mathbb{K}^p$, respectively, and $D = d/dt$ is the ordinary differential operator. In the discrete time case, u, z, y are functions on the time domains $\mathbb{Z}$ or $\mathbb{N}$, and D has to be interpreted as the left shift operator defined by $(Dx)(k) = x(k+1)$. In the sequel we analyze (2.3) in s-domain and use the discrete time interpretation for the purpose of illustration.

Let us consider the following slightly modified version of Fuhrmann's canonical state space model for $\Sigma(s)$ (Fuhrmann (1977)). As state space we choose the quotient module

$$X := Z[s]/P(s)Z[s]$$

and define the state space model $\Sigma_{\text{state}} = (A, B, C, E)$ by

$$\begin{aligned} &A : X \to X, && [z] \mapsto [sz], \\ &B : U \to X, && v \mapsto \pi_p Q(s)v, \\ &C : X \to Y, && [z] \mapsto (VP^{-1}z)_{-1}, \\ &E : U \to Y, && v \mapsto (VP^{-1}Q + W)_0 v, \end{aligned} \tag{2.4}$$

where $\pi_p : Z[s] \to Z[s]/P(s)Z[s]$ is the natural projection, $[z] = \pi_p(z)$ denotes the equivalence class of $z \in Z[s]$ mod $P(s)Z[s]$, $(VP^{-1}z)_{-1}$ and $(VP^{-1}Q + W)_0$ denote the coefficients of s^{-1}(resp. s^0) in the Laurent series representation of $V(s)P(s)^{-1}z(s)$ (resp. $V(s)P(s)^{-1}Q(s) + W(s)$).

The *reachability map* of $\Sigma(s)$ is defined by

$$\varphi_{(P,Q)} : U[s] \to X, \qquad u(s) \mapsto \pi_p Q(s) u(s).$$

Clearly $\varphi_{(P,Q)}$ is a $\mathbb{K}[s]$-module homomorphism associating with any polynomial $u(s) = \sum_{i=0}^{k} u_i s^i \in U[s]$ the state to which the discrete-time system Σ_{state} is steered by the control sequence $(u_k, u_{k-1}, \cdots, u_0)$ from the initial state 0. This justifies the following:

DEFINITION 2.1. If $\Sigma(s)$ is a system matrix of form (2.1) then the $\mathbb{K}[s]$-submodule of $U[s]$,

$$\mu(P, Q) := \operatorname{Ker} \varphi_{(P,Q)} = \{u \in U[s];\ P(s)^{-1} Q(s) u(s) \in Z[s]\},$$

is called its *module of return to zero.*

Obviously, the quotient module $U[s]/\mu(P, Q)$ is isomorphic to the $\mathbb{K}[s]$-module X. Hence $\mu(P, Q)$ is a (free) submodule of $U[s]$ of full rank m. (cf. Rosenbrock (1972), Hautus and Heymann (1978), Münzner and Prätzel-Wolters (1979)).

It follows from the definition that for any polynomial system matrix $\Sigma(s)$, the module of return to zero can be described in terms of the canonical state space model (2.4):

$$u(s) = \sum_{i=0}^{k} v_i s^i \in \mu(P, Q) \Leftrightarrow \sum_{i=0}^{k} A^i B v_i = 0. \tag{2.5}$$

Another characterization of the elements of $\mu(P, Q)$ is obtained if we introduce $b_i := Be_i$, $i = 1, \cdots, m$, where $(e_1, \cdots, e_m)$ is the standard basis of $U = \mathbb{K}^m$. Then

$$u(s) = (u_1(s), \cdots, u_m(s))^T \in \mu(P, Q) \Leftrightarrow \sum_{j=1}^{m} u_j(A) b_j = 0. \tag{2.6}$$

The following proposition which is a direct consequence of (Prätzel-Wolters (1980, Cor. 3.14)) shows how the module $\mu(P, Q)$ can be used to characterize strict system equivalence of reachable systems.

PROPOSITION 2.2. *Let $\Sigma(s)$, $\hat{\Sigma}(s)$ be system matrices with transfer functions $G(s)$* (resp. $\hat{G}(s)$). *Then the following two conditions are equivalent*:

(i) $\Sigma(s)$ *and* $\hat{\Sigma}(s)$ *are* s.s.e.,

(ii) $\mu(P, Q) = \mu(\hat{P}, \hat{Q})$ *and* $G(u) = \hat{G}(u)$ *for all* $u \in \mu(P, Q)$.

By definition, $Gu = (VP^{-1}Qu + Wu) \in Y[s]$ for all $u \in \mu(P, Q)$. Hence

$$\mu(\Sigma) := \left\{ \begin{pmatrix} u \\ y \end{pmatrix};\ u \in \mu(P, Q) \wedge y = Gu \right\} \tag{2.7}$$

is a submodule of the (free) $\mathbb{K}[s]$-module $U[s] \times Y[s]$. Evidently, condition (ii) in Proposition 2.2 can be written equivalently as

$$\mu(\Sigma) = \mu(\hat{\Sigma}). \tag{2.8}$$

This shows that the class of strict system equivalence $[\Sigma(s)]_{\text{s.s.e.}}$ is completely determined by the module $\mu(\Sigma)$. It remains to parametrize the submodules of $U[s] \times Y[s]$.

3. Hermite indices and invariants for s.s.e. With every full submodule $M \subset \mathbb{K}[s]^m$ there is associated the set of basis matrices

$$B(M) := \{D(s) \in \mathbb{K}[s]^{m \times m};\ D(s)\mathbb{K}[s]^m = M\}. \tag{3.1}$$

If $D_0(s) \in B(M)$ then

$$B(M) = \{D_0(s)U(s);\ U(s) \in \mathbb{K}[s]^{m \times m} \text{ unimodular}\}. \tag{3.2}$$

By definition, every matrix in $B(M)$ yields a complete description of M. In order to obtain a suitable parametrization we have to pick out representatives of the sets $B(M)$ in a systematic way.

DEFINITION 3.1. A polynomial matrix $D(s) \in \mathbb{K}[s]^{m\times m}$ is called *of (lower) Hermite form* if for $i = 1, \cdots, m$ and $j = 1, \cdots, m$:

(i) $d_{ij} = 0$ if $i < j$;

(ii) $\deg d_{ij} < \deg d_{ii}$ if $i > j$;

(iii) d_{ii} is monic.

The following proposition is due to Hermite (cf. Newman (1972)).

PROPOSITION 3.2. *For every matrix $D(s) \in \mathbb{K}[s]^{m\times m}$ of full rank m there exists exactly one matrix $H(s) \in \mathbb{K}[s]^{m\times m}$ in Hermite form which is right-equivalent to $D(s)$, i.e., $H(s) = D(s)U(s)$ with $U(s)$ unimodular.*

Nonsquare matrices

$$D(s) = \begin{bmatrix} D_1(s) \\ D_2(s) \end{bmatrix} \in \mathbb{K}[s]^{(m+p)\times m}, \qquad \det D_1(s) \not\equiv 0, \tag{3.3}$$

are called of (lower) Hermite form if $D_1(s)$ fulfills Definition 3.1. Let $\Sigma(s)$ be a polynomial system matrix. Then every matrix $D(s) \in B(\mu(\Sigma))$ is of the form (3.3) with $D_1(s) \in B(\mu(P, Q))$. Therefore, by Propositions 3.2 and 2.2 we have:

PROPOSITION 3.3.

(i) *For every system matrix $\Sigma(s)$ there exists exactly one matrix $H(\Sigma) \in B(\mu(\Sigma))$ of Hermite form.*

(ii) *Two system matrices $\Sigma(s)$, $\hat{\Sigma}(s)$ are* s.s.e. *if and only if $H(\Sigma) = H(\hat{\Sigma})$.*

$H(\Sigma)$ can be written in the form

$$H(\Sigma) = \begin{bmatrix} H_1(s) \\ H_2(s) \end{bmatrix}, \qquad H_1(s) \in \mathbb{K}[s]^{m\times m} \text{ nonsingular.} \tag{3.4}$$

Now consider the system matrix

$$\Sigma_H = \begin{bmatrix} H_1(s) & I_m \\ H_2(s) & 0_{p\times m} \end{bmatrix} \in \mathbb{K}[s]^{(m+p)\times(m+m)}. \tag{3.5}$$

Obviously

$$\mu(H_1(s), I_m) = H_1(s)\mathbb{K}[s]^m = \mu(P, Q),$$

hence

$$\mu(\Sigma) = \mu(\Sigma_H).$$

As a corollary we obtain Rosenbrock's existence and uniqueness theorem for his standard form of polynomial system matrices.

COROLLARY 3.4 (Canonical polynomial form of reachable system matrices). *Every reachable system $\Sigma(s)$ is* s.s.e. *to a uniquely determined system matrix of the form* (3.5) *where $H_1(s)$ is of Hermite form.*

DEFINITION 3.5. For any system matrix $\Sigma(s)$ the uniquely determined matrix $H(\Sigma) = (h_{ij}) \in B(\mu(\Sigma))$ of Hermite form is called the *Hermite matrix* of $\Sigma(s)$ and the family $\delta = (\delta_1, \cdots, \delta_m)$ defined by

$$\delta_i = \deg h_{ii}, \qquad i = 1, \cdots, m,$$

is called list of *Hermite indices* of (P, Q).

It should be noted that the list is a *family* not a *set*: the ordering of the δ_i, $i = 1, \cdots, m$, is important.

Let n be the dimension of the state space of $\Sigma(s)$. Suppose that $H(\Sigma)$ is written in the form (3.4). Since

$$\det H_1(s) = h_{11}(s) \cdot \cdots \cdot h_{mm}(s)$$

and

$$\deg \det H_1(s) = \dim_{\mathbb{K}} \mathbb{K}[s]^m / H_1(s)\mathbb{K}[s]^m = n,$$

it follows that

$$\sum_{i=1}^{m} \delta_i = n. \tag{3.6}$$

In the following we investigate the question of how many scalar parameters are needed in order to determine the Hermite matrix

$$H(\Sigma) = \begin{bmatrix} H_1(s) \\ H_2(s) \end{bmatrix} \in \mathbb{K}[s]^{(m+p)\times m}$$

of a system matrix $\Sigma(s) \in \mathscr{R}(\delta)$ where $\delta = (\delta_1, \cdots, \delta_m)$ is any finite sequence of integers $\delta_i \geqq 0$ with $\delta_1 + \cdots + \delta_m = n$ and $\mathscr{R}(\delta)$ denotes the set of all *reachable* system matrices with Hermite list δ.

Firstly, we will show that the parametrization of the output part $H_2(s)$ can be carried through independently of $H_1(s)$; in particular the number of scalar parameters required for the description of $H_2(s)$ is independent of δ.

The following lemma is easily proved.

LEMMA 3.6. *Let $D(s) \in \mathbb{K}[s]^{m\times m}$ be a column proper matrix with column degrees $\nu_1, \cdots, \nu_m$ and $L(s) \in \mathbb{K}[s]^{p\times m}$. Then the following conditions are equivalent*:

(i) *$L(s)D(s)^{-1}$ is proper rational*;

(ii) *$\begin{bmatrix} D(s) \\ L(s) \end{bmatrix}$ has column degrees $\nu_1, \cdots, \nu_m$.*

In general the Hermite matrix $H_1(s)$ is not column proper and the column degrees of $H_2(s)$ are not necessarily bounded by $\delta_1, \cdots, \delta_m$. Therefore the coefficients of the polynomial entries in $H_2(s)$ are not suitable for the parametrization of the output part: If $H_1(s)$ is fixed these coefficients may *not* be varied independently without eventually violating the condition that $G(s) = H_2(s)H_1(s)^{-1}$ has to be proper rational. Instead we have to consider the coefficients of the polynomial entries of $L(s) = H_2(s)U(s)$ where $U(s)$ is a unimodular matrix in $\mathbb{K}[s]^{m\times m}$ such that $D(s) = H_1(s)U(s)$ is column proper. Since the column degrees of $L(s)$ have to be smaller than or equal to the column degrees $\nu_1, \cdots, \nu_m$ of $D(s)$ (by the preceding lemma) and $\sum_{i=1}^{m} \nu_i = n$ we obtain a complete set of $p \cdot (n+m)$ independent parameters for the description of $H_2(s)$ whenever $H_1(s)$ is given.

Remark 3.7. The parametrization of the output part depends upon the choice of $U(s)$ which is not uniquely determined by the requirement that $H_1(s)U(s)^{-1}$ be column proper. In § 5 we present another canonical basis matrix of $\mu(\Sigma)$ which exhibits directly the free parameters of the output matrix.

Let us now determine a complete set of independent parameters for the input part $H_1(s)$. The entries h_{ij} of $H_1(s)$ are of the form

$$h_{ij} = a_0^{ij} + a_1^{ij} s + \cdots + a_{\delta_i}^{ij} s^{\delta_i}$$

where

$$a_{\delta_i}^{ij} = \begin{cases} 1 & \text{if } j = i, \\ 0 & \text{if } j \neq i. \end{cases}$$

Therefore each entry of the ith row of $H_1(s)$ is determined by δ_i parameters ($i = 1, \cdots, m$). Hence the total number $N(\delta)$ of free parameters of $H_1(s)$ for $H(\Sigma) \in \mathcal{R}(\delta)$ is

$$N(\delta) = \sum_{i=1}^{m} i\delta_i = \sum_{i=1}^{m} \left(n - \sum_{k=1}^{i-1} \delta_k\right) = mn - \sum_{i=1}^{m} (m-i)\delta_i. \tag{3.7}$$

We next analyze the significance of the Hermite matrix (input part) for state space systems and derive the corresponding canonical form of the input pair.

Let (A, B) be the input pair of a *state space* system $\Sigma \in \mathcal{R}(\delta)$. Denote by B_j the matrix $[b_{j+1}, \cdots, b_m]$ of the last $m-j$ column vectors of B. Then the A-invariant subspace $\langle A|\operatorname{Im} B_j\rangle$ generated by $b_{j+1}, \cdots, b_m$ is the set of all states which can be reached from 0 by controlling the system only through the input channels $j+1, \cdots, m$. Put $\langle A|\operatorname{Im} B_m\rangle := \{0\}$. Denote by $\bar{A}_j$ the linear map $X/\langle A|\operatorname{Im} B_j\rangle \to X/\langle A|\operatorname{Im} B_j\rangle$ induced by A and write $[x]_j := x + \langle A|\operatorname{Im} B_j\rangle$ for the elements of these quotient spaces ($x \in X; j = 1, \cdots, m$). Then the Hermite matrix $H_1(\Sigma)$ can be characterized in terms of the input pair (A, B) as follows.

PROPOSITION 3.8. *Let $\delta = (\delta_1, \cdots, \delta_m)$ be a list of nonnegative integers with $\sum_{i=1}^{m} \delta_i = n$ and $\Sigma \in \mathcal{R}(\delta)$ a state space system with input pair (A, B) and Hermite matrix $H(\Sigma) = (h_{ij})$. Then*

(i) *h_{jj} is the minimal monic annihilating polynomial of $[b_j]_j \in X/\langle A|\operatorname{Im} B_j\rangle$ with respect to $\bar{A}_j$, $j = 1, \cdots, m$.*

(ii) *The vectors*

$$b_1, Ab_1, \cdots, A^{\delta_1 - 1}b_1; \quad \cdots; \quad b_m, Ab_m, \cdots, A^{\delta_m - 1}b_m \tag{3.8}$$

form a basis of the state space X.

(iii) *For $j = 1, \cdots, m$, $i > j$, the coefficients a_k^{ij} of the polynomials*

$$h_{ij}(s) = a_0^{ij} + a_1^{ij}s + \cdots + a_{\delta_i - 1}^{ij} s^{\delta_i - 1}$$

are uniquely determined as the coordinates of $h_{jj}(A)b_j$ corresponding to the basis vectors $b_i, Ab_i, \cdots, A^{\delta_i - 1}b_i$:

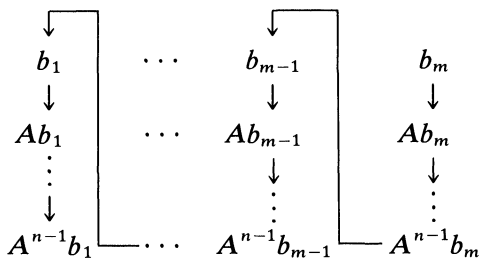
$$h_{jj}(A)b_j = -\sum_{i=j+1}^{m} \sum_{k=0}^{\delta_i - 1} a_k^{ij} A^k b_i. \tag{3.9}$$

Proof. Since the column vectors $h_j = (0, \cdots, 0, h_{jj}, \cdots, h_{mj})^T$ of $H_1(\Sigma)$ belong to $\mu(sI_n - A, B)$ we have

$$\sum_{i=j}^{m} h_{ij}(A)b_i = 0. \tag{3.10}$$

Hence $h_{jj}(A)$ annihilates b_j modulo $\langle A|\operatorname{Im} B_j\rangle$. This implies that in the sequence

(3.11)



no more than the first δ_j vectors in each column $j = m, m-1, \cdots, 1$ are linearly independent of the preceding vectors. However, since $\langle A|\operatorname{Im} B\rangle = X$ and $\sum_{j=1}^{m} \delta_j = n$,

all these n vectors have to be linearly independent. This proves (ii) and also that h_{ij} is a *minimal* annihilating polynomial of $b_j \bmod \langle A|\operatorname{Im} B_j\rangle$ with respect to A for $j=1,\cdots,m$, i.e., (i). Finally (iii) follows from (3.10) and (ii). □

As an immediate consequence, we obtain the following realization of a given system matrix $\Sigma(s)$ in Hermite canonical state space form.

COROLLARY 3.9. (Hermite canonical form of reachable state space systems). *Every system matrix* $\Sigma(s)\in\mathscr{R}(\delta)$ *is* s.s.e. *to a uniquely determined system* $\hat{\Sigma}$ *in state space form with input pair* $(A,B)\in\mathbb{K}^{n\times(n+m)}$ *of the following structure*:

$$A=(A_{ij})_{i,j=1,\cdots,m},\qquad B=(b_j)_{j=1,\cdots,m} \tag{3.12}$$

where

(a) *for* $j=1,\cdots,m$ *with* $\delta_j>0$:

$$A_{jj}=\begin{bmatrix}0 & \cdots & & 0 & -a_0^{jj}\\ 1 & & & & \\ 0 & & & \vdots & \vdots \\ \vdots & & & 0 & \\ 0 & \cdots & 0 & 1 & -a_{\delta_j-1}^{jj}\end{bmatrix}\in\mathbb{K}^{\delta_j\times\delta_j},$$

$$A_{ij}=\begin{bmatrix}0 & & 0 & -a_0^{ij}\\ \vdots & \cdots & \vdots & \vdots \\ 0 & & 0 & -a_{\delta_i-1}^{ij}\end{bmatrix}\in\mathbb{K}^{\delta_i\times\delta_j}\quad \textit{for } i>j,$$

$$A_{ij}=0_{\delta_i\times\delta_j}\quad \textit{for } i<j,$$

$$b_j\in\mathbb{K}^m,\qquad b_{lj}=\begin{cases}1 & \textit{if } l=\sum_{k=1}^{j-1}\delta_k+1,\\ 0 & \textit{else};\end{cases}$$

(b) *for* $j=1,\cdots,m$ *with* $\delta_j=0$:

$$A_{ij}\ \textit{void},\quad b_j\in\mathbb{K}^n,\quad b_{lj}=\begin{cases}0 & \textit{if } l\leqq\sum_{k=1}^{j-1}\delta_k,\\ \textit{arbitrary} & \textit{else}.\end{cases}$$

Proof. Consider the state space model (2.4) and represent the linear operators A,B by their matrices with respect to the basis (3.8) of X. Then A and B have the required form. This proves the existence result. Conversely, let $\hat{\Sigma}$ be a state space system which is strictly system equivalent to Σ and whose input pair is of the above form. It follows from Proposition 3.8 (ii), (iii) and (3.12) that the entries $\hat{h}_{ij}$ of $H_1(\hat{\Sigma})$ are given by

$$\hat{h}_{ij}=\hat{a}_0^{ij}+\hat{a}_1^{ij}s+\cdots+\hat{a}_{\delta_i}^{ij}s^{\delta_i},\ \hat{a}_{\delta_i}^{ij}=\begin{cases}1 & \textit{if } j=i\\ 0 & \textit{if } j\neq i\end{cases}$$

where the coefficients $\hat{a}_k^{ij}$, $0\leqq k<\delta_i$ are taken from $\hat{A}_{ij}$, $i\geqq j$. Because of $H_1(\hat{\Sigma})=H_1(\Sigma)$ we conclude that $\hat{A}=A$. Furthermore $b_j=\hat{b}_j$ for $j\in\{1,\cdots,m\}$ with $\delta_j>0$ (by assumption). Since the vectors A^ib_j, $i=0,\cdots,n-1$, $j\in\{1,\cdots,m\}$, $\delta_j>0$ span X it follows from the similarity of (A,B) and $(\hat{A},\hat{B})$ that $B=\hat{B}$. □

Remark 3.10. The preceding results have been obtained starting from the lower Hermite form. In an analogous way, the upper Hermite form leads to a selection procedure which starts with b_1 instead of b_m. The corresponding state space canonical form coincides with Luenberger's first canonical form (Luenberger [1967]).

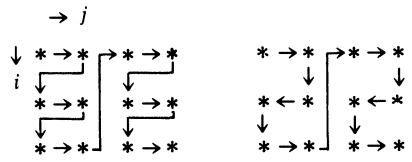
4. Dependence indices and generalized Hermite forms. Let $m,n\in\mathbb{N}$, $m,n\geqq 1$, $\underline{m}=\{1,\cdots,m\}$, $\bar{n}=\{0,1,\cdots,n\}$.

DEFINITION 4.1. A relation of total order $\preccurlyeq$ on $\bar{n} \times \underline{m}$ is called *nice* if

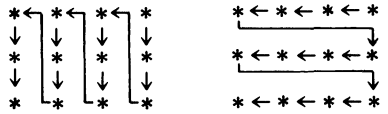
$$i \leqq k \Rightarrow (i, j) \preccurlyeq (k, j) \quad \text{for } j \in \underline{m}, \tag{4.1}$$

$$(k, l) \preccurlyeq (i, j) \Rightarrow (k+1, l) \preccurlyeq (i+1, j) \quad \text{for } k, i < n. \tag{4.2}$$

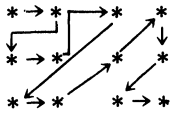
A nice order can be represented graphically by a *nice path* through an $(n+1) \times m$ array of points (Fig. 4.1).



(a) *nice* (b) *not nice*



(c) (*left*) *Hermite path* (d) (*left*) *Kronecker path*



(e) *nice*

FIG. 4.1

Remark 4.2. The properties (4.1), (4.2) imply that two nice orders on $\bar{n} \times \underline{m}$ with the same starting point $(0, j_1)$ coincide if each point $(0, k)$ has the same predecessor (i_k, j_k) with respect to both order relations ($k \in \underline{m}$, $k \neq j_1$).

Proceeding conversely to § 2 we associate with every nice order $\preccurlyeq$ a modified Rosenbrock *deleting procedure* for reachable pairs $(A, B) \in \mathbb{K}^{n \times (n+m)}$ in state space form and derive from it a uniquely determined basis of the module of return to zero $\mu(sI_n - A, B)$.

(4.3) Delete in the family $(A^i b_j)_{(i,j) \in \bar{n} \times \underline{m}}$ ordered by $\preccurlyeq$ every vector $A^i b_j$ which is linearly dependent upon its predecessors.

By (4.1) and the theorem of Cayley–Hamilton all vectors $A^n b_j$, $j \in \underline{m}$ are deleted. By (4.2) the deletion of $A^i b_j$ implies the deletion of $A^{i+k} b_j$ for $k \geqq 0$. Therefore the result of the deletion procedure is a *nice selection* in the sense of Hazewinkel and Kalman (1976), i.e., the vectors remaining after the deleting process form a set of the form

$$\{A^i b_j ; j \in \underline{m}, i < \delta_j\} \tag{4.4}$$

where the integers $\delta_j \geqq 0$ satisfy

$$\sum_{j=1}^{m} \delta_j = n$$

because (A, B) is reachable.

The list $\delta_{(A,B)} = (\delta_1, \cdots, \delta_m)$ is called a list of *dependence indices* of the pair (A, B) with respect to the nice order $\preccurlyeq$. The vectors $A^{\delta_i} b_j$ can be expressed as linear combinations of the $A^k b_i$ with $k < \delta_i$ and $(k, i) \prec (\delta_j, j)$. Hence there exist polynomials $h_{ij} \in \mathbb{K}[s]$ such that for $j \in \underline{m}$:

$$\sum_{i=1}^{m} h_{ij}(A) b_i = 0, \tag{4.5}$$

$$h_{jj} \text{ monic with degree } \delta_j \quad \text{for } j \in \underline{m}, \tag{4.6a}$$

$$\deg h_{ij} < \deg h_{ii} \quad \text{for all } i \in \underline{m}, \quad i \neq j, \tag{4.6b}$$

$$(\deg h_{ij}, i) \preccurlyeq (\deg h_{jj}, j) \quad \text{for } i, j \in \underline{m}, \tag{4.6c}$$

where we define: $(\deg 0, j) \prec (i, k)$ $\forall j, k \in \underline{m}$ and $\forall i \in \bar{n}$. Since the vectors (4.4) are linearly independent the polynomials h_{ij} satisfying (4.5) and (4.6) are uniquely determined. Define $H^{\prec}_{(A,B)} := (h_{ij})_{(i,j) \in \underline{m} \times \underline{m}}$.

LEMMA 4.3. *The columns of $H^{\prec}_{(A,B)}$ form a basis of the module $\mu(sI_n - A, B)$. Two pairs (A, B) and $(\tilde{A}, \tilde{B})$ of dimension $n \times (n+m)$ are similar if and only if $H^{\prec}_{(A,B)} = H^{\prec}_{(\tilde{A},\tilde{B})}$.*

Proof. Because of (4.5) and (2.6), the columns of $H^{\prec}_{(A,B)}$ belong to $\mu(sI_n - A, B)$ and because of (4.6) we have $\deg \det H_{(A,B)} = n$. Thus $H^{\prec}_{(A,B)}$ is a basis matrix for $\mu(sI_n - A, B)$. $H^{\prec}_{(A,B)} = H^{\prec}_{(\tilde{A},\tilde{B})}$ implies $\mu(sI_n - A, B) = \mu(sI_n - \tilde{A}, \tilde{B})$, i.e., similarity of the two reachable pairs. Conversely, if $\tilde{A} = TAT^{-1}$, $\tilde{B} = TB$ with T nonsingular, then $\tilde{A}^i \tilde{b}_j = TA^i b_j$ for all $i \in \bar{n}$, $j \in \underline{m}$. Applying T to (4.5) we conclude that $H^{\prec}_{(A,B)} = H^{\prec}_{(\tilde{A},\tilde{B})}$. □

If $\preccurlyeq$ is the Hermite order (c), then $H^{\prec}_{(A,B)}$ is just of Hermite form. This justifies the following

DEFINITION 4.4. Let $\preccurlyeq$ be a nice order on $\bar{n} \times \underline{m}$ and $\delta = (\delta_1, \cdots, \delta_m) \in \mathbb{N}^m$, $\sum_{i=1}^{m} \delta_i = n$. A matrix $H \in \mathbb{K}[s]^{m \times m}$ is called a *generalized Hermite matrix* of index list δ with respect to $\preccurlyeq$ ($H \in \mathscr{H}^{\prec}(\delta)$) if (4.6) holds.

The following result extends Hermite's theorem to generalized Hermite matrices. Although its content is purely algebraic, we use system theoretic tools for its proof.

THEOREM 4.5. *Let $\preccurlyeq$ be a nice order on $\bar{n} \times \underline{m}$. For every nonsingular matrix $D(s) \in \mathbb{K}[s]^{m \times m}$ with $\deg \det D(s) = n$ there exist a unique index list $\delta = (\delta_1, \cdots, \delta_m) \in \mathbb{N}^m$ with $\sum_{i=1}^{m} \delta_i = n$ and a unique generalized Hermite matrix $H(s) \in \mathscr{H}^{\prec}(\delta)$ such that $D(s)$ is right-equivalent to $H(s)$.*

Proof. Let $U(s)$ be a unimodular matrix such that $\tilde{D}(s)^{-1} := U(s)^{-1} D(s)^{-1}$ is strictly proper and let (A, B, C) be a minimal realization of $G(s) := \tilde{D}(s)^{-1}$. Then the system matrices

$$\begin{bmatrix} sI_n - A & -B \\ C & 0 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} D(s) & -I_m \\ I_m & 0 \end{bmatrix}$$

are s.s.e., hence $\mu(sI_n - A, B) = \mu(\tilde{D}(s), I_m) = \tilde{D}(s)\mathbb{K}[s]^m = D(s)\mathbb{K}[s]^m$ by Proposition 2.2. Define $H := H^{\prec}_{(A,B)}$. Since the columns of $H(s)$ form a basis of $\mu(sI_n - A, B)$ $H(s)$ is right-equivalent to $D(s)$. Since $\det H(s) = \text{const} \cdot \det D(s)$ and $\det D(s) \neq 0$, we have $\delta_i = \deg h_{ii} \geqq 0$ and $\sum_{i=1}^{m} \delta_i = n$. Because of Lemma 4.3 $H(s)$ and δ are uniquely determined by (A, B), hence by $D(s)$. □

As an application of this theorem we derive new complete lists of invariants for system matrices with respect to s.s.e. Let $\preccurlyeq$ be a given nice order on $\bar{n} \times \underline{m}$. For every reachable system matrix

$$\Sigma(s) = \begin{bmatrix} P(s) & -Q(s) \\ V(s) & W(s) \end{bmatrix}$$

of order n there exist a unique list $\delta = (\delta_1, \cdots, \delta_m) \in \mathbb{N}^m$ with $\sum_{i=1}^m \delta_i = n$ and a unique basis matrix $H_{P,Q} \in \mathscr{H}^{\prec}(\delta)$ of $\mu(P, Q)$ by Theorem 4.5. For a fixed list δ denote by $\mathscr{R}^{\prec}(\delta)$ the set of all polynomial input pairs $(P, Q) \in \mathbb{K}[s]^{r\times(r+m)}$, $r \geqq 1$, with $H_{P,Q} \in \mathscr{H}^{\prec}(\delta)$.

PROPOSITION 4.6. (i) *Two reachable system matrices $\Sigma, \tilde{\Sigma}$ of order n with transfer functions $G(s)$, $\tilde{G}(s)$ respectively and with input space $\mathbb{K}^m$ are* s.s.e. *if and only if*

$$H^{\prec}_{P,Q} = H^{\prec}_{\tilde{P},\tilde{Q}} \quad \text{and} \quad G(s)H^{\prec}_{P,Q} = \tilde{G}(s)H^{\prec}_{\tilde{P},\tilde{Q}}.$$

(ii) $$\gamma: \mathscr{R}^{\prec}(\delta) \to \mathscr{H}^{\prec}(\delta), \qquad (P, Q) \to H^{\prec}_{P,Q}$$

is surjective, hence induces a bijection

$$\mathscr{R}^{\prec}(\delta)/_{\text{sse}} \leftrightarrow \mathscr{H}^{\prec}(\delta).$$

Proof. (i) follows from Lemma 4.3 and Proposition 2.2.

(ii) Let $H(s) \in \mathscr{H}^{\prec}(\delta)$ and define $(P, Q) = (H, I_m)$. Then $\mu(P, Q) = H\mathbb{K}[s]^m$ and $\gamma(P, Q)) = H$. □

Proceeding as in § 3 (cf. Corollary 3.4, Corollary 3.9), canonical forms for system matrices and for state space systems can be derived in a straightforward way from this proposition. Thus we can associate with each nice order $\preccurlyeq$ on $\bar{n} \times \underline{m}$ a different *generalized Hermite canonical form.*

Example 4.7. Let

$$D(s) = \begin{bmatrix} s & s+1 & s & s^2 \\ s^3+s^2-s & s^3+2s^2-s & s^2 & s^3+s^2-s \\ 0 & 0 & 2s & 1 \\ -s^2-2s & -s^2-3s-1 & 7s^2-s+3 & -s^2+s \end{bmatrix} \in \mathbb{R}[s]^{4\times 4}.$$

We obtain as associated generalized Hermite matrices:

a) the index list (0, 2, 0, 2) and the Hermite matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ s & s^2-s & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -1 & -s & 2s & s^2+1 \end{bmatrix}$$

for the (left) Hermite order on $\bar{4} \times \underline{4}$;

b) the index list (1, 1, 1, 1) and the Kronecker-matrix

$$\begin{bmatrix} s-1 & 1 & 0 & 0 \\ 0 & s & 0 & 0 \\ 0 & 0 & s & \frac{1}{2} \\ 1 & -1 & -2 & s \end{bmatrix}$$

for the (left) Kronecker order on $\bar{4} \times \underline{4}$;

c) the index list (3, 1, 0, 0) and the $\preccurlyeq$-matrix

$$\begin{bmatrix} s^3-s^2-s-1 & s & -2s^2+2s & s-1 \\ 0 & s & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

for the nice order on $\bar{4} \times \underline{4}$ corresponding to Fig. 4.1(a).

5. Nice minimal bases. A basis $d_1, \cdots, d_k$ of a submodule $M \subset \mathbb{K}[s]^m$ is called *minimal* (of minimal degree) if

$$\sum_{i=1}^{k} \deg d_i \leqq \sum_{i=1}^{k} \deg d_i' \quad \text{for all bases } (d_1', \cdots, d_k') \text{ of } M.$$

It is well known that the set of degrees $\nu_i = \deg d_i$, $i \in \underline{k}$ is independent of the particular minimal basis of M and that a basis $d_1, \cdots, d_k$ of M is minimal if and only if the associated *basis matrix* $D = (d_1, \cdots, d_k)$ is column proper.

Minimal bases of submodules in $\mathbb{K}[s]^m$ play an important role in the analysis of those structural properties of linear systems which are based on the degree structure of $\mu(\Sigma)$ (resp. $\mu(P, Q)$) (cf. Münzner and Prätzel-Wolters (1979)). In particular, we have seen in § 3 that column properness of the basis matrix $H_1(\Sigma)$ is desirable if we want to determine the number of free parameters of the output part of Σ (Lemma 3.6).

For a given submodule $M \subset \mathbb{K}[s]^m$ there exist many different column proper basis matrices, in general. In the following we show how special nice orders can be used to single out uniquely determined representatives among all minimal bases of a full submodule M. These minimal bases are called *nice*.

By definition, every generalized Hermite matrix $H(s) \in \mathbb{K}[s]^{m \times m}$ is row proper with leading row coefficient matrix $[H]_l' = I_m$, but in general is not column proper. In the following we characterize those nice orders which lead to *column proper generalized Hermite matrices.*

A nice order $\preccurlyeq$ with the property

$$(i, j) \prec (i+1, k) \quad \text{for } i \in \overline{n-1}, \quad j, k \in \underline{m} \tag{5.1}$$

is called a *Kronecker order.* Notice that a Kronecker order is completely determined by the ordering of the pairs $(0, j)$, $j \in \underline{m}$. The Kronecker order (cf. Fig. 4.1(d)) with the additional property

$$(i, j) \succ (i, j+1) \quad \text{for } i \in \bar{n}, \quad j \in \underline{m-1} \tag{5.2}$$

is called a *left Kronecker order* and the associated matrices are called *left Kronecker–Hermite matrices.*

PROPOSITION 5.1. (i) *A matrix $D(s) \in \mathbb{K}[s]^{m \times m}$ is a left Kronecker–Hermite matrix if and only if*

$$[D(s)]_l' = I_m \quad \text{and} \quad [D(s)]_l^c = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ * & 1 & 0 & 0 \\ \vdots & & & \vdots \\ * & \cdots & * & 1 \end{bmatrix} \tag{5.3}$$

where $[D(s)]_l^c$ is the leading column coefficient matrix.

(ii) *For every nonsingular matrix $M(s) \in \mathbb{K}[s]^{m \times m}$ there exists a uniquely determined matrix $D(s)$ with the properties* (5.3) *which is right-equivalent to $M(s)$.*

Proof. (i) Suppose that $D(s)$ is a left Kronecker–Hermite matrix with index list δ. Then $[D(s)]_l' = I_m$ (5.1) and (5.2) together with (4.6c) imply

$$\begin{aligned} &\deg d_{ij} < \deg d_{jj} \quad \text{for } i, j \in \underline{m}, \quad i < j, \\ &\deg d_{ij} \leqq \deg d_{jj} \quad \text{for } i, j \in \underline{m}, \quad i \geqq j. \end{aligned} \tag{5.4}$$

Hence $[D(s)]_l^c$ is of lower triangular form and we obtain (5.3) because of (4.6a).

Conversely, suppose (5.3) and let $\preccurlyeq$ denote the left Kronecker order. Then $[D(s)]_l^r = I_m$ implies (4.6a) and (4.6b). If $\deg d_{ij} < \deg d_{jj}$ then

$$(\deg d_{ij}, i) \preccurlyeq (\deg d_{jj}, j) \tag{5.5}$$

by (5.1). On the other hand, if $\deg d_{ij} \geqq \deg d_{jj}$ it follows from (5.3) that $\deg d_{ij} = \deg d_{jj}$ and $i \geqq j$. Hence again (5.5) because of (5.2). This shows that condition (4.6c) is satisfied, i.e., $D(s)$ is a generalized Hermite matrix with respect to the left Kronecker order $\preccurlyeq$.

(ii) follows immediately from (i) and Theorem 4.5. □

By the preceding proposition, every left Kronecker–Hermite matrix is column proper. We will show that this property characterizes the *Kronecker–Hermite matrices*, i.e., those matrices which are generalized Hermite matrices with respect to a suitable Kronecker order on $\bar{n} \times \underline{m}$. Note that to any nice order $\preccurlyeq$ there corresponds a unique Kronecker order $\preccurlyeq_K$ such that

$$(0, j) < (0, k) \Leftrightarrow (0, j) \prec_K (0, k). \tag{5.6}$$

PROPOSITION 5.2. *Let $\preccurlyeq$ be a nice order on $\bar{n} \times \underline{m}$ and $H(s) \in \mathcal{H}^{\prec}(\delta)$ where δ is any index list. Then the following conditions are equivalent:*

(i) *$H(s)$ is column proper.*

(ii) *There exists an $m \times m$-permutation matrix Π such that $\Pi^T H(s)\Pi$ is a left Kronecker–Hermite matrix.*

(iii) *$H(s)$ is a Kronecker–Hermite matrix (with respect to $\preccurlyeq_K$).*

(iv) *δ_j is the j-th column degree of $H(s)$, $j \in \underline{m}$.*

Proof. Let us first note that

$$(0, j) > (0, k) \Rightarrow \deg h_{jk} < \deg h_{kk} \quad \text{for } j, k \in \underline{m}. \tag{5.7}$$

Indeed if $(0, j) > (0, k)$ then $\deg h_{jk} \geqq \deg h_{kk}$ would imply $(\deg h_{jk}, j) > (\deg h_{jk}, k) \geqslant (\deg h_{kk}, k)$ which contradicts (3.6c).

(i)$\Rightarrow$(ii): Suppose that $H(s)$ is column proper. Let $\pi : \underline{m} \to \underline{m}$ be the permutation $k \to j_k$ where

$$(0, j_1) > (0, j_2) > \cdots > (0, j_m)$$

and Π the corresponding permutation matrix. Then $\tilde{H}(s) = \Pi^T H(s)\Pi$ has the entries $\tilde{h}_{ij} = h_{\pi(i)\pi(j)}$ and the leading column coefficient matrix $\tilde{H}_l^c = \Pi^T H_l^c \Pi$. By (5.7)

$$j < k \Rightarrow (0, \pi(j)) > (0, \pi(k)) \Rightarrow \deg \tilde{h}_{jk} < \deg \tilde{h}_{kk} = \delta_{\pi(k)}.$$

Therefore $\tilde{H}_l^c$ is a lower triangular matrix. Since H_l^c is nonsingular by assumption, so is $\tilde{H}_l^c$; hence all the diagonal entries of $\tilde{H}_l^c$ are different from zero and $\delta_{\pi(j)}$ is the degree of the jth column of $\tilde{H}(s)$. This together with (4.6a,b) implies that $\tilde{H}_l^c$ satisfies (5.3) and therefore $\tilde{H}(s)$ is a left Kronecker–Hermite matrix by Proposition 5.1.

(ii)$\Rightarrow$(iii): Suppose (ii). Then $\delta_j = \deg h_{jj}$ is the degree of the jth column of $H(s)$. We have to show that

$$(\deg h_{jk}, j) \preccurlyeq_K (\deg h_{kk}, k) \quad \text{for } j, k \in \underline{m}.$$

If $(0, j) > (0, k)$ this follows from (5.7) and (5.1). If $(0, j) < (0, k)$ this follows from $\deg h_{jk} \leqq \deg h_{kk}$.

(iii)$\Rightarrow$(iv): Suppose that $\preccurlyeq$ is a Kronecker order. (5.1) and (4.6c) imply

$$\deg h_{jk} \leqq \deg h_{kk} \quad \text{for } j, k \in \underline{m}, \tag{5.8}$$

hence (iv).

(iv)$\Rightarrow$(i): Suppose (iv) and define Π and $\tilde{H}$ as above. Because of (iv) and (4.6a), the lower triangular matrix $\tilde{H}_l^c = \Pi^T H_l^c \Pi$ has all diagonal entries equal to 1. Hence H_l^c is nonsingular. □

As a corollary of Proposition 5.2 we obtain the following characterization of *nice minimal basis matrices*, i.e., minimal basis matrices which are in generalized Hermite form for some nice order $\preccurlyeq$.

COROLLARY 5.3. *A minimal basis matrix $H(s)$ of a full submodule M of $\mathbb{K}[s]^m$ is nice if and only if there exists an $m \times m$-permutation matrix Π such that $\Pi^T H(s)\Pi$ is a left Kronecker–Hermite matrix.*

In Proposition 5.2 we have seen that all generalized Hermite matrices corresponding to Kronecker orders are column proper. This property characterizes the Kronecker orders if $n > 1$.

PROPOSITION 5.4. *Let $n > 1$. A nice order $\preccurlyeq$ on $\bar{n} \times \underline{m}$ is a Kronecker order if and only if all generalized Hermite matrices with respect to $\preccurlyeq$ are column proper.*

We omit the easy proof and briefly return to the problem of output parameters. In § 3 it has been noted that the (classical) Hermite matrix $H(\Sigma)$ associated with a reachable system matrix $\Sigma(s)$ does not exhibit the parameters of the output part of $\Sigma(s)$ mod s.s.e. since $H(\Sigma)$ is not necessarily column proper. Now we see that this difficulty is avoided if we employ Kronecker orders instead of Hermite orders. For example, if we select the left Kronecker–Hermite basis matrix of $\mu(P, Q)$ instead of $H_1(\Sigma)$, we obtain a column proper basis matrix

$$K(\Sigma) = \begin{bmatrix} K_1(s) \\ K_2(s) \end{bmatrix}$$

of $\mu(\Sigma)$ (cf. § 3). The indices $\kappa_1, \cdots, \kappa_m$ of $K_1(\Sigma)$ coincide with the reachability indices of the pair (P, Q). Fixing (P, Q), hence K_1, every reachable system matrix $\Sigma(s)$ with input part (P, Q) and proper rational transfer function $G(s)$ is s.s.e. to a unique system matrix in polynomial left Kronecker canonical form

$$K(\Sigma) = \begin{bmatrix} K_1 & I_m \\ K_2 & 0 \end{bmatrix}$$

where the jth column degree of $K_2(s) = G(s)K_1(s)$ is bounded by κ_j, $j \in \underline{m}$ (Lemma 3.6). Counting the number of coefficients of the polynomial entries of $K_2(s)$, we obtain a complete set of $\sum_1^m (\kappa_j + 1)p = (n + m)p$ invariants for the output part.

We conclude the paper with a discussion of Forney's "echelon form". In his article about minimal bases of rational vector spaces (Forney (1975)) he exposed several basic concepts and results which proved to be very fruitful for a systematic analysis of degree structures in linear system theory. In particular, he emphasized the importance of a unique minimal basis for linear subspaces of $\mathbb{K}(s)^m$, and proposed the matrices in echelon form for this purpose[1] (cf. also Eckberg (1974), Dickinson, Morf and Kailath (1974)). Since his definition is intricate and since the details will become important in the following, we recall Forney's original definition (in "column form").

DEFINITION 5.5 (Forney). (i) Let $D(s) \in \mathbb{K}[s]^{m \times m}$ be column proper with ordered column indices $\nu_1 \leqq \cdots \leqq \nu_m$. The ith *pivot index* γ_i of $D(s)$ is the least integer such that the matrix D_i formed by the intersection of rows $\gamma_1, \cdots, \gamma_i$ with the columns of $D(s)$ of index $\leqq \nu_i$ has leading column coefficient matrix $[D_i]_l^c$ of rank i.

[1] A different type of uniquely determined column proper basis matrices $D(s) \in \mathbb{K}[s]^{m \times m}$ had already been proposed by V. M. Popov (1970).

(ii) A column proper matrix $D(s) \in \mathbb{K}[s]^{m\times m}$ is said to be in *echelon form* if

$$\nu_1 \leqq \cdots \leqq \nu_m, \tag{5.9}$$

$$d_{\gamma_i i} \text{ is monic of degree } \nu_i, \tag{5.10}$$

$$\deg d_{\gamma_i j} < \nu_i \quad \text{for any } i \text{ and } j,\ i \neq j \quad \text{s.t. } \nu_i \leqq \nu_j. \tag{5.11}$$

This definition requires a slight modification. Indeed, the following example shows that there are submodules $M \subset \mathbb{K}[s]^m$ of full rank m for which there does *not* exist a basis matrix in echelon form, according to the above formulation.

EXAMPLE 5.6. Let

$$D(s) = \begin{bmatrix} s & 1 & s^4 \\ 1 & s^3 & s \\ s^3 & s & 1 \end{bmatrix}.$$

The pivot indices of $D(s)$ are $\gamma_1 = 1$, $\gamma_2 = 2$ and $\gamma_3 = 3$. If $E(s) = D(s)U(s)$, $U(s)$ unimodular, is any other column proper basis matrix of the module $M = D(s)\mathbb{K}[s]^3$ with ordered column degrees $\nu_1 \leqq \nu_2 \leqq \nu_3$, then necessarily $\nu_1 = 3$, $\nu_2 = 3$, $\nu_3 = 4$. Therefore the first column E_1 of $E(s)$ is a linear combination of $(s, 1, s^3)^T$ and $(1, s^3, s)^T$ with coefficients $\alpha_1, \alpha_2 \in \mathbb{K}$, $\alpha_1 \neq 0$ or $\alpha_2 \neq 0$. It follows that the first pivot index of $E(s)$ is $\gamma_1 = 1$ and $\deg e_{\gamma_1 1} \leqq 1 < 3 = \nu_1$, which contradicts (5.10), hence $E(s)$ is not in echelon form.

The existence of basis matrices in echelon form can be maintained for every submodule $M \subset \mathbb{K}[s]^m$ if the definition of pivot indices in Definition 5.5 is modified as follows:

DEFINITION 5.5(i'). The ith pivot index γ_i is defined to be the least integer such that the coefficient matrix D_i formed by the intersection of rows $\gamma_1, \cdots, \gamma_i$ of $[D(s)]_l^c$ with the columns of $[D(s)]_l^c$ corresponding to indices $\leqq \nu_i$ has full rank i.

It appears that Forney assumed this version to be equivalent to Definition 5.5(i). However, if we apply the modified definition to Example 5.6, we obtain the new pivot indices $\gamma_1 = 2$, $\gamma_2 = 3$ and $\gamma_3 = 1$, and it is easily verified that the (modified) echelon form is obtained by permuting the first two columns of $D(s)$.

From now on we suppose pivot indices to be defined by the revised version, Definition 5.5(i').

PROPOSITION 5.7. *A matrix $D(s) \in \mathbb{K}[s]^{m\times m}$ with ordered column degrees $\nu_1 \leqq \cdots \leqq \nu_m$ is of echelon form if and only if*

$$[D(s)]_l^r = P \quad \text{and} \quad [D(s)]_l^c = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ * & 1 & 0 & \cdots & \vdots \\ \vdots & & & & 0 \\ * & \cdots & & * & 1 \end{bmatrix} P \tag{5.12}$$

where P is a permutation matrix of size $m \times m$ with entries 1 at (β_j, j) such that for $k, j \in \underline{m}$

$$k < j \quad \text{and} \quad \nu_k = \nu_j \Rightarrow \beta_k < \beta_j. \tag{5.13}$$

Proof. Suppose that $D(s)$ is in echelon form and has pivot indices $\gamma_1, \cdots, \gamma_m$. We will show that $D(s)$ satisfies (5.12) and (5.13) with $\beta_j = \gamma_j$. Consider the γ_jth row, $j \in \underline{m}$. $d_{\gamma_j j}$ is monic of degree ν_j. (5.9) and (5.11) imply $\deg d_{\gamma_j k} < \nu_j$ for $k = j+1, \cdots, m$. If $k < j$, $\nu_k < \nu_j$, then $\deg d_{\gamma_j k} < \nu_j$ is a consequence of $\deg d_{\gamma_j k} \leqq \nu_k$. If $k < j$, $\nu_k = \nu_j$, then again (5.11) implies $\deg d_{\gamma_j k} < \nu_j$. This shows that $P := [D(s)]_l^r$ has entries $P_{\gamma_j j} = 1$, $j \in \underline{m}$, and $P_{ij} = 0$ elsewhere. Since $\{\gamma_1, \cdots, \gamma_m\} = \underline{m}$, P is a permutation matrix. Now

consider the jth column of $D(s)$, $j \in \underline{m}$. $d_{\gamma_i j}$ is monic of degree ν_j. By (5.11)

$$\deg d_{\gamma_i j} < \nu_i \leqq \nu_j \quad \text{for } i < j. \tag{5.14}$$

Let D_j be the submatrix of $[D(s)]_l^c$ consisting of all the columns k corresponding to indices $\nu_k \leqq \nu_j$. By (5.14) the γ_kth row has zero entries in the jth column for $k < j$. Suppose $\deg d_{ij} \geqq \nu_j$. Then the rows $\gamma_1, \cdots, \gamma_{j-1}, i$ of D_j are linearly independent. By definition of γ_j we conclude that $i \geqq \gamma_j$. Hence

$$\deg d_{ij} < \nu_j \quad \text{for } i = 1, \cdots, \gamma_j - 1,$$

i.e., $[D(s)]_l^c P^{-1}$ has lower triangular form.

Finally, if $\nu_k = \nu_j$ and $k < j$, then $D_k = D_j$ and γ_k is the first row of this matrix which is linearly independent of the rows $\gamma_1, \cdots, \gamma_{k-1}$, while γ_j is the first row of D_j which is linearly independent of the rows $\gamma_1, \cdots, \gamma_k, \cdots, \gamma_{j-1}$ (cf. 5.5(i')). This proves (5.13) for $\beta_j = \gamma_j$, $j \in \underline{m}$.

Conversely, assume (5.12) together with (5.13). Then $D(s)$ is column proper and satisfies (5.9) by assumption. Furthermore, the second equality of (5.12) implies that $d_{\beta_j j}$ is monic of degree ν_j for $j \in \underline{m}$, and it follows from the first equality that

$$\deg d_{\beta_j k} < \nu_j \quad \text{for all } k \in \underline{m}, \quad k \neq j. \tag{5.15}$$

It only remains to prove (by induction) that β_k is the kth pivot index of $D(s)$ for $k \in \underline{m}$. Suppose this has been shown for $k = 1, \cdots, j-1$, $j \in \underline{m}$. Define D_j as above and let j_1, respectively j_2, be the first, respectively last, integer such that $\nu_{j_1} = \nu_j = \nu_{j_2}$. By assumption of induction, the rows $\beta_1, \cdots, \beta_{j-1}$ are linearly independent. Applying (5.15) we conclude that all rows $\beta_1, \cdots, \beta_{j-1}$ of D_j have zeros in the columns j, $j+1, \cdots, j_2$. It follows that the β_jth row of D_j is linearly independent of the rows $\beta_1, \cdots, \beta_{j-1}$ and that every row i of D_j which is linearly independent of the rows $\beta_1, \cdots, \beta_{j-1}$ necessarily has some nonzero entries in the columns $j, \cdots, j_2$. This implies by (5.12) that $i \geqq \beta_r$ for some $r \in \{j, \cdots, j_2\}$, hence $i \geqq \beta_j$ because of (5.13). Altogether, we see that β_j is the first row of D_j which is linearly independent of the rows $\beta_1, \cdots, \beta_{j-1}$, i.e., β_j is the jth pivot index of $D(s)$. This completes the proof by induction. □

Proposition 5.7 yields a characterization of the echelon form in terms of leading coefficient matrices. This criterion is easier to verify and to handle than Forney's original definition based on the concept of pivot indices. Nevertheless, comparing (5.12) and (5.13) with (5.3), we see that the analogous characterization of left Kronecker–Hermite matrices is definitely simpler. From this point of view the left Kronecker–Hermite form seems to be preferable to the echelon form.

Corollary 5.8, below specifies the relationship between both forms. Corollary 5.9 is the module theoretic analogue of Forney's existence and uniqueness result (cf. Forney [1975, p. 500]) for echelon bases of rational vector spaces.

COROLLARY 5.8. (i) *If $D(s) \in \mathbb{K}[s]^{m \times m}$ is in echelon form, then there exist a unique left Kronecker–Hermite matrix $H(s)$ and a unique permutation matrix P of size $m \times m$ satisfying* (5.13) *such that $D = HP$.*

(ii) *If $H(s) \in \mathbb{K}[s]^{m \times m}$ is a left Kronecker–Hermite matrix, then there exists a unique permutation matrix P such that HP is in echelon form.*

COROLLARY 5.9. *Every full submodule M of $\mathbb{K}[s]^m$ has a unique minimal basis in echelon form.*

The concept of the Kronecker–Hermite matrix can be generalized to rectangular matrices. By applying this concept, the above corollary can be deduced for arbitrary submodules $M \subset \mathbb{K}[s]^m$ (Hinrichsen, Münzner and Prätzel-Wolters, (1981, Thm. 3.1)).

REFERENCES

P. BRUNOVSKY, (1966) *On stabilization of linear systems under a certain class of persistent perturbations*, J. Differential Equations, 2, pp. 401–405.

M. J. DENHAM, (1974), *Canonical forms for the identification of multivariable linear systems*, IEEE Trans. Automat. Control, AC-19, pp. 646–656.

B. W. DICKINSON, M. MORF AND T. KAILATH, (1974), *A minimal realization algorithm for matrix sequences*, IEEE Trans. Automat. Control, AC-19, pp. 31–38.

A. E. ECKBERG, JR. (1974), *A characterization of linear systems via polynomial matrices and module theory*, Report ESL-R-528, Electronic Systems Laboratory, Electrical Engineering Department, Massachusetts Institute of Technology, Cambridge, MA, 1974.

G. D. FORNEY, (1975), *Minimal bases of rational vector spaces with applications to multivariable linear systems*, this Journal, 13, pp. 493–520.

P. A. FUHRMANN, (1977), *On strict system equivalence and similarity*, Internat. J. Control, 25, pp. 5–10.

K. GLOVER AND J. C. WILLEMS, (1973), *Parametrization of linear dynamical systems: canonical forms and idenifiability*, IEEE Trans. Automat. Control, AC-19, pp. 640–645.

M. L. J. HAUTUS AND M. HEYMANN, (1978), *Feedback—an algebraic approach*, this Journal, 16, pp. 83–105.

M. HAZEWINKEL AND R. E. KALMAN, (1976), *Invariants, canonical forms and moduli for linear, constant, finite dimensional, dynamical systems*, in Proc. Symposium on Algebraic System Theory, Udine, 1975, Marchesini, G., Mitter, S. K., eds., Lecture Notes on Economics and Math. Systems 131, Springer-Verlag, New York, pp. 48–60.

M. HAZEWINKEL, (1977), *Moduli and canonical forms for linear dynamical systems* II: *The topological case*, J. Math. Systems Theory, 10, pp. 363–385.

D. HINRICHSEN, H. F. MÜNZNER AND D. PRÄTZEL-WOLTERS, (1981), *Parametrization of* (C, A)*-invariant subspaces*, Systems and Control Lett. 3, pp. 192–199.

D. G. LUENBERGER, (1967), *Canonical forms for linear multivariable systems*, IEEE Trans. Automat. Control, AC-12, pp. 290–293.

D. Q. MAYNE, (1972), *A canonical model for identification of multivariable linear systems*, IEEE Trans. Automat. Control, AC-17, pp. 728–729.

H. F. MÜNZNER AND D. PRÄTZEL-WOLTERS, (1979), *Minimal bases of polynomial modules, structural indices and Brunovsky-transformations*, Internat. J. Control, 30, pp. 291–318.

M. NEWMAN, (1972), *Integral Matrices*, Academic Press, New York and London.

V. M. POPOV, (1970), *Some properties of the control systems with irreducible matrix-transfer functions*, in Seminar on Differential Equations and Dynamical Systems II, Springer-Verlag, Berlin, Heidelberg, New York.

———, (1972), *Invariant description of linear time-invariant controllable systems*, this Journal, 10, pp. 252–264.

D. PRÄTZEL-WOLTERS, (1981), *Brunovsky equivalence of system matrices—The reachable case*, IEEE Trans. Automat. Control, AC-26, pp. 429–434.

J. RISSANEN, (1974), *Basis of invariants and canonical forms for linear dynamical systems*, Automatica, 10, pp. 175–182.

H. H. ROSENBROCK, (1970), *State Space and Multivariable theory*, Nelson-Wiley, New York.

———, (1972), *Modules and the definition of state*, Internat. J. Control, 16, pp. 433–435.

H. WEINERT AND J. ANTON, (1972), *Canonical forms for multivariable system identification*, IEEE Conf. on Decision and Control.

W. A. WOLOVICH, (1974), *Linear Multivariable Systems*, Springer-Verlag, Berlin, Heidelberg, New York.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 2, March 1983

0363-0129/83/2102-0010 $01.25/0

ON THE FUNCTION SPACE CONTROLLABILITY OF LINEAR NEUTRAL SYSTEMS*

DANIEL A. O'CONNOR† AND T. J. TARN†

Abstract. A criterion is derived for approximate controllability of linear autonomous neutral functional differential equations in the Sobolev space $W_2^{(1)}[-h, 0: \mathbb{E}^n]$. Controllability conditions are based on an abstract evolution equation representation of the system. An abstract criterion applicable to a general class of neutral systems is obtained and for the special case of matrix systems of the form

$$\frac{d}{dt}(x(t)-A_1x(t-h))=E_0x(t)+E_1x(t-h)+Bu(t)$$

this abstract criterion leads to an algebraic criterion.

Key words. linear neutral system, boundary control system, approximate controllability, spectral completeness

1. Introduction. In this paper we derive a criterion for approximate controllability of linear autonomous neutral functional differential equations in the Sobolev space $W_2^{(1)}[-h, 0: \mathbb{E}^n]$. Our approach is based on an abstract evolution equation representation of the system in the state space. We first obtain an abstract controllability criterion for a very general class of neutral systems and then obtain a more specific, easier to use, algebraic controllability criterion for the class of matrix neutral systems of the form

$$\frac{d}{dt}(x(t)-A_1x(t-h))=E_0x(t)+E_1x(t-h)+Bu(t) \tag{1.1}$$

where h is a positive constant, A_1, E_0, E_1 are $n \times n$ matrices and B is an $n \times m$ matrix.

The main result of this paper is the boundary control model of a neutral system in § 3. This model admits an abstract variation of constants formula for the state of the neutral system. In § 4 we obtain abstract controllability conditions directly from this representation of the state. For matrix systems of the form (1.1) we are able to sharpen these conditions and obtain the following result: System (1.1) is approximately controllable in $W_2^{(1)}[-h, 0: \mathbb{E}^n]$ if and only if

$$\text{a)} \qquad \operatorname{rank}_{\mathbb{C}}[A_1\lambda+E_1, B]=n, \tag{1.2}$$

$$\text{b)} \qquad \operatorname{rank}[\Delta(\lambda), B]=n, \qquad \text{for all } \lambda \in \mathbb{C}, \tag{1.3}$$

where $\Delta(\lambda)$ is the characteristic matrix of (1.1).

The techniques for converting the abstract controllability conditions to conditions (1.2)–(1.3) were developed by Manitius and Triggiani in [11] and Manitius in [14]. In these papers, the authors use an abstract representation of the state to develop conditions for approximate controllability of retarded functional differential equations in the state space $\mathbb{R}^n \times L_2[-h, 0: \mathbb{R}^n]$. Our work extends some results of [11], [14] to systems of neutral type.

The first investigations of state controllability for neutral systems in $W_2^{(1)}[-h, 0: \mathbb{R}^n]$ are detailed in the series of papers [2], [8], [18]. In this work, a criterion for exact controllability is obtained for systems of type (1.1) using techniques

* Received by the editors March 3, 1981, and in final revised form March 5, 1982. This research was supported in part by the National Science Foundation under grant nos. ECS-8017184, ENG 79-08090 and INT 7902976.

† Department of Electrical Engineering, Clarkson College of Technology, Potsdam, New York 13676.
‡ Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

of the operational calculus. Although in our work we focus on approximate controllability, our results compliment and extend some results of [2], [8], [18]. Specifically, our Corollary 5.8 extends the exact controllability condition of [8], [18] to multi-input systems. Furthermore, for system (1.1), the relationship between the three concepts—spectral controllability, approximate controllability and exact controllability—is made clear. The weakest concept of the three, spectral controllability, is known to hold if and only if (1.3) holds [16]. Condition (1.2) applied to an already spectrally controllable system yields a necessary and sufficient condition for approximate controllability. The strongest concept, exact controllability, holds if and only if condition (1.3) holds and $\operatorname{rank}[B, A_1B, \cdots, A_1^{n-1}B] = n$.

Notation. We take $\mathbb{R}$ to be the real line, $\mathbb{C}$ the complex plane and $\mathbb{E}^n$ the space of complex column n-vectors with norm $|x| = (\bar{x}^T x)^{1/2}$ and inner product $\langle x, y\rangle = \bar{x}^T y$, where x^T denotes transpose and $\bar{x}$ denotes complex conjugate.

$L^2[a, b: \mathbb{E}^n]$ is the Lebesgue space of $\mathbb{E}^n$-valued functions on $[a, b]$ with norm $\|f\|_{2,[a,b]} = (\int_a^b |f(\xi)|^2\, d\xi)^{1/2}$. $L^2_{\text{loc}}[0, \infty: \mathbb{E}^n]$ is the space of $\mathbb{E}^n$-valued functions on $[0, \infty)$ whose restrictions to finite intervals are square integrable. $W_2^{(k)}[a, b: \mathbb{E}^n]$ is the Sobolev space of $\mathbb{E}^n$-valued absolutely continuous functions on $[a, b]$ with square integrable kth derivatives on $[a, b]$. $W_2^{(k)} \triangleq W_2^{(k)}[-h, 0: \mathbb{E}^n]$, $k = 1, 2, \cdots$; $W_{2,\text{loc}}^{(k)}[0, \infty: \mathbb{E}^n]$ is the space of $\mathbb{E}^n$-valued absolutely continuous functions on $[0, \infty)$ whose kth derivatives are square integrable on all finite subintervals of $[0, \infty)$. $M_2[-h, 0: \mathbb{E}^n]$ is the product space $\mathbb{E}^n \times L^2[-h, 0: \mathbb{E}^n]$.

We will denote the derivative of a function y by $\dot{y}$ and the formal operation of differentiation by D, thus for functions in $W_2^{(1)}[-h, 0: \mathbb{E}^n]$

$$[D_\theta \phi](\theta) = \dot{\phi}(\theta), \qquad \theta \in [-h, 0]$$

and for functions f in $W_2^{(1)}[a, b: \mathbb{E}^n]$

$$[D_t f](t) = \dot{f}(t), \qquad t \in [a, b].$$

We use the inner product $(\cdot|\cdot)$ in $W_2^{(1)}$

$$(\phi|\psi) = \langle \phi(0), \psi(0)\rangle + \int_{-h}^{0} \langle \dot{\phi}(\theta), \dot{\psi}(\theta)\rangle\, d\theta$$

and we define the norm $\|\phi\| = (\phi|\phi)^{1/2}$.

We let $L(X, Y)$ represent the set of all linear operators (not necessarily bounded) mapping the Banach space X into the Banach space Y. For any operator $A \in L(X, Y)$ with domain $\mathcal{D}(A) \subseteq X$, A^* is its adjoint and $N(A)$ and $\operatorname{Im}(A)$ are its null space and image, respectively. For any subspace $M \subseteq X$, $M^\perp$ represents its orthogonal compliment.

For matrices $P[\lambda]$ whose elements are polynomials in λ, $\operatorname{rank}_{\mathbb{C}} P[\lambda]$ denotes the rank of $P[\lambda]$ as a matrix over the ring of polynomials. For matrices $R(\lambda)$ whose elements are analytic functions in λ, $R^{(k)}(\lambda *) \triangleq (d^k R/d\lambda^k)(\lambda)|_{\lambda = \lambda *}$, $k = 1, 2, \cdots$.

2. Preliminaries. Throughout most of this paper we shall consider a system of linear neutral functional differential equations of the form

$$\begin{aligned} &\frac{d}{dt}\Big(x(t) - \sum_{k=1}^{N} A_k x(t-h_k)\Big) = \sum_{k=0}^{N} E_k x(t-h_k) + \int_{-h}^{0} E(\theta) x(t+\theta)\, d\theta + Bu(t), \\ &x(t) = \phi(t), \qquad t \in [-h, 0]. \end{aligned} \tag{2.1}'$$

In the above N is a positive integer, h is a fixed finite delay, $0<h<\infty$, $h=h_N>h_{N-1}>\cdots>h_1>h_0=0$. The A_i's and E_i's are real $n\times n$ matrices, B is a real $n\times m$ matrix and $E(\cdot)$ is a real $n\times n$ matrix-valued square integrable function.

The initial data $\phi(\cdot)$ is an element in $W_2^{(1)}$ and the control term $u(\cdot)$ is an element in $L^2_{\text{loc}}[0,\infty:\mathbb{E}^m]$.

For compactness in notation we may rewrite (2.1)′ in terms of Stieltjes integrals

$$\begin{aligned}&\frac{d}{dt}\left(x(t)-\int_{-h}^{0}d\mu(\theta)x(t+\theta)\right)=\int_{-h}^{0}d\eta(\theta)x(t+\theta)+Bu(t),\\ &x(t)=\phi(t),\qquad t\in[-h,0],\end{aligned}\tag{2.1}$$

where $\mu(\cdot)$, $\eta(\cdot)$ are $n\times n$ matrix functions of bounded variation, continuous from the left, defined by

$$\mu(\theta)=-\sum_{k=1}^{N}A_k\chi_k(\theta),\qquad \eta(\theta)=-\sum_{k=0}^{N}E_k\chi_k(\theta)+\int_{-h}^{\theta}E(\xi)\,d\xi,$$

where $\chi_k(\cdot)$ is the characteristic function of the interval $(-\infty,-h_k]$.

Given any initial data ϕ, an element in $W_2^{(1)}$ and control $u(\cdot)$, an element in $L^2_{\text{loc}}[0,\infty:\mathbb{E}^m]$, a solution of system (2.1) is a locally absolutely continuous function $x(t;\phi,u)$ on $[-h,\infty)$ which satisfies (2.1) almost everywhere on $[0,\infty)$ and $x(t;\phi,u)=\phi(t)$ for $-h\leqq t\leqq 0$.

The solutions of (2.1) can also be viewed as $W_2^{(1)}$-valued functions. We define the function $x_t(\cdot;\phi,u)$ on $[-h,0]$, called the state of system (2.1) at time $t\geqq 0$ as,

$$x_t(\cdot;\phi,u)(\theta)=x(t+\theta;\phi,u),\qquad \theta\in[-h,0].$$

The existence and uniqueness of solutions of (2.1) with initial data in $W_2^{(1)}$ was studied in [7]. There it was shown that under the above conditions (2.1) has a unique solution, for all $t\geqq 0$, $x_t(\cdot;\phi,u)$ is in $W_2^{(1)}$, and there exist positive constants M and K depending only on the coefficient matrices such that for $t\geqq 0$

$$\|x_t(\cdot;\phi,u)\|\leqq M\left\{\|\phi\|+\left(\int_0^t|u(\xi)|^2\,d\xi\right)^{1/2}\right\}e^{Kt}.\tag{2.2}$$

It was further demonstrated in [7] that if we define the solution operator $T(t)$ by

$$T(t)\phi=x_t(\cdot;\phi,0),\qquad \phi\in W_2^{(1)}$$

then:

1) The family of operators $\{T(t),t\geqq 0\}$ is a strongly continuous semigroup of bounded linear operators on $W_2^{(1)}$.

2) The infinitesimal generator $\tilde{A}$ of $\{T(t),t\geqq 0\}$ is given by

$$\mathcal{D}(\tilde{A})=\left\{\psi\in W_2^{(1)}\,\middle|\,\dot{\psi}\in W_2^{(1)},\ \dot{\psi}(0)=\int_{-h}^{0}d\mu(\theta)\dot{\psi}(\theta)+\int_{-h}^{0}d\eta(\theta)\psi(\theta)\right\}$$

and

$$\tilde{A}\psi=\dot{\psi}\quad\text{for }\psi\in\mathcal{D}(\tilde{A}).$$

3) The spectrum of $\tilde{A}$, $\sigma(\tilde{A})$, is countably infinite and coincides with the roots of the characteristic equation, $\det\Delta(\lambda)=0$, where the characteristic matrix $\Delta(\lambda)$ is defined by

$$\Delta(\lambda)\equiv\lambda\left(I-\int_{-h}^{0}d\mu(\theta)\,e^{\lambda\theta}\right)-\int_{-h}^{0}d\eta(\theta)\,e^{\lambda\theta}.$$

4) For all λ not in the spectrum $\sigma(\tilde{A})$ the resolvent of $\tilde{A}$, $(\lambda I - \tilde{A})^{-1}$, is a bounded linear operator in $W_2^{(1)}$ characterized by

$$(2.3) \qquad ((\lambda I - A)^{-1}\psi)(\theta) = e^{\lambda\theta}\Delta^{-1}(\lambda)b + \int_\theta^0 e^{\lambda(\theta-\xi)}\psi(\xi)\,d\xi \quad \text{on } [-h, 0],$$

where

$$b = \psi(0) - \int_{-h}^0 d\mu(\theta)\psi(\theta) + \lambda \int_{-h}^0 d\mu(\theta) \int_\theta^0 e^{\lambda(\theta-\xi)}\psi(\xi)\,d\xi + \int_{-h}^0 d\eta(\theta) \int_\theta^0 e^{\lambda(\theta-\xi)}\psi(\xi)\,d\xi.$$

In the next section we will make use of the following lemma which characterizes some smooth solutions of (2.1).

LEMMA 2.1. *If the forcing function u is an element in $W^{(1)}_{2,\mathrm{loc}}[0, \infty: \mathbb{E}^m]$ and the initial data ϕ is an element in $W_2^{(2)}$ with*

$$\dot{\phi}(0) = \int_{-h}^0 d\mu(\xi)\dot{\phi}(\xi) + \int_{-h}^0 d\eta(\xi)\phi(\xi) + Bu(0),$$

then the following remarks hold:

a) *$x_t(\cdot\,; \phi, u)$ is an element in $W_2^{(2)}$ for each $t > 0$;*

b) *the $W_2^{(1)}$-valued function, $t \to x_t(\cdot\,; \phi, u)$, is strongly continuously differentiable on $(0, \infty)$ and for $t > 0$,*

$$(2.4) \qquad \frac{d}{dt}x_t(\cdot\,; \phi, u) = D_\theta x_t(\cdot\,; \phi, u).$$

Proof. a) For $0 \leqq t \leqq h_1$, (h_1 is the shortest delay), the hypotheses on u and ϕ imply that $x(t; \phi, u)$ is in $W_2^{(2)}[-h, h_1: \mathbb{E}^n]$. By a straightforward application of the method of steps, (e.g. [3, Thm. 5.1]), the solution can be continued indefinitely and $x(t; \phi, u)$ is in $W^{(2)}_{2,\mathrm{loc}}[-h, \infty: \mathbb{E}^n]$.

By the definition of state, it follows that for all $t > 0$, $x_t(\cdot\,; \phi, u)$ is in $W_2^{(2)}$ and furthermore

$$[D_\theta x_t(\cdot\,; \phi, u)](\theta) = \dot{x}(t + \theta; \phi, u), \qquad \theta \in [-h, 0].$$

b) Since the $\mathbb{E}^n$-valued function, $\dot{x}(t; \phi, u)$, is absolutely continuous one can easily show that the $W_2^{(1)}$-valued function, $t \to D_\theta x_t(\cdot\,; \phi, u)$ is continuous in $W_2^{(1)}$ norm.

To verify statement (2.4), we show that for all $t > 0$

$$(2.5) \qquad \lim_{\tau \to 0} \left\| \frac{x_{t+\tau}(\cdot\,; \phi, u) - x_t(\cdot\,; \phi, u)}{\tau} - D_\theta x_t(\cdot\,; \phi, u) \right\| = 0.$$

Expanding the square of the above norm we obtain

$$\left\| \frac{x_{t+\tau}(\cdot\,; \phi, u) - x_t(\cdot\,; \phi, u)}{\tau} - D_\theta x_t(\cdot\,; \phi, u) \right\|^2$$

$$= \left| \frac{x(t+\tau; \phi, u) - x(t; \phi, u)}{\tau} - \dot{x}(t; \phi, u) \right|^2 + \int_{-h}^0 \left| \frac{\dot{x}(t+\tau+\theta; \phi, u) - \dot{x}(t+\theta; \phi, u)}{\tau} - \ddot{x}(t+\theta; \phi, u) \right|^2 d\theta.$$

The first term on the right tends to zero as $\tau \to 0$ since $x(t; \phi, u)$ is continuously differentiable. Moreover, since $\dot{x}(t; \phi, u)$ is absolutely continuous with $\ddot{x}(t; \phi, u)$ in $L^2_{\text{loc}}[0, \infty: \mathbb{E}^n]$, we have

$$\begin{aligned}
&\int_{-h}^{0} \left| \frac{\dot{x}(t+\tau+\theta; \phi, u) - \dot{x}(t+\theta; \phi, u)}{\tau} - \ddot{x}(t+\theta; \phi, u) \right|^2 d\theta \\
&= \int_{-h}^{0} \left| \int_0^{\tau} \frac{\ddot{x}(t+\theta+\xi; \phi, u) - \ddot{x}(t+\theta; \phi, u)}{\tau} d\xi \right|^2 d\theta \\
&\leqq \int_{-h}^{0} \int_0^{\tau} \left| \frac{\ddot{x}(t+\theta+\xi; \phi, u) - \ddot{x}(t+\theta; \phi, u)}{\tau} \right|^2 d\xi \, d\theta \\
&= \frac{1}{\tau} \int_0^{\tau} \left[\int_{-h}^{0} |\ddot{x}(t+\theta+\xi; \phi, u) - \ddot{x}(t+\theta; \phi, u)|^2 \, d\theta \right] d\xi \\
&\to 0, \text{ as } \tau \to 0, \text{ for almost all } t.
\end{aligned}$$

In the last step, we have used the fact that the expression in brackets is an integrable function of ξ on any finite interval.

Therefore, the limit (2.5) and hence (2.4) hold for almost all $t > 0$. Since $t \to D_\theta x_t(\cdot; \phi, u)$ is continuous, (2.4) must in fact hold for all $t > 0$.

3. The abstract boundary control system and a variation of constants formula. System (2.1) can be associated with the following abstract boundary control system on $W_2^{(1)}$:

$$\begin{aligned}
&\frac{d}{dt} z(t) = \Lambda z(t), \quad t > 0, \qquad z(0) \equiv \phi, \\
&\Xi z(t) = Bu(t), \qquad t \geqq 0,
\end{aligned} \tag{3.1}$$

where $\Lambda \in L(W_2^{(1)}, W_2^{(1)})$ is a closed linear operator defined by

$$\Lambda \psi = D_\theta \psi, \quad \text{for all } \psi \in \mathcal{D}(\Lambda) = W_2^{(2)};$$

$\Xi \in L(W_2^{(1)}, \mathbb{E}^n)$, is a linear operator defined by

$$\Xi \psi = \dot{\psi}(0) - \int_{-h}^{0} d\mu(\theta) \dot{\psi}(\theta) - \int_{-h}^{0} d\eta(\theta) \psi(\theta), \quad \text{for all } \psi \in \mathcal{D}(\Xi) = W_2^{(2)}.$$

B is an $n \times m$ real matrix, the control term $u(\cdot)$ is an element in $L^2_{\text{loc}}[0, \infty: \mathbb{E}^m]$ and the initial data $\phi(\cdot)$ is an element in $W_2^{(1)}$.

System (3.1) is a special case of the general abstract boundary control system defined in [5]. Equations of this form were originally used to study controllability and optimal control problems for systems of partial differential equations with control on the boundary, (e.g., [1], [21], [20]). We will show that solutions of (2.1) are also solutions of the abstract boundary control problem. The state of system (2.1) will be represented by a variation of constants formula for solutions of (3.1).

The association of control problems for functional differential equations with problems of boundary control for partial differential equations has been known for some time, [22]. In the case of retarded functional differential equations and some parabolic partial differential equations with boundary control, the state evolution can be expressed on a product space. Specifically, it is well known, [4], that for retarded systems the semigroup $(T(t), t \geqq 0)$, can be extended to the space $M_2[-h, 0: \mathbb{E}^n]$ and

the evolution of the system state $\tilde{x}(t)$ expressed by the "distributed control" equation

$$\frac{d}{dt}\tilde{x}(t) = \hat{A}\tilde{x}(t) + \hat{B}u(t), \tag{3.2}$$

where $\hat{A}$ is the infinitesimal generator of a strongly continuous semigroup, and $\hat{B}$ is a bounded linear operator.

Since we choose to work with neutral systems in the space $W_2^{(1)}$ it is not possible to write state evolution equations in the distributed control form since there is no well-defined operator $\hat{B}$ with range in $W_2^{(1)}$ which characterizes the effect of the control action on the system. Thus we consider abstract evolution equations of boundary control type.

DEFINITION 3.1. A *strong solution* of the abstract boundary control system (3.1) in $t \geqq 0$ is a $W_2^{(1)}$-valued continuously differentiable function $z(t)$ defined in $t \geqq 0$ such that $z(t) \in W_2^{(2)}$ for $t \geqq 0$, $z(0) = \phi$ and (3.1) holds everywhere.

THEOREM 3.2. *If the assumptions of Lemma* 2.1 *are satisfied the state function* $x_t(\cdot\,; \phi, u)$ *is a strong solution of system* (3.1).

Proof. Under the assumptions of Lemma 2.1 we have shown the function $t \to x_t(\cdot\,; \phi, u)$ is strongly continuously differentiable and clearly $x_0(\cdot\,; \phi, u) = \phi$.

Moreover, for each $t > 0$, $x_t(\cdot\,; \phi, u)$ is in $W_2^{(2)}$ and

$$\frac{d}{dt}x_t(\cdot\,; \phi, u) = D_\theta x_t(\cdot\,; \phi, u) = \Lambda x_t(\cdot\,; \phi, u).$$

Therefore the first equation in (3.1) is satisfied. Next consider the boundary equation. Since $x(t; \phi, u)$ is a solution of (2.1),

$$\begin{aligned}\Xi[x_t(\cdot\,; \phi, u)] &= \dot{x}(t; \phi, u) - \int_{-h}^{0} d\mu(\theta)\dot{x}(t+\theta; \phi, u) - \int_{-h}^{0} d\eta(\theta)x(t+\theta; \phi, u) \\ &= Bu(t) \quad \text{for almost all } t > 0.\end{aligned}$$

This equation also holds at $t = 0$ since, by hypothesis,

$$\Xi(\phi) = \dot{\phi}(0) - \int_{-h}^{0} d\mu(\theta)\dot{\phi}(\theta) - \int_{-h}^{0} d\eta(\theta)\phi(\theta) = Bu(0);$$

thus for almost all $t \geqq 0$

$$\Xi[x_t(\cdot\,; \phi, u)] = Bu(t). \tag{3.3}$$

However it follows from the smoothness assumptions that both sides of (3.3) are continuous and therefore (3.3) holds for all $t \geqq 0$. All requirements of a strong solution being met, the theorem is proved.

Given that under smoothness assumptions $x_t(\cdot\,; \phi, u)$ is a solution to (3.1) we will now construct a representation of $x_t(\cdot\,; \phi, u)$ in terms of the semigroup $\{T(t), t \geqq 0\}$. To this end we make the following definition.

DEFINITION 3.3. The *auxiliary boundary operator* $\tilde{B}_\lambda$ associated with system (3.1) is a bounded linear mapping from $\mathbb{E}^m$ into $W_2^{(1)}$ defined by

$$(\tilde{B}_\lambda u)(\theta) = e^{\lambda\theta}\Delta^{-1}(\lambda)Bu, \qquad \theta \in [-h, 0],$$

where λ is in the resolvent of $\tilde{A}$, $\rho(\tilde{A}) = \mathbb{C} - \sigma(\tilde{A})$ and $\Delta(\lambda)$ is the characteristic matrix of (2.1).

LEMMA 3.4. *The range of* $\tilde{B}_\lambda$ *lies in* $W_2^{(2)}$ *and for all* $\nu \in \mathbb{E}^m$

$$\Xi(\tilde{B}_\lambda \nu) = B\nu.$$

Proof. It is obvious that $\tilde{B}_\lambda \nu$ is in $W_2^{(2)}$ for all $\nu \in \mathbb{E}^m$. That $\Xi(\tilde{B}_\lambda \nu) = B\nu$ can be shown by straightforward calculation.

THEOREM 3.5. *Let the control* $u(\cdot)$ *be twice continuously differentiable, the initial data* $\phi(\cdot)$ *be an element in* $W_2^{(2)}$ *and*

$$\dot{\phi}(0) = \int_{-h}^{0} d\mu(\xi)\dot{\phi}(\xi) + \int_{-h}^{0} d\eta(\xi)\phi(\xi) + Bu(0).$$

Then the state $x_t(\cdot\,; \phi, u)$ *is given by the following variation of constants formula*:

$$x_t(\cdot\,; \phi, u) = T(t)\phi + (\lambda I - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\, d\xi. \tag{3.4}$$

Proof. First we shall construct a solution for the abstract boundary control system (3.1). Assume that the solution $z(t)$ can be expressed by

$$z(t) = v(t) + \tilde{B}_\lambda u(t), \qquad t \geqq 0$$

where $u(\cdot)$ is the control term and $v(\cdot)$ is some $W_2^{(1)}$-valued continuously differentiable function with $v(t) \in W_2^{(2)}$ for each t.

Then the following equations must be satisfied:

$$\begin{aligned} &\frac{d}{dt}[v(t) + \tilde{B}_\lambda u(t)] = \Lambda[v(t) + \tilde{B}_\lambda u(t)], \\ &\Xi[v(t) + \tilde{B}_\lambda u(t)] = Bu(t), \\ &v(0) + \tilde{B}_\lambda u(0) = \phi. \end{aligned} \tag{3.5}$$

It follows from Lemma 3.4 that equations (3.5) are equivalent to the following system of equations for $v(\cdot)$:

$$\begin{aligned} &\frac{d}{dt}v(t) = \Lambda v(t) + \left(\Lambda\tilde{B}_\lambda u(t) - \tilde{B}_\lambda \frac{d}{dt}u(t)\right), \\ &\Xi v(t) = 0, \\ &v(0) = \phi - \tilde{B}_\lambda u(0). \end{aligned} \tag{3.6}$$

Since the infinitesimal generator $\tilde{A}$ of the semigroup $(T(t), t \geqq 0)$ is characterized by

$$\tilde{A}\psi = D_\theta \psi \quad \text{for } \psi \in W_2^{(2)} \text{ such that } \Xi\psi = 0,$$

(3.6) is equivalent to

$$\begin{aligned} &\frac{d}{dt}v(t) = \tilde{A}v(t) + \left(\Lambda\tilde{B}_\lambda u(t) - \tilde{B}_\lambda \frac{d}{dt}u(t)\right), \\ &v(0) = \phi - \tilde{B}_\lambda u(0). \end{aligned} \tag{3.7}$$

By Lemma 3.3 the range of $\tilde{B}_\lambda$ is contained in the domain of Λ, hence by the closed graph theorem $\Lambda\tilde{B}_\lambda$ is a bounded linear operator, in fact for all $\nu \in \mathbb{E}^m$

$$\Lambda\tilde{B}_\lambda \nu = \lambda\tilde{B}_\lambda \nu.$$

By our smoothness hypothesis on $u(\cdot)$ both $\Lambda\tilde{B}_\lambda u(\cdot)$ and $\tilde{B}_\lambda\, du(\cdot)/dt$ are continuously differentiable.

System (3.7) is a well-posed ordinary differential equation in $W_2^{(1)}$ with continuously differentiable forcing term and, moreover, the initial data $v(0)$ is an element in $\mathscr{D}(\tilde{A})$, since by hypothesis

$$\phi - \tilde{B}_\lambda u(0) \in W_2^{(2)} \quad \text{and} \quad \Xi[\phi - \tilde{B}_\lambda u(0)] = \Xi\phi - Bu(0) = 0.$$

Under these assumptions it is well known [17], that (3.7) admits a unique continuously differentiable solution which is expressed by the following variation of constants formula:

$$v(t) = T(t)[\phi - \tilde{B}_\lambda u(0)] + \int_0^t T(t-\xi)\left(\lambda \tilde{B}_\lambda u(\xi) - \tilde{B}_\lambda \frac{d}{d\xi} u(\xi)\right) d\xi.$$

Therefore a solution of system (3.1) is given by

$$z(t) = T(t)[\phi - \tilde{B}_\lambda u(0)] + \int_0^t T(t-\xi)\left(\lambda \tilde{B}_\lambda u(\xi) - \tilde{B}_\lambda \frac{d}{d\xi} u(\xi)\right) d\xi + \tilde{B}_\lambda u(t). \tag{3.8}$$

The derivative term in the integrand of (3.8) may be eliminated by the abstract integration by parts formula [9]. Therefore under the hypothesis that u is twice continuously differentiable (3.8) may be rewritten,

$$z(t) = T(t)\phi + (\lambda I - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\, d\xi.$$

Under the stated assumptions on u and ϕ, an application of Theorem 3.2 shows that the state $x_t(\cdot\,; \phi, u)$ of system (2.1) satisfies the abstract boundary control problem and therefore can be represented by (3.4).

The representation of $x_t(\cdot\,; \phi, u)$ has so far been valid only under smoothness assumptions on $u(\cdot)$ and $\phi(\cdot)$. However, the variation of constants formula (3.4) is well defined for all ϕ in $W_2^{(1)}$ and all u in $L^2_{\text{loc}}[0, \infty: \mathbb{E}^m]$ if we interpret the integral on the right as a Bochner integral. In the following theorem we prove that (3.4) is indeed a valid representation of $x_t(\cdot\,; \phi, u)$ for all ϕ in $W_2^{(1)}$ and u in $L^2_{\text{loc}}[0, \infty: \mathbb{E}^m]$.

THEOREM 3.6. *For all $u \in L^2_{\text{loc}}[0, \infty: \mathbb{E}^m]$ and $\phi \in W_2^{(1)}$*

$$x_t(\cdot\,; \phi, u) = T(t)\phi + (\lambda I - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\, d\xi, \qquad t \geqq 0.$$

Proof. By linearity, it is sufficient to show that for all $u \in L^2_{\text{loc}}[0, \infty: \mathbb{E}^m]$

$$x_t(\cdot\,; 0, u) = (\lambda I - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\, d\xi, \qquad t \geqq 0.$$

For any fixed $t \geqq 0$, with u an element in $L^2_{\text{loc}}[0, \infty: \mathbb{E}^m]$, we can find a sequence $\{y_k\}_{k \geqq 0}$ of twice continuously differentiable $\mathbb{E}^m$-valued functions with compact support on $[0, t]$ such that as $k \to \infty$,

$$\|y_k - u\|_{2,[0,t]} \to 0.$$

By Theorem 3.5, for each $k \geqq 0$

$$x_t(\cdot\,; 0, y_k) = (\lambda I - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda y_k(\xi)\, d\xi. \tag{3.9}$$

Since $\lambda \in \rho(\tilde{A})$, we can operate on both sides of (3.9) to obtain

$$(\lambda I - \tilde{A})^{-1} x_t(\cdot\,; 0, y_k) = \int_0^t T(t-\xi)\tilde{B}_\lambda y_k(\xi)\, d\xi.$$

The mapping $u(\cdot)\to\int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\,d\xi$, which takes $L^2[0,t:\mathbb{E}^m]$ into $W_2^{(1)}$, is easily shown to be continuous and hence

$$\lim_{k\to\infty}\int_0^t T(t-\xi)\tilde{B}_\lambda y_k(\xi)\,d\xi=\int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\,d\xi.$$

On the other hand the mapping of $L^2[0,t:\mathbb{E}^m]$ into $W_2^{(1)}$ defined by

$$u(\cdot)\to x_t(\cdot;0,u)$$

is continuous by (2.2). Therefore since $y_k\to u$ in $L^2[0,t:\mathbb{E}^m]$

$$\lim_{k\to\infty}x_t(\cdot;0,y_k)=x_t(\cdot;0,u).$$

By continuity of $(\lambda I-\tilde{A})^{-1}$, this implies,

$$\int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\,d\xi=(\lambda I-\tilde{A})^{-1}x_t(\cdot;0,u),$$

hence

$$\int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\,d\xi\in\mathcal{D}(\tilde{A})$$

and

$$x_t(\cdot;\phi,u)=(\lambda I-\tilde{A})\int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\,d\xi,$$

as required.

Remarks on the variation of constants formula (3.4). (i) The variation of constants formula (3.4) is independent of the choice of $\lambda\in\rho(\tilde{A})$.

(ii) We cannot differentiate (3.4) to obtain an abstract differential equation for the state $x_t(\cdot;\phi,u)$ since an unbounded operator $(\lambda I-\tilde{A})$ operates on the integral term. However we can associate the state $x_t(\cdot;\phi,u)$ with the differential equation in $W_2^{(1)}$

$$\begin{aligned}&\frac{d}{dt}y(t)=\tilde{A}y(t)+\tilde{B}_\lambda u(t),\qquad t>0,\\&y(0)=(\lambda I-\tilde{A})^{-1}\phi\end{aligned}\tag{3.10}$$

and the output equation $x_t(\cdot;\phi,u)=(\lambda I-\tilde{A})y(t)$. That is, if we define $y(t)\equiv(\lambda I-\tilde{A})^{-1}x_t(\cdot;\phi,u)$,

$$y(t)=T(t)(\lambda I-\tilde{A})^{-1}\phi+\int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\,d\xi\tag{3.11}$$

then if the control term $u(\cdot)$ is sufficiently smooth, (3.11) can be differentiated to yield (3.10).

(iii) In the retarded case, $(A_k=0,k=1,\cdots,N)$, it is well known, [4], that the state $x_t(\cdot;\phi,u)$ can be represented by an element $\tilde{x}(t)$ of the product space $M_2[-h,0:\mathbb{E}^n]=\mathbb{E}^n\times L^2[-h,0:\mathbb{E}^n]$,

$$\tilde{x}(t)=(x(t;\phi,u),x_t(\cdot;\phi,u)).$$

Moreover $\tilde{x}(t)$ satisfies the variation of constants formula in $M_2[-h, 0: \mathbb{E}^n]$

$$\tilde{x}(t) = \hat{T}(t)(\phi(0), \phi) + \int_0^t \hat{T}(t-\xi)\hat{B}u(\xi)\, d\xi, \tag{3.12}$$

where $\{\hat{T}(t), t \geqq 0\}$ is the semigroup extension of $\{T(t), t \geqq 0\}$ on $M_2[-h, 0: \mathbb{E}^n]$. The infinitesimal generator $\hat{A}$, of $\{\hat{T}(t), t \geqq 0\}$ is defined by

$$\hat{A}(\psi^0, \psi^1) = \left(\sum_{k=0}^{N} E_k \psi^1(-h_k) + \int_{-h}^{0} E(\theta)\psi^1(\theta)\, d\theta, \dot{\psi}^1(\cdot) \right)$$

for all elements in $\mathcal{D}(\hat{A}) = \{(\psi^0, \psi^1) \in M_2[-h, 0: \mathbb{E}^n] | \psi^1 \in W_2^{(1)}, \psi^1(0) = \psi^0\}$. $\hat{B}$ is the bounded linear operation from $\mathbb{E}^m$ into $M_2[-h, 0: \mathbb{E}^n]$ defined by

$$\hat{B}u = (Bu, 0).$$

The formula (3.12) can be seen to be the extension of (3.4) to the space $M_2[-h, 0: \mathbb{E}^n]$. Indeed, let i be the canonical injection of $W_2^{(1)}$ into $M_2[-h, 0: \mathbb{E}^n]$,

$$i(\phi) = (\phi(0), \phi(\cdot)).$$

Then it is easy to show that

$$\hat{T}(t)i(\phi) = i(T(t)\phi) \qquad \text{for } \phi \in W_2^{(1)}, \tag{3.13}$$

$$i(\tilde{A}\phi) = \hat{A}i(\phi) \qquad \text{for } \phi \in \mathcal{D}(\tilde{A}), \tag{3.14}$$

$$(\lambda I - \hat{A})^{-1}\hat{B}u = i(\tilde{B}_\lambda u) \qquad \text{for } u \in \mathbb{E}^m. \tag{3.15}$$

Recall that the state $x_t(\cdot\,; \phi, u)$ in $W_2^{(1)}$ satisfies,

$$x_t(\cdot\,; \phi, u) = T(t)\phi + (\lambda I - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\, d\xi.$$

If we embed $x_t(\cdot\,; \phi, u)$ into $M_2[-h, 0: \mathbb{E}^n]$ using (3.13), (3.14) we obtain,

$$\tilde{x}(t) = i(x_t(\cdot\,; \phi, u)) = i(T(t)\phi) + i\left((\lambda I - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\, d\xi \right)$$

$$= \hat{T}(t)i(\phi) + (\lambda I - \hat{A}) \int_0^t \hat{T}(t-\xi) i(\tilde{B}_\lambda u(\xi))\, d\xi.$$

Since by (3.15) $i(\tilde{B}_\lambda u(\xi))$ is in the domain of $\hat{A}$ the term $(\lambda I - \hat{A})$ can be brought inside the integral to yield

$$\tilde{x}(t) = \hat{T}(t)i(\phi) + \int_0^t (\lambda I - \hat{A})\hat{T}(t-\xi) i(\tilde{B}_\lambda u(\xi))\, d\xi.$$

Since a semigroup commutes with its generator and by (3.15) we obtain

$$\tilde{x}(t) = \hat{T}(t)i(\phi) + \int_0^t \hat{T}(t-\xi)\hat{B}u(\xi)\, d\xi.$$

Thus, in the retarded case, the variation of constants formula (3.4) is a restriction of the well-known variation of constants formula (3.12) for retarded systems in $M_2[-h, 0: \mathbb{E}^n]$.

4. Abstract operator conditions for state controllability. We now apply the results of the previous sections to the function space controllability problem in $W_2^{(1)}$ for the linear neutral control system (2.1).

DEFINITION 4.1. System (2.1) is said to be *approximately controllable* if and only if for every ϕ, ψ in $W_2^{(1)}$ and all $\varepsilon > 0$, there exists a time $t_1 > 0$ (t_1 possibly depending on ϕ, ψ, ε), and a control $u(\cdot)$ an element in $L^2[0, t_1: \mathbb{E}^m]$, such that

$$\|x_{t_1}(\cdot\,; \phi, u) - \psi\| < \varepsilon.$$

As usual we can restate this definition in terms of properties of attainable states.

DEFINITION 4.2. The attainable states at time t of system (2.1) is the linear manifold K_t in $W_2^{(1)}$ defined by

$$K_t = \{x_t(\cdot\,; 0, u) | u \in L^2[0, t: \mathbb{E}^m]\}.$$

By (3.4) we can express K_t as

$$K_t = \left\{(\lambda - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\, d\xi \,\middle|\, u \in L^2[0, t: \mathbb{E}^m]\right\}.$$

It is easily seen that $K_{t_1} \subset K_{t_2}$ if $t_2 > t_1$ and we define

$$K_\infty = \bigcup_{t>0} K_t.$$

A simple argument shows that system (2.1) is approximately controllable if and only if K_∞ is dense in $W_2^{(1)}$. Necessary and sufficient conditions for density of K_∞ can now be found by applying a general characterization of $K_\infty^\perp$ for boundary control systems found in [5]. We thus obtain the following result.

THEOREM 4.1. *System* (2.1) *is approximately controllable if and only if for all* $\psi \in W_2^{(1)}$

$$\left(\psi(0) + \int_{-h}^0 \lambda e^{\lambda\theta}\dot{\psi}(\theta)\, d\theta\right)^T \Delta^{-1}(\lambda)B = 0, \quad \text{for all } \lambda \in \rho(\tilde{A}) \tag{4.1}$$

implies $\psi \equiv 0$.

Proof. K_∞ is dense in $W_2^{(1)}$ if and only if the orthogonal complement $K_\infty^\perp$ is the trivial subspace $\{0\}$. A function ψ in $K_\infty^\perp$ is characterized by

$$\left(\psi \middle| (\lambda - \tilde{A}) \int_0^t T(t-\xi)\tilde{B}_\lambda u(\xi)\, d\xi\right) = 0$$

for all $t > 0$ and all infinitely differentiable $\mathbb{E}^m$-valued functions $u(\cdot)$ on $[0, t]$ with $u(0) = 0$.

For $t > 0$, let $f(\xi) \triangleq u(t-\xi)$, $\xi \in [0, t]$. Using the abstract integration by parts formula [9], the above condition can be replaced by the following condition: ψ is in $K_\infty^\perp$ if and only if

$$\left(\psi \middle| \int_0^t T(\xi)\{\lambda\tilde{B}_\lambda f(\xi) + \tilde{B}_\lambda \dot{f}(\xi)\}\, d\xi + \tilde{B}_\lambda f(0)\right) = 0 \tag{4.2}$$

for all $t > 0$ and all infinitely differentiable functions f on $[0, t]$ such that $f(t) = 0$.

Another integration by parts allows us to write (4.2) in the form,

$$\begin{aligned}\int_0^t \left(\left\{\tilde{B}_\lambda^* T^*(\xi) - \bar{\lambda}\tilde{B}_\lambda^* \int_0^\xi T^*(r)\, dr - \tilde{B}_\lambda^*\right\}\psi \middle| \dot{f}(\xi)\right) d\xi \\ + \left(\psi \middle| \int_0^t T(\xi)\, d\xi\, \lambda\tilde{B}_\lambda f(t) + \tilde{B}_\lambda f(t)\right) = 0.\end{aligned} \tag{4.3}$$

Expression (4.3) can hold for all $t>0$, all infinitely differentiable $f(\cdot)$ on $[0, t]$ such that $f(t)=0$ if and only if for all $\xi \geqq 0$

$$\left\{\tilde{B}_{\lambda}^{*} T^{*}(\xi)-\bar{\lambda} \tilde{B}_{\lambda}^{*} \int_{0}^{\xi} T^{*}(r)\, dr-\tilde{B}_{\lambda}^{*}\right\} \psi=0. \tag{4.4}$$

If we take the Laplace transform of (4.4) we obtain

$$\{\bar{\lambda} \tilde{B}_{\lambda}^{*}(\mu I-\tilde{A}^{*})^{-1}-\mu \tilde{B}_{\lambda}^{*}(\mu I-\tilde{A}^{*})^{-1}+\tilde{B}_{\lambda}^{*}\} \psi=0 \tag{4.5}$$

for all μ with sufficiently large real part, and by analytic continuation we can extend (4.5) to all $\mu \in \rho(\tilde{A})$.

On the other hand (4.5) implies (4.3) by the uniqueness of the Laplace transform; thus ψ is an element in $K_{\infty}^{\perp}$ if and only if (4.5) is satisfied.

A straightforward calculation (see Appendix) shows that

$$(\lambda-\bar{\mu})(\bar{\mu} I-\tilde{A})^{-1} \tilde{B}_{\lambda}=\tilde{B}_{\bar{\mu}}-\tilde{B}_{\lambda}.$$

Taking adjoints, this implies that

$$(\bar{\lambda}-\mu) \tilde{B}_{\lambda}^{*}(\mu I-\tilde{A}^{*})^{-1}=\tilde{B}_{\bar{\mu}}^{*}-\tilde{B}_{\lambda}^{*}.$$

Thus for all $\lambda \in \rho(\tilde{A})$, (4.5) can be rewritten

$$\tilde{B}_{\bar{\mu}}^{*} \psi=0 \quad \text{for all } \mu \in \rho(\tilde{A}).$$

The operator $\tilde{B}_{\bar{\mu}}^{*}$ is characterized by (see Appendix)

$$\tilde{B}_{\bar{\mu}}^{*} \psi=B^{T} \Delta^{-T}(\mu)\left\{\psi(0)+\int_{-h}^{0} \mu\, e^{\mu \theta} \dot{\psi}(\theta)\, d\theta\right\}.$$

Therefore $\psi \in K_{\infty}^{\perp}$ if and only if

$$B^{T} \Delta^{-T}(\mu)\left\{\psi(0)+\int_{-h}^{0} \mu\, e^{\mu \theta} \dot{\psi}(\theta)\, d\theta\right\}=0 \quad \text{for all } \mu \in \rho(\tilde{A}).$$

Since approximate controllability will hold if and only if $K_{\infty}^{\perp}$ is the trivial subspace and taking transposes, the theorem is proved.

Remarks on Condition (4.1).

(1) Approximate controllability condition (4.1) is similar to a condition derived in [11] for retarded systems in the space $M_2[-h, 0: \mathbb{E}^n]$. Moreover, it can be shown [15], that for retarded systems, approximate controllability in the state spaces $W_2^{(1)}$ and $M_2[-h, 0: \mathbb{E}^n]$ are equivalent.

(2) If in addition to approximate controllability, the attainable set K_{∞} is closed in $W_2^{(1)}$ the system is said to be exactly controllable. That is, any state in $W_2^{(1)}$ can be reached exactly from the zero state in finite time. For neutral systems of the following type:

$$\frac{d}{dt}[x(t)-A_1 x(t-h)]=E_0 x(t)+E_1 x(t-h)+Bu(t), \tag{4.6}$$

closedness of the attainable set and exact controllability in $W_2^{(1)}$ has been studied [2], [8], [18]. It is shown in [8], that system (4.6) has a closed attainable set if the pair (A_1, B) is controllable and that (A_1, B) being controllable is a necessary condition for exact controllability in $W_2^{(1)}$. Therefore necessary and sufficient conditions for (4.6)

to be exactly controllable are

i) (A_1, B) is a controllable pair,

ii) (4.6) is approximately controllable.

(3) Conditions similar to (4.1) for approximate controllability of neutral systems with delay in control can be obtained [15].

(4) Our interest lies in systems with real coefficients. However, since we must make extensive use of the spectrum and resolvent which are complex-valued, we developed our results in the complex space $W_2^{(1)}[-h, 0: \mathbb{E}^n]$ rather than in the corresponding real space. This in no way limits the applicability of our results to real systems. Indeed the following corollary is an obvious consequence of Theorem 4.1.

COROLLARY 4.2. *System* (2.1) *is approximately controllable* (*with* $L_{\text{loc}}^2[0, \infty: \mathbb{R}^m]$ *controls*) *in the real linear space* $W_2^{(1)}[-h, 0: \mathbb{R}^n]$ *if and only if for all* ψ *in* $W_2^{(1)}[-h, 0: \mathbb{R}^n]$, (4.1) *implies* $\psi \equiv 0$.

5. Algebraic conditions for approximate controllability. In this section we will limit our attention to neutral systems with a single point delay in the state:

$$\begin{aligned} &\frac{d}{dt}(x(t)-A_1x(t-h)) = E_0x(t)+E_1x(t-h)+Bu(t), \\ &x(t) = \phi(t), \qquad t \in [-h, 0]. \end{aligned} \tag{5.1}$$

We will develop necessary and sufficient conditions for approximate controllability of (5.1) based on the abstract controllability condition (4.1). This will extend the results of [11], [14] for retarded systems to systems of neutral type.

From (4.1) we have that (5.1) is approximately controllable if and only if there is no nonzero $\psi \in W_2^{(1)}$ such that

$$\left(\psi(0)+\lambda\int_{-h}^{0} e^{\lambda\theta}\dot{\psi}(\theta)\,d\theta\right)^T \Delta^{-1}(\lambda)B = 0 \quad \text{for all } \lambda \in \rho(\tilde{A}),$$

where

$$\Delta(\lambda) = \lambda I - A_1\lambda\, e^{-\lambda h} - E_0 - E_1\, e^{-\lambda h}.$$

It is convenient to express this condition in the well-known notation of [11]. First we define

$$\begin{aligned} \text{FLT}[0, h] = \{q(\lambda) | q(\lambda) &\text{ is an } n \times 1 \text{ entire function which is the} \\ &\text{finite Laplace transform of an } L_2[0, h: \mathbb{E}^n] \text{ function}\}. \end{aligned}$$

Then system (5.1) is approximately controllable if and only if for $x \in \mathbb{E}^n$, $q \in \text{FLT}[0, h]$

$$[x+\lambda q(\lambda)]^T \Delta^{-1}(\lambda)B = 0$$

for all $\lambda \in \rho(\tilde{A})$ implies $x = 0$, $q \equiv 0$.

Equivalently, approximate controllability holds if and only if for $x \in \mathbb{E}^n$, $q \in \text{FLT}[0, h]$

$$[x+\lambda q(\lambda)]^T \text{ adj } \Delta(\lambda)B = 0$$

for all $\lambda \in \mathbb{C}$ implies $x = 0$, $q \equiv 0$.

It is convenient to expand adj $\Delta(\lambda)B$ in powers of $e^{-\lambda h}$,

$$\text{adj } \Delta(\lambda)B = \sum_{k=0}^{n-1} P_k(\lambda)(e^{-\lambda h})^k B.$$

Here, $P_k(\lambda)$ are $n \times n$ matrix polynomials of degree at most $(n-1)$. We can also write,

$$\operatorname{adj} \Delta(\lambda)B = P[\lambda]V(e^{-\lambda h}),$$

where $P[\lambda] = [P_0(\lambda)B, \cdots, P_{n-1}(\lambda)B]$ is an $n \times mn$ polynomial matrix and $V(e^{-\lambda h})$ is the $mn \times m$ matrix

$$V(e^{-\lambda h}) = (I_m, I_m e^{-\lambda h}, \cdots, I_m e^{-(n-1)\lambda h})^T, \qquad I_m = m \times m \text{ identity.}$$

Adopting this further notation, we see that system (5.1) is approximately controllable if and only if for all $x \in \mathbb{E}^n$, $q \in \mathrm{FLT}[0, h]$

$$[x + \lambda q(\lambda)]^T P[\lambda] V(e^{-\lambda h}) \equiv 0 \quad \text{for all } \lambda \in \mathbb{C} \tag{5.2}$$

implies $x = 0$, $q \equiv 0$.

The matrix $P[\lambda]$ plays an important role in determining the structural properties of system (5.1). Properties of $P[\lambda]$ determine the following necessary conditions for approximate controllability.

PROPOSITION 5.1. *A necessary condition for approximate controllability is that* $\operatorname{rank}_{\mathbb{C}} P[\lambda] = n$.

This result is proved in [11] for retarded systems and using (5.2) the result is easily extended to neutral systems. Next we will state two conditions for $\operatorname{rank}_{\mathbb{C}} P[\lambda] = n$ which are straightforward generalizations of [11, Thm. 3.4 and Thm. 3.6] to (5.1).

PROPOSITION 5.2. i) $\operatorname{rank}_{\mathbb{C}} P[\lambda] = n$ *is equivalent to*

$$\operatorname{rank}[(\lambda_1 I - E_0)^{-1}B, ((\lambda_1 I - E_0)^{-1}(E_1 + \lambda_1 A_1))(\lambda_1 I - E_0)^{-1}B, \cdots, ((\lambda_1 I - E_0)^{-1}(E_1 + \lambda_1 A_1))^{n-1}(\lambda_1 I - E_0)^{-1}B] = n$$

for some λ_1 *not an eigenvalue of* E_0.

ii) *A necessary condition for* $\operatorname{rank}_{\mathbb{C}} P[\lambda] = n$ *is that* $\operatorname{rank}_{\mathbb{C}}[\lambda A_1 + E_1, B] = n$.

Finding easily computable sufficient conditions for approximate controllability based on criterion (5.2) seems to be a difficult problem. In order to obtain a sufficient condition we will not work with (5.2) but consider a related concept of controllability.

We say that system (5.1) is spectrally controllable if and only if for each $\lambda \in \sigma(\tilde{A})$ the projection of (5.1) onto the generalized eigenspace of λ, $\mathcal{M}_\lambda$, (e.g., [6]), is a completely controllable finite dimensional system. As is well known [16], (5.1) will be spectrally controllable if and only if for each $\lambda \in \sigma(\tilde{A})$

$$\operatorname{rank}[\Delta(\lambda), B] = n. \tag{5.3}$$

Spectral controllability is in general a weaker concept than approximate controllability. However the two concepts are equivalent in the case where the system of eigenprojections associated with (5.1) is complete in $W_2^{(1)}$.

DEFINITION 5.3. System (5.1) is *spectrally complete* in $W_2^{(1)}$ if

$$\operatorname{span}\{\mathcal{M}_\lambda | \lambda \in \sigma(\tilde{A})\} \text{ is a dense subset of } W_2^{(1)},$$

where $\mathcal{M}_\lambda = \bigcup_{n \geq 0} N[(\lambda - \tilde{A})^n]$ is the generalized eigenspace of λ.

Since spectral controllability implies that every element in $\operatorname{span}\{\mathcal{M}_\lambda | \lambda \in \sigma(\tilde{A})\}$ can be reached, spectral controllability and spectral completeness imply that the set of attainable states is dense in $W_2^{(1)}$, hence the system is approximately controllable.

Conditions for spectral completeness of retarded systems in various spaces have been obtained in [10], [12], [13]. We now give some conditions for spectral completeness of neutral systems in $W_2^{(1)}$.

THEOREM 5.4. *System* (2.1) *is spectrally complete if and only if there is no* ϕ *in* $W_2^{(1)}$, $\phi \neq 0$ *such that*

$$\left[\phi(0)+\int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}(\theta)\,d\theta\right]^T \Delta^{-1}(\lambda)$$

is entire.

Proof. The set $\sigma(\tilde{A})$ is a countable set of points whose only limit point is at ∞. We can order the elements of $\sigma(\tilde{A})$ to obtain a sequence $\{\lambda_k\}_{k\geq 1}$. Each eigenvalue λ_k is a root of $\det \Delta(\lambda)=0$ of order m_k.

We will now prove the theorem in two steps.

Step 1. System (2.1) is spectrally complete if and only if

$$(\phi|\psi_k)=0 \quad \text{for } k=1,2,\cdots$$

implies $\phi=0$, where ψ_k is an element of the generalized eigenspace of λ_k, $\mathcal{M}_{\lambda_k}$.

The elements ψ_k are characterized by

$$\psi_k(\theta)=\sum_{j=0}^{m_k-1}\frac{\theta^j}{j!}(e^{\theta\lambda_k}\gamma_{j+1}),$$

where $\gamma=(\gamma_1^T,\gamma_2^T,\cdots,\gamma_{m_k}^T)^T$ is an element of $N(\hat{H}_k)$, and $\hat{H}_k$ is the $(nm_k\times nm_k)$ matrix

$$\hat{H}_k=\begin{bmatrix} \Delta(\lambda_k) & \Delta^{(1)}(\lambda_k) & \cdots & \dfrac{1}{(m_k-1)!}\Delta^{(m_k-1)}(\lambda_k) \\ 0 & \Delta(\lambda_k) & & \vdots \\ 0 & 0 & \ddots & \\ \vdots & & & \Delta^{(1)}(\lambda_k) \\ 0 & \cdots & & \Delta(\lambda_k) \end{bmatrix}.$$

See for example [6, Chapt. 7].

If we define the $1\times n$ vector $F(\lambda)$,

$$F(\lambda)=\left[\phi(0)+\int_{-h}^{0}\lambda e^{\lambda\theta}\dot{\phi}(\theta)\,d\theta\right]^T$$

then it is easy to show that for all k

$$(\bar{\phi}|\psi_k)=\sum_{j=0}^{m_k-1}\frac{1}{j!}F^{(j)}(\lambda_k)\gamma_{j+1}.$$

Therefore system (2.1) is spectrally complete if and only if

$$\sum_{j=0}^{m_k-1}\frac{1}{j!}F^{(j)}(\lambda_k)\gamma_{j+1}=0 \quad \text{for } k=1,2,\cdots \tag{5.4}$$

mplies $\phi\equiv 0$.

Writing (5.4) in matrix form, we have

$$\theta_k^T=\left[F(\lambda_k),F^{(1)}(\lambda_k),\cdots,\frac{1}{(m_k-1)!}F^{(m_k-1)}(\lambda_k)\right] \quad \text{and} \quad \theta_k^T\gamma=0.$$

Since γ can be any element in $N(\hat{H}_k)$, system (2.1) is spectrally complete if and only if there is no nonzero ϕ in $W_2^{(1)}$ such that the sequence of $(nm_k\times 1)$ vectors $\{\theta_k\}_{k\geq 1}$ satisfies

$$\theta_k\in(N(\hat{H}_k))^{\perp} \quad \text{for } k=1,2,\cdots. \tag{5.5}$$

Step 2. We will show that (5.5) holds if and only if $F(\lambda)\Delta^{-1}(\lambda)$ is entire.

If $F(\lambda)\Delta^{-1}(\lambda)$ is entire, then there exists a $1\times n$ vector function $G(\lambda)$, whose elements are entire functions and

$$F(\lambda)=G(\lambda)\Delta(\lambda) \quad \text{for all } \lambda\in\mathbb{C}. \tag{5.6}$$

If we evaluate the first m_k coefficients of the Taylor series expansion of (5.6) about $\lambda=\lambda_k$ we obtain for $j=1,2,\cdots,m_k$, $k=1,2,\cdots$

$$\frac{1}{j!}F^{(j)}(\lambda_k)=\sum_{i=0}^{j}\left(\frac{1}{(j-i)!}G^{(j-i)}(\lambda_k)\right)\left(\frac{1}{i!}\Delta^{(i)}(\lambda_k)\right);$$

hence for $k=1,2,\cdots$

$$\theta_k^T=\left[G(\lambda_k),G^{(1)}(\lambda_k),\cdots,\frac{1}{(m_k-1)!}G^{(m_k-1)}(\lambda_k)\right]\hat{H}_k.$$

Therefore

$$\theta_k^T\in \operatorname{Im}(\hat{H}_k^T)=(N(\hat{H}_k))^{\perp} \quad \text{for } k=1,2,\cdots.$$

Alternatively, suppose (5.5) holds; we will show that $F(\lambda)\Delta^{-1}(\lambda)$ is entire. We know that $F(\lambda)\Delta^{-1}(\lambda)$ is a $1\times n$ vector-valued function on $\mathbb{C}$ whose elements are analytic in $\rho(\tilde{A})$ and have a possibly infinite number of poles at $\lambda=\lambda_k$, $k=1,2,\cdots$. We claim that if (5.5) holds, the zeros of $F(\lambda)$ will cancel the poles of $\Delta^{-1}(\lambda)$ and therefore $F(\lambda)\Delta^{-1}(\lambda)$ will be entire. More precisely, for each k we will show that if the Laurent expansion of $\Delta^{-1}(\lambda)$ at $\lambda=\lambda_k$ is given by

$$\Delta^{-1}(\lambda)=\frac{U_1}{(\lambda-\lambda_k)}+\frac{U_2}{(\lambda-\lambda_k)^2}+\cdots+\frac{U_{r_k}}{(\lambda-\lambda_k)^{r_k}}+\sum_{j=0}^{\infty}Q_j(\lambda-\lambda_k)^j,$$

where $r_k\leqq m_k$, U_j, Q_j are $n\times n$ constant matrices, then

$$F(\lambda)[U_1(\lambda-\lambda_k)^{r_k-1}+U_2(\lambda-\lambda_k)^{r_k-2}+\cdots+U_{r_k}]=(\lambda-\lambda_k)^{r_k}Q(\lambda), \tag{5.7}$$

where $Q(\lambda)$ is a $1\times n$ vector-valued function whose elements are all entire.

Clearly the entire function

$$F(\lambda)[U_1(\lambda-\lambda_k)^{r_k-1}+U_2(\lambda-\lambda_k)^{r_k-2}+\cdots+U_{r_k}] \tag{5.8}$$

will satisfy (5.7) if the first r_k coefficients in the Taylor series expansion of (5.8) are zero. That is if for $j=1,2,\cdots,r_k-1$

$$\frac{1}{j!}\frac{d^j}{d\lambda^j}[F(\lambda)(U_1(\lambda-\lambda_k)^{r_k-1}+\cdots+U_{r_k})]\Big|_{\lambda=\lambda_k}=0. \tag{5.9}$$

Expression (5.9) may be written as the matrix equation,

$$\theta_k^T\hat{Q}_k=0, \tag{5.10}$$

where

$$\hat{Q}_k = \begin{bmatrix} \overbrace{\begin{matrix} U_{r_k} & U_{r_{k-1}} & \cdots & U_1 \\ 0 & U_{r_k} & \cdots & U_2 \\ \vdots & & & \vdots \\ 0 & & & U_{r_{k-1}} \\ 0 & \cdots & 0 & U_{r_k} \end{matrix}}^{nr_k} \\ \begin{matrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{matrix} \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ b \\ c \\ d \\ e \end{matrix}}\right\} nr_k \\ \left.\vphantom{\begin{matrix} a \\ b \\ c \end{matrix}}\right\} n(m_k - r_k) \end{matrix}$$

By repeated differentiation of the identity, $\Delta(\lambda)\Delta(\lambda)^{-1} = I$, it is easily seen that $\hat{H}_k\hat{Q}_k = 0$, $k = 1, 2, \cdots$. Therefore,

$$\text{Im}\,(\hat{Q}_k) \subset N(\hat{H}_k) \quad \text{and} \quad (\text{Im}\,(\hat{Q}_k))^{\perp} \supset (N(\hat{H}_k))^{\perp}.$$

By the assumption (5.5),

$$\theta_k \in (N(\hat{H}_k))^{\perp} \subset (\text{Im}\,(\hat{Q}_k))^{\perp} \quad \text{for } k = 1, 2, \cdots;$$

therefore

$$\theta_k^T \hat{Q}_k y = 0 \quad \text{for all } y \in \mathbb{E}^{r_k \cdot n}$$

and thus (5.10) holds and $F(\lambda)\Delta^{-1}(\lambda)$ is entire; the theorem is proved.

THEOREM 5.5. *System* (5.1) *is spectrally complete if and only if* $\det(\lambda A_1 + E_1)$ *is not identically zero.*

Proof. First, suppose that $\det(\lambda A_1 + E_1)$ is not identically zero and $G(\lambda) = [\phi(0) + \int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}(\theta)\, d\theta]^T \Delta^{-1}(\lambda)$ is entire. We will show that $\phi \equiv 0$; therefore by Theorem 5.4 the system is spectrally complete.

Let $F(\lambda) = [\phi(0) + \int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}(\theta)\, d\theta]^T$; we denote the elements of $G(\lambda)$ and $F(\lambda)$ by $g_j(\lambda)$ and $f_j(\lambda)$, respectively. Let α be a real number such that the set, $\{\lambda \,|\det(I - A_1 e^{-\lambda h}) = 0\}$, lies in the open half plane $\text{Re}\,\lambda < \alpha$ (for example, let $\alpha > 1/h \ln(|\lambda^*(A_1)|)$, where $\lambda^*(A_1)$ is the eigenvalue of A_1 with largest modulus).

Next, we obtain growth estimates on G and F. Let $\lambda = \upsilon + i\omega$, where υ and ω are real numbers.

(1) *Estimate on F.*

(1a) For $\omega = 0$, $\upsilon > 0$, υ large

$$|f_j(\upsilon)| = O(\upsilon^{1/2}) \quad \text{as } \upsilon \to \infty.$$

(1b) For $\upsilon = \alpha$, $|\omega|$ large

$$|f_j(\alpha + i\omega)| = O(|\omega|) \quad \text{as } |\omega| \to \infty.$$

(1c) For $\omega = 0$, $\upsilon > 0$, υ large

$$|f_j(-\upsilon)| = O(\upsilon^{1/2} e^{\upsilon h}) \quad \text{as } \upsilon \to \infty.$$

(2) *Estimate on G*:

(2a) For $\omega = 0$, $\upsilon > 0$, υ large

$$G(\upsilon) = \left[\frac{F(\upsilon)}{\upsilon}\right]\left[\frac{\Delta(\upsilon)}{\upsilon}\right]^{-1}. \tag{5.11}$$

Since $(\Delta(\upsilon)/\upsilon)\to I$ as $\upsilon\to\infty$, then $[\Delta(\upsilon)/\upsilon]^{-1}$ is bounded as $\upsilon\to\infty$. The estimate (1a) and (5.11) give the estimate

$$|g_j(\upsilon)| = O(\upsilon^{-1/2}) \quad \text{as } \upsilon\to\infty.$$

(2b) For $\upsilon=\alpha$, $|\omega|$ large

$$G(\alpha+i\omega) = \left[\frac{F(\alpha+i\omega)}{\alpha+i\omega}\right]\left[\frac{\Delta(\alpha+i\omega)}{\alpha+i\omega}\right]^{-1}. \tag{5.12}$$

One has that

$$\frac{\Delta(\alpha+i\omega)}{\alpha+i\omega} - (I - A_1 e^{-(\alpha+i\omega)h}) \to 0 \quad \text{as } |\omega|\to\infty.$$

Since $\det(I-A_1 e^{-(\alpha+i\omega)h})$ is uniformly bounded away from zero for $-\infty<\omega<\infty$, $\det[\Delta(\alpha+i\omega)/(\alpha+i\omega)]$ is bounded away from zero as $|\omega|\to\infty$. The elements of $\Delta(\alpha+i\omega)/(\alpha+i\omega)$ are bounded for large $|\omega|$, therefore $[\Delta(\alpha+i\omega)/(\alpha+i\omega)]^{-1}$ is bounded as $|\omega|\to\infty$. The estimate (1b) and (5.12) yield the estimate

$$|g_j(\alpha+i\omega)| = O(1) \quad \text{as } \omega\to\infty.$$

(2c) For $\omega=0$, $\upsilon>0$, υ large

$$G(-\upsilon) = e^{-\upsilon h}F(-\upsilon)[e^{-\upsilon h}\Delta(-\upsilon)]^{-1}. \tag{5.13}$$

Since $\det(A_1\lambda+E_1)$ is not identically zero we can write

$$e^{-\upsilon h}\Delta(-\upsilon) = (\upsilon A_1 - E_1)[I - (\upsilon A_1 - E_1)^{-1}(\upsilon I + E_0)\, e^{-\upsilon h}].$$

The elements of $(\upsilon A_1 - E_1)^{-1}$ are rational functions so the last term in the expression tends to zero as $\upsilon\to\infty$. Hence, the inverse of the bracketed expression is bounded. The inverse of $(\upsilon A_1 - E_1)$ is $O(e^{\varepsilon\upsilon})$ for all $\varepsilon>0$ and in view of the estimate (1c) and (5.13) we obtain

$$|g_j(-\upsilon)| = O(e^{\upsilon\varepsilon}) \quad \text{for all } \varepsilon>0, \text{ as } \upsilon\to\infty.$$

We define the function G_α, $G_\alpha(\lambda)=G(\lambda+\alpha)$, for all $\lambda\in\mathbb{C}$. Let $g_{\alpha_j}(\lambda)$ denote an element of $G_\alpha(\lambda)$.

Clearly, $G_\alpha(\lambda)$ is entire and satisfies the estimates (2a), (2c) and

(2b)′ For $\upsilon=0$, $|\omega|$ large

$$|g_{\alpha_j}(\omega)| = O(1) \quad \text{as } |\omega|\to\infty.$$

The estimates, (2a), (2b)′, (2c) and an application of the Phragmen–Lindelof theorem, [10], [19], imply that $G_\alpha(\lambda)$ is uniformly bounded in $\mathbb{C}$. Since, by hypothesis, G_α is entire, it follows from Liouville's theorem that G_α is a constant. Moreover, since each element in $G_\alpha(\lambda)$ satisfies the estimate (2a), this constant must be zero.

Thus, for $\lambda\in\mathbb{C}$,

$$\left[\phi(0) + \int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}(\theta)\, d\theta\right]^T \Delta^{-1}(\lambda) = G(\lambda) = G_\alpha(\lambda-\alpha) = 0.$$

This implies, that for all $\lambda\in\rho(\tilde{A})$,

$$\left[\phi(0) + \int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}(\theta)\, d\theta\right] = 0, \tag{5.14}$$

and by analytic continuation (5.14) can be shown to hold for all $\lambda\in\mathbb{C}$.

Uniqueness of the Laplace transform and (5.14) imply that $\phi \equiv 0$.

On the other hand, suppose that $\det(\lambda A_1 + E_1) \equiv 0$. Then there exists an $n \times 1$ polynomial vector $p(\lambda)$, not identically zero, such that,

$$p^T(\lambda)(A_1\lambda + E_1) \equiv 0. \tag{5.15}$$

Let $q^T(\lambda) \triangleq p^T(\lambda)(\lambda I - E_0)$. We can find a function ϕ_1 in $W_2^{(1)}$, ϕ_1 not identically zero, such that,

$$\left[\phi_1(0) + \int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}_1(\theta)\, d\theta\right] = q(\lambda) \int_{-h}^{0} \lambda e^{\lambda\theta} g(\theta)\, d\theta,$$

where g is an element in $L_2[-h, 0: \mathbb{E}]$. This follows from [11, Lemma 3.3] by taking $\phi_1(0) = 0$ with $\dot{\phi}_1$ playing the role of η [11, p. 609].

If $\operatorname{Re}\lambda > \beta$, for some $\beta > 0$ sufficiently large, we can write

$$\Delta^{-1}(\lambda) = (\lambda I - E_0)^{-1}[I - (A_1\lambda + E_1)(\lambda I - E_0)^{-1} e^{-\lambda h}]^{-1}.$$

In fact, picking β such that,

$$\det[\lambda I - E_0] \neq 0 \quad \text{for } \operatorname{Re}\lambda \geqq \beta,$$

and

$$|(A_1\lambda + E_1)(\lambda I - E_0)^{-1} e^{-\lambda h}| < 1 \quad \text{for } \operatorname{Re}\lambda \geqq \beta,$$

we can write

$$\Delta^{-1}(\lambda) = (\lambda I - E_0)^{-1} \sum_{i=0}^{\infty} [(A_1\lambda + E_1)(\lambda I - E_0)^{-1} e^{-\lambda h}]^i$$

for $\operatorname{Re}\lambda \geqq \beta$.

Therefore, in the half plane $\operatorname{Re}\lambda \geqq \beta$

$$\begin{aligned}\left[\phi_1(0) + \int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}_1(\theta)\, d\theta\right]^T \Delta^{-1}(\lambda) \\ = \left(\int_{-h}^{0} \lambda e^{\lambda\theta} g(\theta)\, d\theta\right) p^T(\lambda) \sum_{i=0}^{\infty} [(A_1\lambda + E_1)(\lambda I - E_0)^{-1} e^{-\lambda h}]^i \\ = \left(\int_{-h}^{0} \lambda e^{\lambda\theta} g(\theta)\, d\theta\right) p^T(\lambda),\end{aligned} \tag{5.16}$$

where in the last step we used (5.15).

To conclude the argument we extend the identity

$$\left[\phi_1(0) + \int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}_1(\theta)\, d\theta\right]^T \Delta^{-1}(\lambda) = \left(\int_{-h}^{0} \lambda e^{\lambda\theta} g(\theta)\, d\theta\right) p^T(\lambda), \tag{5.17}$$

to all of $\mathbb{C}$.

Each element of the $1 \times n$ function

$$d(\lambda) = \left[\phi_1(0) + \int_{-h}^{0} \lambda e^{\lambda\theta}\dot{\phi}_1(\theta)\, d\theta\right]^T - \left(\int_{-h}^{0} \lambda e^{\lambda\theta} g(\theta)\, d\theta\right) p^T(\lambda)\Delta(\lambda)$$

is entire, and in view of (5.16), each element is zero in the half plane $\operatorname{Re}\lambda > \beta$. Since a nonzero entire function can have only isolated zeros this implies that $d(\lambda) \equiv 0$. Post multiplying $d(\lambda)$ by $\Delta^{-1}(\lambda)$ implies that (5.17) holds for all $\lambda \in \rho(\tilde{A})$ and by analytic continuation we can extend (5.17) to all of $\mathbb{C}$.

Identity (5.17) and Theorem 5.4 imply that the system is not spectrally complete.

COROLLARY 5.6. *A sufficient condition for approximate controllability of system* (5.1) *is*

(a) $\det(\lambda A_1 + E_1)$ *is not identically zero,*

(b) $\operatorname{rank}[\Delta(\lambda), B] = n$, *for all* $\lambda \in \sigma(\tilde{A})$.

Proof. The corollary follows immediately from Theorem 5.5 and the remarks following Proposition 5.2.

The requirements of Corollary 5.6 are quite restrictive since many systems of interest are not spectrally complete. Moreover, it is not required that a system be spectrally complete for spectral controllability and approximate controllability to be equivalent.

In his theory of controllability for retarded systems in $M_2[-h, 0: \mathbb{E}^n]$ Manitius [14] showed that under certain reasonable conditions a system that is not spectrally complete may be related by feedback to a spectrally complete system. Moreover, he demonstrated that both spectral controllability and approximate controllability are invariant under feedback. Adopting these ideas to system (5.1) we obtain the following necessary and sufficient conditions for approximate controllability.

THEOREM 5.7. *System* (5.1) *is approximately controllable if and only if*

(a) $\operatorname{rank}_{\mathbb{C}}[A_1\lambda + E_1, B] = n$,

(b) $\operatorname{rank}[\Delta(\lambda), B] = n$, *for all* $\lambda \in \sigma(\tilde{A})$.

Proof. Sufficiency. First we show there is a $m \times n$ matrix K such that

$$\operatorname{rank}_{\mathbb{C}}[A_1\lambda + E_1 + BK] = n. \tag{5.18}$$

To see this, suppose there is no $m \times n$ matrix K such that (5.18) holds. Then we can find an $n \times 1$ vector $p(\lambda)$ whose elements are polynomials $p(\lambda) \not\equiv 0$ and for all $\lambda \in \mathbb{C}$

$$p^T(\lambda)[A_1\lambda + E_1 + BK] = 0$$

for any $m \times n$ matrix K. In particular if $K = 0_{m\times n}$,

$$p^T(\lambda)[A_1\lambda + E_1] = 0 \quad \text{for all } \lambda \in \mathbb{C}. \tag{5.19}$$

Therefore, for all $m \times n$ matrices K,

$$p^T(\lambda)BK = 0 \quad \text{for all } \lambda \in \mathbb{C}.$$

Let $B = [b_1, b_2, \cdots, b_m]$ where b_j are the column vectors of B and for $j = 1, 2, \cdots, m$

$$K_j = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 0 \\ 1 & 1 & & 1 \\ 0 & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \leftarrow j\text{th row.}$$

Then for $j = 1, 2, \cdots, m$,

$$\text{hence,} \quad \begin{aligned} &p^T(\lambda)BK_j = 0 \quad \text{for all } \lambda \in \mathbb{C}; \\ &p^T(\lambda)B = 0 \quad \text{for all } \lambda \in \mathbb{C}. \end{aligned} \tag{5.20}$$

But conditions (5.19), (5.20), imply $\operatorname{rank}_{\mathbb{C}}[\lambda A_1 + E_1, B] < n$, contradicting hypothesis (a).

Next, let K be an $m \times n$ matrix such that (5.18) holds. Consider the neutral system

$$\dot{x}(t) = A_1 \dot{x}(t-h) + E_0 x(t) + (E_1 + BK)x(t-h) + Bv(t). \tag{5.21}$$

In view of (5.18), system (5.21) is spectrally complete and hence approximately controllable if

$$\operatorname{rank}[\Delta(\lambda) - BKe^{-\lambda h}, B] = n \tag{5.22}$$

for all $\lambda \in \sigma(\tilde{A})$.

However it is easy to see that (5.22) holds if and only if

$$\operatorname{rank}[\Delta(\lambda), B] = n$$

for all $\lambda \in \sigma(\tilde{A})$. Thus, in view of hypothesis (b), system (5.21) is approximately controllable.

Finally, we show that approximate controllability of (5.21) implies approximate controllability of (5.1). Indeed, let $x_t^1(\cdot; \phi, u)$, $x_t^2(\cdot; \phi, u)$ be the state trajectories of systems (5.1) and (5.21), respectively. For any $\psi \in W_2^{(1)}$ and $\varepsilon > 0$, there is a time $t > 0$ and control $w \in L^2[0, t: \mathbb{E}^m]$ such that

$$\|x_t^2(\cdot; 0, w) - \psi\| < \varepsilon.$$

But if we define the function $u \in L^2[0, t: \mathbb{E}^m]$

$$u(t) = Kx^2(t-h; 0, u) + w(t), \qquad t \geqq 0$$

then

$$x^1(t; 0, u) = x^2(t; 0, w) \quad \text{and} \quad \|x_t^1(\cdot; 0, u) - \psi\| < \varepsilon,$$

therefore system (5.1) is approximately controllable.

Necessity. Approximate controllability is a stronger concept than spectral controllability [16], therefore (b) is clearly necessary. Propositions 5.1 and 5.2 imply that (a) is also a necessary condition.

COROLLARY 5.8. *System* (5.1) *is exactly controllable if and only if*

(a) $\operatorname{rank}[B, A_1 B, \cdots, A_1^{n-1} B] = n$,

(b) $\operatorname{rank}[\Delta(\lambda), B] = n$, *for all* $\lambda \in \sigma(\tilde{A})$.

Proof. Sufficiency. By [8, Lemma 4.1] condition (a) implies $\operatorname{rank}_{\mathbb{C}} P[\lambda] = n$. Thus, by our Proposition 5.2, the conditions of Theorem 5.7 are satisfied and hence, the system is approximately controllable. By Remark (2) after Theorem 4.1, the system is exactly controllable.

Necessity. Clearly (b) is necessary for exact controllability. The necessity of (a) was demonstrated in [8, Proposition 2.2].

6. Examples. We will now illustrate the controllability criteria with the following examples.

(i) Consider the scalar neutral system

$$\frac{d^n}{dt^n} x(t) = \sum_{i=0}^{n} b_i \frac{d^i}{dt^i} x(t-h) + \sum_{i=0}^{n-1} a_i \frac{d^i}{dt^i} x(t) + u(t), \tag{6.1}$$

where a_i, b_i are real constants. The system can be put in matrix form (5.1) with

$$A_1=\begin{bmatrix}0 & 0 & \cdots & 0\\ \vdots & & & \vdots\\ 0 & 0 & \cdots & b_n\end{bmatrix},\qquad E_0=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & & & \vdots\\ 0 & 0 & & \cdots & 1\\ a_0 & a_1 & & \cdots & a_{n-1}\end{bmatrix},$$

$$E_1=\begin{bmatrix}0 & 0 & \cdots & 0\\ \vdots & & & \vdots\\ 0 & 0 & \cdots & 0\\ b_0 & b_1 & \cdots & b_{n-1}\end{bmatrix},\qquad B=\begin{bmatrix}0\\ 0\\ \vdots\\ 1\end{bmatrix}.$$

It is easy to check that rank $[\Delta(\lambda), B]=n$ for all $\lambda\in\mathbb{C}$. Therefore the system is spectrally controllable. However, when we compute the matrix $P[\lambda]$,

$$P[\lambda]=\begin{bmatrix}1 & 0 & \cdots & 0\\ \lambda & 0 & \cdots & 0\\ \vdots & \vdots & & \vdots\\ \lambda^{n-2} & 0 & \cdots & 0\\ \lambda^{n-1} & 0 & \cdots & 0\end{bmatrix},$$

we see that $\operatorname{rank}_{\mathbb{C}} P[\lambda]<n$ for $n>1$ and thus, by Proposition 5.1, the system is not approximately controllable.

(ii) Consider the matrix neutral system

$$\dot{x}(t)=\begin{bmatrix}0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 0\end{bmatrix}\dot{x}(t-h)+\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}x(t)+\begin{bmatrix}-1 & 1 & 0\\ 1 & 0 & 1\\ 0 & 1 & 1\end{bmatrix}x(t-h)+\begin{bmatrix}0 & 0\\ 1 & 0\\ 0 & 1\end{bmatrix}u(t).$$

If we compute the matrix $P[\lambda]$

$$P[\lambda]=\begin{bmatrix}0 & 0 & (\lambda+1)(\lambda-1) & 0 & -(\lambda+1) & (\lambda+1)^2\\ (\lambda-1)^2 & 0 & 0 & (\lambda+1)(\lambda-1) & -1 & (\lambda+1)\\ 0 & (\lambda-1)^2 & (\lambda-1) & (\lambda-1) & 1 & -(\lambda+1)\end{bmatrix},$$

then since the 3×3 minor

$$\begin{bmatrix}0 & 0 & -(\lambda+1)\\ 0 & (\lambda+1)(\lambda-1) & -1\\ (\lambda-1)^2 & (\lambda-1) & 1\end{bmatrix}$$

has determinant

$$(\lambda+1)^2(\lambda-1)^3\neq 0,$$

$\operatorname{rank}_{\mathbb{C}} P[\lambda]=3$ and the necessary condition for approximate controllability is met.

By Proposition 5.2 and Theorem 5.7 a sufficient condition for approximate controllability is that

$$\operatorname{rank}\begin{bmatrix}\lambda+e^{-\lambda h}-1 & -e^{-\lambda h}(\lambda+1) & 0 & 0 & 0\\ -e^{-\lambda h} & \lambda-1 & -e^{-\lambda h}(\lambda+1) & 1 & 0\\ 0 & -e^{-\lambda h} & \lambda-e^{-\lambda h}-1 & 0 & 1\end{bmatrix}=3$$

for all $\lambda\in\mathbb{C}$.

By checking the 3×3 minors we see that rank $[\Delta(\lambda^*), B]<3$ for some $\lambda^*\in\mathbb{C}$ if and only if

$$\lambda^*+e^{-\lambda^* h}-1=0 \quad \text{and} \quad \lambda^*+1=0.$$

This pair of equations has a solution only if $h=\ln 2$. Thus, if the delay h is equal to $\ln 2$ the system is not approximately controllable (or spectrally controllable). On the other hand, if $h\neq \ln 2$ the system is approximately controllable. In fact since

$$\text{rank}\,[B, A_1B, A_1^2B]=3,$$

it follows from Corollary 5.8 that the system is exactly controllable.

Appendix. We state below the proofs of results used in Theorem 4.1

(A1) $(\mu-\lambda)(\lambda I-\tilde{A})^{-1}\tilde{B}_\mu=\tilde{B}_\lambda-\tilde{B}_\mu$.

For any $u\in\mathbb{E}^m$, $\tilde{B}_\lambda u-\tilde{B}_\mu u$ is in $\mathcal{D}(\tilde{A})$. We can therefore write,

$$(\lambda I-\tilde{A})(\tilde{B}_\lambda u-\tilde{B}_\mu u)=(\mu-\lambda)\tilde{B}_\mu u.$$

Operating on both sides of the above expression with the resolvent $(\lambda I-\tilde{A})^{-1}$ yields the identity (A1).

(A2) *The adjoint of* $\tilde{B}_\lambda$. Let $u\in\mathbb{E}^m$, $\psi\in W_2^{(1)}$ and consider

$$\begin{aligned}(\psi|\tilde{B}_\lambda u)&=\langle\psi(0), \Delta^{-1}(\lambda)Bu\rangle+\int_{-h}^{0}\langle\dot{\psi}(\theta), \lambda\, e^{\lambda\theta}\Delta^{-1}(\lambda)Bu\rangle\, d\theta\\&=\left\langle B^T\Delta^{-T}(\bar{\lambda})\psi(0)+\int_{-h}^{0}B^T\Delta^{-T}(\bar{\lambda})\bar{\lambda}\, e^{\bar{\lambda}\theta}\dot{\psi}(\theta)\, d\theta, u\right\rangle\\&=(\tilde{B}_\lambda^*\psi|u).\end{aligned}$$

Thus

$$\tilde{B}_\lambda^*\psi=B^T\Delta^{-T}(\bar{\lambda})\left\{\psi(0)+\int_{-h}^{0}\bar{\lambda}\, e^{\bar{\lambda}\theta}\dot{\psi}(\theta)\, d\theta\right\}.$$

Acknowledgment. The authors wish to thank the reviewers for their thorough review and helpful comments. In particular, we appreciate the suggestions on how to improve Theorem 5.5 and the Appendix.

REFERENCES

[1] A. BALAKRISHNAN, *Identification and stochastic control of a class of distributed systems with boundary noise*, Proc. of the International Conference on Control Theory, June 1974, Springer-Verlag, New York.

[2] H. T. BANKS, M. W. JACOBS AND C. E. LANGENHOP, *Characterization of the controlled states in* $W_2^{(1)}$ *of linear hereditary systems*, this Journal, 13 (1975), pp. 611–649.

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

[4] A. BENSOUSSAN, M. C. DELFOUR AND S. K. MITTER, *Optimal control of linear hereditary systems with a quadratic cost criterion*, MIT, Rep. ESL-P-603, Cambridge, MA, 1975.

[5] H. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.

[6] J. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[7] D. HENRY, *Linear autonomous functional differential equations in the Sobolev space* $W_2^{(1)}$, Tech. Rep., Dept. Mathematics, Univ. of Kentucky, Lexington, 1974.

[8] M. Q. JACOBS AND C. E. LANGENHOP, *Criteria for function space controllability of linear neutral systems*, this Journal, 14 (1976), pp. 1009–1048.

[9] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[10] N. LEVINSON AND C. MCCALLA, *Completeness and independence of the exponential solutions of some functional differential equations*, Stud. Appl. Math., 53 (1974), pp. 1–15.

[11] 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.

[12] A. MANITIUS, *Function space controllability of retarded systems: Some new algebraic conditions*, Proc. Allerton Conference, October 1976.

[13] ———, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, Rep. CRM-755, University of Montreal, 1978.

[14] ———, *Necessary and sufficient conditions of approximate controllability for general linear retarded systems*, this Journal, 19 (1981), pp. 516–532.

[15] D. O'CONNOR, *State controllability and observability for linear neutral systems*, D.Sc. Thesis, Washington University, St. Louis, MO, 1978.

[16] L. PANDOLFI, *Stabilization of neutral functional differential equations*, J. Optim. Theory Appl., 20 (1976), pp. 191–204.

[17] A. PAZY, *Semi-groups of linear operators and applications to partial differential equations*, Lecture notes for a course at the University of Maryland, 1973.

[18] 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.

[19] E. C. TITCHMARSH, *The Theory of Functions*, Clarendon Press, Oxford, 1932.

[20] R. TRIGGIANI, *A cosine operator approach to modelling boundary input hyperbolic systems*, Proc. IFIP Conference, Modelling and Identification of Distributed Parameter Systems, June 1976, Springer-Verlag, New York.

[21] D. WASHBURN, *A semigroup theoretic approach to modelling of boundary input problems*, Proc. IFIP Conference, Modelling and Identification of Distributed Parameter Systems, June 1976, Springer-Verlag, New York.

[22] J. ZABCYZK, *A semigroup approach to boundary value control*, Proc. 2nd IFAC Symposium on Control of Distributed Parameter Systems, Coventry, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0001 $01.25/0

DISCRETE-TIME OBSERVERS AND PARAMETER DETERMINATION FOR DISTRIBUTED PARAMETER SYSTEMS WITH DISCRETE-TIME INPUT-OUTPUT DATA*

TOSHIHIRO KOBAYASHI†

Abstract. The aim of this paper is to study the estimation of unknown states and unknown input distribution functions for distributed parameter systems with discrete-time input–output data. First we construct finite dimensional discrete-time state observers to estimate unknown states and give the error estimates for distributed parameter systems with unknown inputs. Next we consider the determination of unknown input distribution functions using these estimated states. We discuss the relationship between observability and identifiability. The problem of determination of unknown functions is not well posed in general even if the distributed parameter systems are identifiable. We present and discuss a feasible approximation method by regularization which gives a constructive procedure to obtain approximately a true input distribution function. We also investigate limit properties of approximate solutions as the number of sampling periods tends to $+\infty$.

Key words. distributed parameter system, discrete-time observer, parameter determination, identifiability, well-posed approximation method

1. Introduction. The construction of a mathematical model for a system using input–output data is a very important problem from a theoretical and practical point of view. Even if system equations have been postulated, there still remains an equally important problem of determining initial states and system parameters from the input–output data.

Many control components deliver their outputs in discrete, or sampled-data, form. Whenever a digital computer constitutes a part of a control system, the continuous signal must be discretized in order to be digestible by the computer. Discrete-time control theory is of great interest because of its application in computer control. Thus it is very important to determine system states and system parameters from discrete-time input–output data. These problems are closely related to system observability and parameter identifiability. Discrete-time observability and discrete-time identifiability are investigated in [4], [5] for a class of distributed parameter systems.

In [6] the authors study the estimation of unknown states and unknown parameters for distributed parameter systems with continuous-time input–output data. They construct finite dimensional continuous-time observers to estimate unknown states and present a feasible approximation method by regularization to determine unknown input distribution functions. The theory presented in this paper parallels the theory of [6]. In continuous-time observer theory we cannot construct a finite-time settling observer, because it is impossible to set any characteristic value of the observer $-\infty$. In discrete-time observer theory, however, we can construct a finite-time settling observer by setting all the characteristic values of the observer 0. This is an essential difference between a continuous-time observer and a discrete-time observer.

The problem of continuous-time identification of parameters can be formulated as that of minimizing a functional on a finite time interval. A discrete-time distributed parameter system is never identifiable from only finitely many observations. Thus the problem of discrete-time identifications of parameters must be formulated as that of minimizing a functional on an infinite time interval. This imposes several restrictions on system operators. From a practical point of view, we should determine parameters

* Received by the editors November 25, 1980, and in revised form January 26, 1982.
† Department of Control Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu 804, Japan.

from only finitely many observations. Therefore, in the case of discrete-time identification of parameters, it is important to discuss limit properties of approximate solutions as the number of sampling periods tends to $+\infty$.

In this paper we consider distributed parameter systems with unknown initial states and unknown functions in the case of discrete-time observations. An initial state as well as an unknown function cannot be determined uniquely from the input–output data. Therefore we first construct finite-dimensional discrete-time observers using only discrete-time measurement data to estimate the system states at sampling times for the distributed parameter systems with unknown input sources. Next we consider the determination of unknown input distribution functions using these estimated states. We investigate the relationship between observability and identifiability for a more general class of systems than the system treated in [5]. We present and discuss a feasible approximation method to determine approximately unknown functions using discrete-time input–output data. We also investigate limit properties of approximate solutions as a number of intervals tends to $+\infty$.

2. System description. We consider the system described by the first-order evolution equation on a Hilbert space H

$$\begin{aligned} &\frac{du(t)}{dt} = Au(t) + F(t), \qquad t > 0, \\ &u(0) = u_0, \end{aligned} \tag{2.1}$$

where $u(t)$ is the system state vector and $F \in L_1(0, \infty; H)$ is the unknown input vector. The operator A is the infinitesimal generator of a strongly continuous semigroup $U(t)$ on H. The unknown initial state u_0 is assumed to be in H. Then there exists the unique mild solution $u(t)$ of the system (2.1) such that

$$u(t) = U(t)u_0 + \int_0^t U(t-s)F(s)\,ds \tag{2.2}$$

and $u \in C(0, \infty; H)$ ([1, p. 31]).

We consider the following discrete-time outputs

$$z_k = Mu(kT), \qquad k = 0, 1, \cdots, \tag{2.3}$$

where T is a sampling period and M is a bounded linear operator from H to the observation space E (E being a Hilbert space). To assume M to be bounded may be restrictive, because the boundary or point observations in H are not bounded in general. However using the abstract formulation ([1, Chap. 8]) the theory stated here can be generalized to include such cases. From (2.2) we obtain

$$u(kT) = U(kT)u_0 + \int_0^{kT} U(kT-s)F(s)\,ds, \qquad k = 1, 2, \cdots. \tag{2.4}$$

Since $U(t)$ is a strongly continuous semigroup on H, we can transform the right-hand side of (2.4) as follows

$$\begin{aligned} u(kT) &= U(T)U(\overline{k-1}T)u_0 + U(T)\int_0^{(k-1)T} U(\overline{k-1}T-s)F(s)\,ds \\ &\quad + \int_{(k-1)T}^{kT} U(kT-s)F(s)\,ds \\ &= U(T)u(\overline{k-1}T) + \int_{(k-1)T}^{kT} U(kT-s)F(s)\,ds. \end{aligned}$$

Putting

$$F_k = \int_{kT}^{(k+1)T} U(\overline{k+1}T - s)F(s)\,ds, \qquad k = 1, 0, \cdots, \tag{2.5}$$

we have the discrete-time system

$$\begin{aligned} &u(\overline{k+1}T) = U(T)u(kT) + F_k, \qquad k = 0, 1, \cdots, \\ &u(0) = u_0. \end{aligned} \tag{2.6}$$

In § 4 we consider the input to a continuous system (2.1) as a sequence of constants f_j. That is, $F(t) = gf_j$, $jT \leqq t < (j+1)T$, $j = 0, 1, \cdots$. Then $F(t)$ is also written

$$F(t) = g \sum_{j=0}^{[t]} f_j(Y(t - jT) - Y(t - \overline{j+1}T)), \tag{2.7}$$

where $g \in H$ is the unknown input distribution function and $\{f_j\} \in l_2$ is an input signal. $Y(t)$ is the heaviside function and $[\cdot]$ is the Gauss bracket. In this case F also belongs to $L_2(0, \infty; H)$ and F_k becomes

$$F_k = \int_0^T U(T - s)g\,ds \cdot f_k = V(T)gf_k, \qquad k = 0, 1, \cdots. \tag{2.8}$$

Here $V(T) \in \mathscr{L}(H:H)$ is defined by

$$V(T)g = \int_0^T U(T - s)g\,ds \quad \text{for } g \in H. \tag{2.9}$$

Now let us assume that the operator A in (2.1) satisfies the spectrum decomposition assumption ([1, p. 75], [3, p. 171]), then there exists the orthogonal projection P such that

$$H = PH + QH, \qquad Q = I - P,$$

and PH, QH form A invariant subspaces of H. From the viewpoints of system analysis and synthesis, it is practical and interesting to take PH as a finite dimensional space. We shall assume henceforth that PH is an n dimensional subspace. Let A_P and A_Q be the restrictions of A on PH and QH, respectively. We also denote by $U_P(t)$ and $U_Q(t)$ the strongly continuous semigroups on PH and QH generated by A_P and A_Q. Actually A_P is bounded on PH and $U_P(t)$ is a uniformly continuous analytic semigroup.

In this case we have from (2.1)

$$\frac{dPu(t)}{dt} = A_P Pu(t) + PF(t), \qquad Pu(0) = Pu_0, \tag{2.10}$$

$$\frac{dQu(t)}{dt} = A_Q Qu(t) + QF(t), \qquad Qu(0) = Qu_0, \tag{2.11}$$

$$u(t) = Pu(t) + Qu(t),$$

$$\|u(t)\| = \|Pu(t)\| + \|Qu(t)\|. \tag{2.12}$$

The solutions $Pu(t)$ and $Qu(t)$ are given by

$$Pu(t) = U_P(t)Pu_0 + \int_0^t U_P(t - s)PF(s)\,ds, \tag{2.13}$$

$$Qu(t) = U_Q(t)Qu_0 + \int_0^t U_Q(t - s)QF(s)\,ds. \tag{2.14}$$

The outputs z_k are written as

$$z_k = MPu(kT) + MQu(kT), \qquad k = 0, 1, \cdots. \tag{2.15}$$

Moreover (2.6) becomes

$$Pu(\overline{k+1}T) = U_P(T)Pu(kT) + PF_k, \tag{2.16}$$

$$Qu(\overline{k+1}T) = U_Q(T)Qu(kT) + QF_k, \qquad k = 0, 1, \cdots, \tag{2.17}$$

where

$$\begin{aligned} PF_k &= \int_{kT}^{(k+1)T} U_p(\overline{k+1}T - s)PF(s)\, ds, \\ QF_k &= \int_{kT}^{(k+1)T} U_Q(\overline{k+1}T - s)QF(s)\, ds, \end{aligned} \qquad k = 0, 1, \cdots. \tag{2.18}$$

We refer to the state $Pu(kT)$ governed by the n-dimensional system (2.16) as the estimated modes of the system (2.6) and the state $Qu(kT)$ governed by the infinite dimensional system (2.17) as the residual (or unestimated)modes of the system (2.6).

We assume that the semigroup $U(t)$ is uniformly bounded, that is,

$$\|U(t)\| \leqq L_1, \qquad t > 0 \tag{2.19}$$

and also assume that the semigroup $U_Q(T)$ satisfies the condition

$$r(U_Q(T)) < 1, \tag{2.20}$$

where $r(U_Q(T))$ is the spectral radius of $U_Q(T)$.

Remark 1. If A is a symmetric operator with compact resolvent and lower semibounded spectrum, then there exists a sequence $\{\lambda_m, \phi_m;\ m = 1, 2, \cdots\}$ of eigenvalues and corresponding orthonormal eigenfunctions such that for a constant c

$$c \geqq \lambda_1 \geqq \lambda_2 \cdots \geqq \lambda_m \geqq \cdots, \qquad \lim_{m\to\infty} \lambda_m = -\infty$$

and

$$A\phi_m = \lambda_m \phi_m, \qquad m = 1, 2, \cdots,$$

for every vector u in H has a unique representation

$$u = \sum_{m=1}^{\infty} u_m \phi_m, \qquad u_m = (u, \phi_m)_H.$$

The semigroup $U(t)$ is given by

$$U(t)u = \sum_{m=1}^{\infty} u_m \exp(\lambda_m t)\phi_m \quad \text{if } u \in H.$$

Then

$$\|U(t)u\|^2 = \sum_{m=1}^{\infty} u_m^2 \exp(2\lambda_m t) \leqq \exp(2\lambda_1 t)\|u\|^2.$$

If $\lambda_1 \leqq 0$, $\|U(t)u\| \leqq \|u\|$, that is, $\|U(t)\| \leqq 1 (= L_1)$, $t \geqq 0$.

Next the orthogonal projections P and Q are defined by

$$Pu = \sum_{m=1}^{n} u_m \phi_m \quad \text{if } u \in H, \qquad Qu = \sum_{m=n+1}^{\infty} u_m \phi_m \quad \text{if } u \in H.$$

Then $U_Q(T)$ is given by

$$U_Q(T)Qu = \sum_{m=n+1}^{\infty} (Qu, \phi_m) \exp(\lambda_m T)\phi_m \quad \text{if } u \in H.$$

The eigenvalues of $U_Q(T)$ are $\exp(\lambda_{n+1}T)$, $\exp(\lambda_{n+2}T), \cdots$. Thus if $\lambda_{n+1} < 0$, $r(U_Q(T)) < 1$. In this case we have also $\|U_Q(T)\| < 1$.

Under the assumption (2.19) we can show the following lemmas.

LEMMA 1. *If F belongs to $L_1(0, \infty; H)$, then $\{F_k\}_{k=0}^{\infty}$ belongs to $l_1(H)$.*

Proof. From (2.5) we obtain

$$\begin{aligned}
\|\{F_k\}\|_{l_1(H)} = \sum_{k=0}^{\infty} \|F_k\| &= \sum_{k=0}^{\infty} \left\| \int_{kT}^{(k+1)T} U(\overline{k+1}T - s)F(s)\, ds \right\| \\
&\leqq \sum_{k=0}^{\infty} \int_{kT}^{(k+1)T} \|U(\overline{k+1}T - s)\| \cdot \|F(s)\|\, ds \\
&\leqq L_1 \sum_{k=0}^{\infty} \int_{kT}^{(k+1)T} \|F(s)\|\, ds \\
&= L_1 \int_0^{\infty} \|F(s)\|\, ds < \infty.
\end{aligned}$$

LEMMA 2. *If F belongs to $L_1(0, \infty; H)$, then $\{F_k\}_{k=0}^{\infty}$ belongs to $l_2(H)$.*

Proof. From (2.5) we obtain

$$\begin{aligned}
\|\{F_k\}\|_{l_2(H)}^2 = \sum_{k=0}^{\infty} \|F_k\|^2 &= \sum_{k=0}^{\infty} \left\| \int_{kT}^{(k+1)T} U(\overline{k+1}T - s)F(s)\, ds \right\|^2 \\
&\leqq \sum_{k=0}^{\infty} \left(\int_{kT}^{(k+1)T} \|U(\overline{k+1}T - s)\| \cdot \|F(s)\|\, ds \right)^2 \\
&\leqq L_1^2 \sum_{k=0}^{\infty} \left(\int_{kT}^{(k+1)T} \|F(s)\|\, ds \right)^2 \\
&\leqq L_1^2 \int_0^{\infty} \int_0^{\infty} \|F(s)\| \cdot \|F(t)\|\, ds\, dt \\
&= L_1^2 \left(\int_0^{\infty} \|F(s)\|\, ds \right)^2 < \infty.
\end{aligned}$$

Moreover under the assumption (2.20), we can obtain the following lemmas.

LEMMA 3. *If $\{F_k\}_{k=0}^{\infty}$ belongs to $l_1(H)$, then $\{Qu(kT)\}_{k=0}^{\infty}$ also belongs to $l_1(H)$.*

Proof. From (2.17) we get

$$Qu(kT) = U_Q(T)^k Qu_0 + \sum_{i=0}^{k-1} U_Q(T)^{k-i-1} PF_i. \tag{2.21}$$

By means of (2.20) we have

$$\sum_{k=0}^{\infty} \|U_Q(T)^k Qu_0\| \leqq \sum_{K=0}^{\infty} \|U_Q(T)^k\| \cdot \|Qu_0\| = \|Qu_0\| \sum_{k=0}^{\infty} \|U_Q(T)^k\| < \infty,$$

which means that $\{U_Q(T)^k Qu_0\}_{k=0}^{\infty} \in l_1(H)$. Further we get

$$\sum_{k=1}^{\infty} \left\| \sum_{i=0}^{k-1} U_Q(T)^{k-i-1} PF_i \right\| \leqq \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} \|U_Q(T)^{k-i-1}\| \cdot \|PF_i\|$$

$$= \sum_{k=1}^{\infty} \sum_{i=0}^{k-1} \|U_Q(T)^i\| \cdot \|PF_{k-i-1}\|$$

(and reversing the order of summation)

$$= \sum_{i=0}^{\infty} \|U_Q(T)^i\| \sum_{k=i+1}^{\infty} \|PF_{k-i-1}\|$$

$$\leqq \sum_{i=0}^{\infty} \|U_Q(T)^i\| \sum_{k=0}^{\infty} \|F_k\| < \infty,$$

which implies that $\{\sum_{i=0}^{k-1} U_Q(T)^{k-i-1} PF_i\}_{k=1}^{\infty} \in l_1(H)$. Consequently we obtain $\{Qu(kT)\}_{k=0}^{\infty} \in l_1(H)$.

LEMMA 4. *If $\{F_k\}_{k=0}^{\infty}$ belongs to $l_2(H)$, then $\{Qu(kT)\}_{k=0}^{\infty}$ also belongs to $l_2(H)$.*

Proof. From (2.21) we get

$$\|Qu(kT)\| \leqq \|PF_{k-1}\| + \|U_Q(T)\| \cdot \|PF_{k-2}\| + \cdots$$
$$+ \|U_Q(T)^{k-1}\| \cdot \|PF_0\| + \|U_Q(T)^k\| \cdot \|Qu_0\|,$$

where $F_{-1} = 0$. Hence we may write

$$(\|Qu(0)\|, \|Qu(T)\|, \cdots) \leqq (1, \|U_Q(T)\|, \|U_Q(T)^2\| \cdots)$$
$$* (\|Qu_0\|, \|PF_0\|, \|PF_1\|, \cdots)$$

where $*$ denotes the convolution transform. Using the Young's inequality ([2, p. 951]) we have

$$\left(\sum_{k=0}^{\infty} \|Qu(kT)\|^2 \right) \leqq \left(\sum_{k=0}^{\infty} \|U_Q(T)^k\| \right)^2 \left(\|Qu_0\|^2 + \sum_{k=0}^{\infty} \|PF_k\|^2 \right)$$

$$\leqq \left(\sum_{k=0}^{\infty} \|U_Q(T)^k\| \right)^2 \left(\|u_0\|^2 + \sum_{k=0}^{\infty} \|F_k\|^2 \right) < \infty,$$

which implies that $\{Qu(kT)\}_{k=0}^{\infty}$ belongs to $l_2(H)$.

3. Finite dimensional discrete-time observers. In this section we construct finite dimensional discrete-time observers to approximate the estimated modes $Pu(kT)$ of the system (2.1) with unknown input $F(t)$. For the system (2.16), we introduce a discrete-time state observer on the n-dimensional subspace PH ([7, p. 525], [8, p. 308])

$$w(k+1) = Dw(k) + Bz_k, \qquad k = 0, 1, \cdots, w(0) = 0, \tag{3.1}$$

where D is a bounded linear operator on PH and B is a bounded linear operator from E to PH. We assume that D satisfies the identity observer condition

$$D = U_P(T) - BM. \tag{3.2}$$

From (2.16), (3.1) and (3.2) we have

$$Pu(\overline{k+1}T) - w(k+1) = D(Pu(kT) - w(k)) + PF_k - BMQu(kT), \tag{3.3}$$

$$Pu(0) - w(0) = Pu_0. \tag{3.4}$$

The solution of this error equation is given by

$$(3.5)\quad Pu(kT)-w(k)=D^kPu_0+\sum_{i=0}^{k-1}D^{k-i-1}[PF_i-BMQu(iT)],\qquad k=1,2,\cdots.$$

If the estimated modes Pu are observable, there exists a unique output feedback operator B such that all the eigenvalues of D can be arbitrarily assigned. If all the eigenvalues of D are less than 1 in their absolute values, the error system (3.3) is asymptotically stable. Moreover, in the case of discrete-time systems, it is interesting and practical to make all the eigenvalues of D be 0 from the viewpoint of finite time setting. In this special case, (3.5) becomes for $k=n, n+1, \cdots$

$$(3.6)\quad Pu(kT)-w(k)=\sum_{i=k-n}^{k-1}D^{k-i-1}[PF_i-BMQu(iT)].$$

If $r(D)<1$ and $\{F_k\}_{k=0}^{\infty}\in l_1(H)$ (or $l_2(H)$), it follows from Lemma 3 (or Lemma 4) that $\{Qu(kT)\}_{k=0}^{\infty}\in l_1(H)$ (or $l_2(H)$). Moreover using the same arguments as in the proof of Lemma 3 (or Lemma 4) we can show from (3.5) that $\{Pu(kT)-w(k)\}_{k=0}^{\infty}\in l_1(H)$ (or $l_2(H)$). Therefore we have

$$(3.7)\quad \lim_{k\to\infty}\|Pu(kT)-w(k)\|=0.$$

Now if we estimate the state $u(kT)$ of the system (2.1) by

$$(3.8)\quad \hat{u}(kT)=w(k),$$

the estimation error $u(kT)-\hat{u}(kT)$ is given by

$$u(kT)-\hat{u}(kT)=(Pu(kT)-\hat{u}(kT))+Qu(kT).$$

Then

$$\begin{aligned}\|u(kT)-\hat{u}(kT)\|&\leqq\|Pu(kT)-\hat{u}(kT)\|+\|Qu(kT)\|\\&\leqq\|Pu(kT)-w(k)\|+\|Qu(kT)\|.\end{aligned}$$

Since from Lemma 3 (or Lemma 4), $\{Qu(kT)\}_{k=0}^{\infty}\in l_1(H)$ (or $l_2(H)$), $\lim_{k\to\infty}\|Qu(kT)\|=0$. From this and (3.7), we obtain the following result.

THEOREM 1. *If the unknown input F belongs to $L_1(0,\infty;H)$ and $r(D)<1$,*

$$\lim_{k\to\infty}\|u(kT)-\hat{u}(kT)\|=0.$$

This theorem implies that we can construct the finite dimensional discrete-time observer (3.1) for the system (2.1) with the unknown input $F(t)$ in the case of discrete-time observation. We can estimate the state $u(kT)$ at the kth sampling time of the system (2.1) by $w(k)$ for some large k.

Moreover in order to obtain explicit error estimates, let us assume that

$$(3.9)\quad \|U_Q(T)\|\leqq q<1,$$

$$(3.10)\quad \|D^k\|\leqq L_2a^k,\qquad a<1,$$

$$(3.11)\quad \|F(t)\|\leqq L_3\exp(-\beta t),\qquad \beta>0,\quad t>0,$$

where $\|D^k\|$ is the Euclidean norm of the matrix D^k and L_2, L_3, a, q and β are constants.

Remark 2. The assumption (3.9) was discussed in Remark 1. The assumption (3.10) says that all the eigenvalues of D are less than 1 in their absolute values. Let

us illustrate this by an example and consider the case of D having n distinct eigenvalues $\gamma_1, \cdots, \gamma_n$ and corresponding orthonormal eigenvectors $\psi_1, \cdots, \psi_n$. Then for any n-dimensional vector w

$$D^k w = \gamma_1^k w_1 \psi_1 + \cdots + \gamma_n^k w_n \psi_n, \qquad w_i = (w, \psi_i)_{R^n}, \quad i = 1, \cdots, n$$

holds. From this we have

$$\begin{aligned} \|D^k w\| &\leqq |\gamma_1^k w_1| \cdot \|\psi_1\| + \cdots + |\gamma_n^k w_n| \cdot \|\psi_n\| \\ &= |\gamma_1^k w_1| + \cdots + |\gamma_n^k w_n| \\ &\leqq \Big(\max_i |\gamma_i|\Big)^k \sum_{i=1}^n |w_i| \leqq \sqrt{n}\Big(\max_i |\gamma_i|\Big)^k \Big(\sum_{i=1}^n w_i^2\Big)^{1/2}. \end{aligned}$$

Thus if we take $L_2 = \sqrt{n}$ and $a = \max_i |\gamma_i|$, then the assumption (3.10) holds.

First let us estimate $\|F_k\|$. From (2.5) we have

$$\begin{aligned} \|F_k\| &= \Big\| \int_{kT}^{(k+1)T} U(\overline{k+1}T - s) F(s)\, ds \Big\| \\ &\leqq \int_{kT}^{(k+1)T} \|U(\overline{k+1}T - s)\| \cdot \|F(s)\|\, ds \\ &\leqq L_1 L_3 \int_{kT}^{(k+1)T} \exp(-\beta s)\, ds \\ &= \frac{L_1 L_3}{\beta} (1 - \exp(-\beta T)) \exp(-k\beta T) = L_4 b^k, \end{aligned} \tag{3.12}$$

where

$$L_4 = \frac{L_1 L_3}{\beta}(1-b), \qquad b = \exp(-\beta T).$$

For example, when $a < q < b$, we can obtain an explicit error estimate as follows. From (2.21) we get

$$\begin{aligned} \|Qu(kT)\| &\leqq \|U_Q(kT)\| \cdot \|Qu_0\| + \sum_{i=0}^{k-1} \|U_Q^{k-i-1}\| \cdot \|PF_i\| \\ &\leqq q^k \|Qu_0\| + \sum_{i=0}^{k-1} q^{k-i-1} L_4 b^i \\ &\leqq q^k \|u_0\| + L_4 b^k \frac{b^k - q^k}{b-q} \leqq K_1 b^k, \end{aligned} \tag{3.13}$$

where

$$K_1 = \|u_0\| + L_4 \frac{1}{b-q}.$$

Moreover we have

$$\begin{aligned} &\|Pu(kT) - w(k)\| \\ &\leqq \|D^k P u_0\| + \Big\| \sum_{i=0}^{k-1} D^{k-i-1} [PF_i - BMQu(iT)] \Big\| \\ &\leqq L_2 a^k \|Pu_0\| + \sum_{i=0}^{k-1} L_2 a^{k-i-1} (\|PF_i\| + \|BM\| \cdot \|Qu(iT)\|) \\ &\leqq L_2 a^k \|u_0\| + L_2 \sum_{i=0}^{k-1} a^{k-i-1} (L_4 b^i + \|BM\| K_1 b^i) \end{aligned} \tag{3.14}$$

$$= L_2\|u_0\|a^k + L_2(L_4 + \|BM\|K_1)\sum_{i=0}^{k-1} a^{k-i-1}b^i$$

$$= L_2\|u_0\|a^k + L_2(L_4 + \|BM\|K_1)\frac{b^k - a^k}{b-a} \leqq K_2 b^k,$$

where

$$K_2 = L_2\|u_0\| + \frac{L_2(L_4 + \|BM\|K_1)}{b-a}.$$

In the special case of $r(D) = 0$, for $k \geqq n$ we can estimate $\|Pu(kT) - w(k)\|$ as follows

$$\begin{aligned}\|Pu(kT) - w(k)\| &\leqq \sum_{i=k-n}^{k-1} \|D^{k-i-1}\| \cdot \|PF_i - BMQu(iT)\| \\ &\leqq L_2 \sum_{i=k-n}^{k-1} a^{k-i-1}(L_4 b^i + \|BM\|K_1 b^i) \\ &= L_2(L_4 + \|BM\|K_1)\sum_{i=k-n}^{k-1} a^{k-i-1}b^i \\ &= \frac{L_2(L_4 + \|BM\|K_1)}{b-a}\left(1 - \left(\frac{a}{b}\right)^n\right) b^k = K_2 b^k,\end{aligned} \tag{3.15}$$

where

$$K_2 = \frac{L_2(L_4 + \|BM\|K_1)}{b-a}\left(1 - \left(\frac{a}{b}\right)^n\right).$$

Consequently we have an explicit error estimate

$$\begin{aligned}\|u(kT) - \hat{u}(kT)\| &\leqq \|Pu(kT) - w(k)\| + \|Qu(kT)\| \\ &\leqq (K_1 + K_2)b^k.\end{aligned} \tag{3.16}$$

We can similarly obtain explicit error estimates when $0 < a, b, q < 1$.

THEOREM 2. *When $U_Q(T)$, D and $F(t)$ satisfy* (3.9), (3.10) *and* (3.11), *respectively, the explicit error estimate is given by*

$$\|u(kT) - \hat{u}(kT)\| \leqq \text{Const. } p^k, \qquad k = 1, 2, \cdots, \tag{3.17}$$

where $p = \max(a, b, q)$.

Before ending this section we shall give a simple example to illustrate the presented theory.

Example 1. Let us consider the system

$$\frac{\partial u(t,x)}{\partial t} = 0.1\frac{\partial^2 u(t,x)}{\partial x^2} - 0.1u(t,x) + r(t)g(x), \qquad x \in (0,1), \quad t > 0,$$

$$\frac{\partial u(t,0)}{\partial x} = \frac{\partial u(t,1)}{\partial x} = 0,$$

$$u(0,x) = u_0(x),$$

where the input $r(t)$, the input distribution function $g(x)$ and the initial state $u_0(x)$ are unknown.

For this example, we take $H = L_2(0, 1)$ and $Au = 0.1\Delta u - 0.1u$. Here Δ is the Laplacian with Neumann conditions at $x = 0, 1$. The eigenfunctions are

$$\phi_1(x) = 1, \qquad \phi_m(x) = \sqrt{2}\cos(m-1)\pi x, \qquad m = 2, 3, \cdots,$$

which constitute an orthonormal basis for $L_2(0, 1)$ and the eigenvalues $\lambda_m = -0.1(m-1)^2\pi^2 - 0.1$, $m = 1, 2, \cdots$. The output of the system is given by

$$z_k = \int_0^1 h(x)u(kT, x)\,dx = \sum_{m=1}^{\infty} h_m u_m(kT), \qquad h \in L_2(0, 1), \quad k = 0, 1, \cdots,$$

where $h_m = \int_0^1 h(x)\phi_m(x)\,dx$, $u_m(kT) = \int_0^1 u(kT, x)\phi_m(x)\,dx$, $m = 1, 2, \cdots$. If $h_m \neq 0$, $m = 1, \cdots, n$, the first n modes are observable [4].

Using eigenfunctions expansion, we obtain the discrete-time systems

$$u_m(\overline{k+1}T) = \exp(\lambda_m T)u_m(kT) + g_m r_{mk}, \qquad m = 1, \cdots, n,$$

corresponding to (2.16) and

$$u_m(\overline{k+1}T) = \exp(\lambda_m T)u_m(kT) + g_m r_{mk}, \qquad m = n+1, n+2, \cdots,$$

corresponding to (2.17), where $g_m = \int_0^1 g(x)\phi_m(x)\,dx$, $m = 1, 2, \cdots$, and $r_{mk} = \int_{kT}^{(k+1)T} \exp(\lambda_m(\overline{k+1}T - s))r(s)\,ds$, $m = 1, 2, \cdots$, $k = 0, 1, \cdots$. From Remark 1 we find $L_1 = 1$ in (2.19).

Taking $n = 5$, we construct a 5-dimensional discrete-time observer

$$w(k+1) = Dw(k) + Bz_k, \qquad w(0) = 0, \quad k = 0, 1, \cdots,$$

where D is a 5×5 matrix and B is a 5-dimensional feedback vector which satisfies the identity observer condition

$$D = \begin{bmatrix} \exp(\lambda_1 T) & & 0 \\ & \ddots & \\ 0 & & \exp(\lambda_5 T) \end{bmatrix} - \begin{bmatrix} B_1 \\ \vdots \\ B_5 \end{bmatrix} [h_1 \cdots h_5].$$

If the first 5 modes $u_1, \cdots, u_5$ are observable, we can determine uniquely a feedback vector B such that the matrix D has specified eigenvalues. If all the eigenvalues of D are less than 1 in their absolute values, the error system

$$\begin{bmatrix} u_1(\overline{k+1}T) \\ \vdots \\ u_5(\overline{k+1}T) \end{bmatrix} - w(k+1) = D\left(\begin{bmatrix} u_1(kT) \\ \vdots \\ u_5(kT) \end{bmatrix} - w(k)\right) + \begin{bmatrix} g_1 r_{1k} \\ \vdots \\ g_5 r_{5k} \end{bmatrix} - \begin{bmatrix} B_1 \\ \vdots \\ B_5 \end{bmatrix} \sum_{m=6}^{\infty} h_m u_m(kT)$$

is asymptotically stable and then

$$\lim_{k\to\infty} \left(\begin{bmatrix} u_1(kT) \\ \vdots \\ u_5(kT) \end{bmatrix} - w(k)\right) = 0.$$

In the case of $r(t) = \exp(-t)$ and D having eigenvalues 0.02, 0.01, 0, −0.01, −0.02 we have $q = \exp(\lambda_6 T) = \exp(-(2.5\pi^2 + 0.1)T)$, $L_2 = \sqrt{5}$, $a = 0.02$, $L_3 = (\int_0^1 g^2(x)\,dx)^{1/2}$, $\beta = 1$ and $b = \exp(-T)$.

4. Determination of unknown input distribution functions. In this section we consider the problem of determining unknown input distribution functions for the

systems with discrete-time input–output data such that

$$\frac{du(t)}{dt} = Au(t) + g \sum_{j=0}^{[t]} f_j (Y(t-jT) - Y(t-\overline{j+1}T)), \tag{4.1}$$

$$\begin{aligned} &u(0) = u_0, \\ &z_k = Mu(kT), \qquad k = 0, 1, \cdots, \end{aligned} \tag{4.2}$$

where $g \in H$ is the unknown input distribution function to be determined. The input function $\{f_j\}_{j=0}^{\infty}$ belongs to l_2 and is known. Though the initial state u_0 is unknown, we can estimate the system state $u(NT)$ at time NT by constructing a finite dimensional discrete-time observer such as (3.1), where we can especially choose $f_j = 0$, $j = 0, 1, \cdots, N-1$.

We denote by $J(g)$ a functional which measures the distance between the observation $z = \{z_k\}_{N+1}^{\infty}$ and the output $\{M\bar{u}(kT)\}_{N+1}^{\infty}$ computed for each g from the system (4.1) with the initial state $u(NT) = \hat{u}(NT)$ estimated by (3.8). We take the functional $J(g)$ to be

$$J(g) = \sum_{k=N+1}^{\infty} \|z_k - M\bar{u}(kT)\|_E^2. \tag{4.3}$$

Here

$$\begin{aligned} \bar{u}(kT) &= U_p(kT-NT)\hat{u}(NT) + \int_{NT}^{kT} U(kT-s) \left(g \sum_{j=N}^{[s]} f_j [Y(s-jT) - Y(s-\overline{j+1}T)] \right) ds \\ &= U_p(T)^{k-N}\hat{u}(NT) + \sum_{j=N}^{k-1} \int_{jT}^{(j+1)T} U(kT-s) g \, ds \, f_j \end{aligned}$$

(using $V(T)$ defined by (2.9))

$$= U_p(T)^{k-N}\hat{u}(NT) + \sum_{j=N}^{k-1} U(T)^{k-j-1} V(T) g f_j \quad \text{for } k \geqq N+1. \tag{4.4}$$

Moreover we can write

$$z_k = M\bar{u}(kT) + M(U(T)^{k-N}u(NT) - U_p(T)^{k-N}\hat{u}(NT)) \quad \text{for } k \geqq N+1. \tag{4.5}$$

The second term on the right-hand side seems to be an observation error. The problem of unknown function determination can be formulated as that of minimizing $J(g)$ with respect to g under the constraint (4.1). If the system (4.1) and (4.2) is identifiable, the solution of this problem is unique as shown later.

First we investigate the relationship between observability and identifiability of the system (4.1) and (4.2). The system (4.1) and (4.2) is said to be observable (N-step observable) if the knowledge of the input $\{gf_j\}$ and the observation $\{z_k\}_1^{\infty}(\{z_k\}_1^N)$ implies that the initial state $u(0)$ is uniquely determined. The system is said to be identifiable (N-step identifiable) if the knowledge of the initial state $u(0)$, $\{f_j\}$ and $\{z_k\}_1^{\infty}(\{z_k\}_1^N)$ implies that g is uniquely determined.

From the definitions and the linearity, the system (4.1) and (4.2) is observable (N-step observable) if and only if $MU(T)^k u_0 = 0$, $k = 1, 2, \cdots (k = 1, \cdots, N)$ implies that $u_0 = 0$. The system (4.1) and (4.2) is identifiable (N-step identifiable) if and only if

$$M \sum_{j=0}^{k-1} U(T)^{k-j-1} V(T) g f_j = 0, \qquad k = 1, 2, \cdots (k = 1, \cdots, N),$$

implies that $g = 0$.

Let us consider whether or not the system (4.1) and (4.2) can be N-step observable. Define an operator W by

$$Wu_0 = \{MU(T)^k u_0\}_{k=1}^N, \qquad u_0 \in H.$$

W is the bounded linear operator from H to $l_2(\sigma_N; E)$, where $\sigma_N = \{1, \cdots, N\}$. If the observation space E is finite dimensional, the space $l_2(\sigma_N; E)$ is also finite dimensional. W is the bounded linear operator from the infinite dimensional space to the finite dimensional space. The operator W can never be injective. Thus the condition for N-step observability is never met for any finite N. In the similar way it is shown that the system (4.1) and (4.2) is never N-step identifiable for any finite N.

Remark 3. Since the system (4.1) and (4.2) is never N-step identifiable for any finite N, the functional $J(g)$ in (4.3) must be defined on an infinite time interval. In the continuous-time case, $J(g)$ can be defined on a finite-time interval [6]. This is an essential difference between the problem of continuous-time identification of parameters and the problem of discrete-time identification of parameters.

Moreover we can show the following theorem.

THEOREM 3. *Let* $f = \{f_j, j = 0, 1, \cdots\} \not\equiv 0$ *and the nullspace of* $V(T)$ *be* $\{0\}$. *Then the system* (4.1) *and* (4.2) *is identifiable if and only if the system* (4.1) *and* (4.2) *is observable.*

Proof. To prove the sufficiency, we suppose that

$$M \sum_{j=0}^{k-1} U(T)^{k-j-1} V(T) g f_j = 0, \qquad k = 1, 2, \cdots. \tag{4.6}$$

Applying the z-transformation to this equation, we have

$$MF(z)Z[U(T)^j V(T)g] = 0$$

for any $|z| > \max(\xi, \eta)$, where ξ and η are the convergence coordinates of $F(z) = Z[f_j]$ and $Z[U(T)^j V(T)g]$, respectively. From the assumption we get $F(z) \neq 0$ for almost every z such that $|z| > \xi$. Thus we obtain

$$Z[U(T)^j V(T)g] = 0.$$

Since the inverse z-transform of 0 is 0, we get

$$U(T)^j V(T)g = 0, \qquad j = 1, 2, \cdots.$$

From the observability of the system, it follows that $V(T)g = 0$. Since the nullspace of $V(T)$ is $\{0\}$, $g = 0$. Thus the system is identifiable.

To prove the necessity, suppose that the system (4.1) and (4.2) is not observable. Then (4.6) has a nonzero solution g in H. Therefore the system is not identifiable.

Remark 4. Theorem 3 is given in [5] in the case of A being a symmetric operator with compact resolvent and lower semibounded spectrum. In this case we can show that the nullspace of $V(T)$ is $\{0\}$. From Remark 1 we have

$$U(t)g = \sum_{m=1}^{\infty} g_m \exp(\lambda_m t)\phi_m \quad \text{if } g \in H.$$

Moreover we get

$$V(T)g = \int_0^T U(T-t)g\, dt = \sum_{m=1}^{\infty} g_m \frac{\exp(\lambda_m T) - 1}{\lambda_m} \phi_m \quad \text{if } g \in H,$$

where for $\lambda_m = 0$ $(\exp(\lambda_m T) - 1)/\lambda_m$ must be replaced by T. In order to show that the nullspace of $V(T)$ is $\{0\}$ suppose that

$$\sum_{m=1}^{\infty} g_m \frac{\exp(\lambda_m T) - 1}{\lambda_m} \phi_m = 0.$$

Taking the inner product with ϕ_m and using the orthonormality of $\{\phi_m\}$, we obtain

$$\frac{\exp(\lambda_m T) - 1}{\lambda_m} g_m = 0, \qquad m = 1, 2, \cdots$$

which implies $g_m = 0$, $m = 1, 2, \cdots$, that is, $g = 0$. Thus the nullspace of $V(T)$ is $\{0\}$.

Now let us define a linear operator $S: H \to l_2(\sigma; E)$ $(\sigma = \{N+1, N+2, \cdots\})$ by

$$Sg = \left\{ M \sum_{j=N}^{k-1} U(T)^{k-j-1} V(T) g f_j \right\}_{k \in \sigma}, \quad \text{if } g \in H. \tag{4.7}$$

Then we have

LEMMA 5. *If $r(U(T)) < 1$, then S is bounded.*

Proof. From (4.7) we have

$$\begin{aligned} \|(Sg)_k\| &\leqq \|MV(T)gf_{k-1}\| + \|MU(T)V(T)gf_{k-2}\| + \cdots + \|MU(T)^{k-N-1}V(T)gf_N\| \\ &\leqq \|M\| \cdot \|V(T)g\| \cdot |f_{k-1}| + \|M\| \cdot \|V(T)g\| \cdot \|U(T)\| \cdot |f_{k-2}| + \cdots \\ &\quad + \|M\| \cdot \|V(T)g\| \cdot \|U(T)^{k-N-1}\| \cdot |f_N|. \end{aligned}$$

Hence we may write

$$\begin{aligned} (\|(Sg)_{N+1}\|, \|(Sg)_{N+2}\|, \cdots) &\leqq (1, \|U(T)\|, \|U(T)^2\|, \cdots) \\ &\quad * (\|M\| \cdot \|V(T)g\| \cdot |f_N|, \|M\| \cdot \|V(T)g\| \cdot |f_{N+1}|, \cdots) \end{aligned}$$

where $*$ denotes the convolution transform. Using Young's inequality we have

$$\begin{aligned} \|Sg\|^2_{l_2(\sigma;E)} = \sum_{k=N+1}^{\infty} \|(Sg)_k\|^2 &\leqq \left(\sum_{k=0}^{\infty} \|U(T)^k\| \right)^2 \left(\|M\|^2 \cdot \|V(T)g\|^2 \sum_{k=N}^{\infty} f_k^2 \right) \\ &\leqq \left(\sum_{k=0}^{\infty} \|U(T)^k\| \right)^2 \left(\|M\|^2 \cdot \|V(T)\|^2 \sum_{k=N}^{\infty} f_k^2 \right) \|g\|^2. \end{aligned}$$

Since

$$\sum_{k=0}^{\infty} \|U(T)^k\| < +\infty, \quad \sum_{k=0}^{\infty} f_k^2 < +\infty \quad \text{and} \quad \|V(T)\| < +\infty,$$

S is bounded.

The system (4.1) and (4.2) is identifiable with $f_j = 0$, $0 \leqq j \leqq N-1$ if and only if the nullspace of S is $\{0\}$, that is, S^*S is positive on H. Here $(\cdot)^*$ denotes the adjoint operator of an operator $(\cdot)$. Henceforth we assume that the system (4.1) and (4.2) is identifiable with $f_j = 0$, $0 \leqq j \leqq N-1$.

Back to the problem of minimizing the functional $J(g)$, we get

$$J(g) = \|y - Sg\|^2_{l_2(\sigma;E)},$$

where $y = \{y_k\}_{k \in \sigma}$, $y_k = z_k - MU_P(T)^{k-N}\hat{u}(NT)$. Since the operator S is continuous, the functional $J(g)$ is Fréchet differentiable and convex. Hence the necessary condition for optimality is

$$J'(g)h = 0 \quad \text{for any } h \in H.$$

From this the minimizing solution g_0 of $J(g)$ must satisfy

$$S^*Sg = S^*y. \tag{4.8}$$

Since the system (4.1) and (4.2) is identifiable with $f_j = 0$, $0 \leqq j \leqq N-1$, the optimal solution g_0 is uniquely determined by

$$g_0 = (S^*S)^{-1}S^*y. \tag{4.9}$$

However a positive self-adjoint operator S^*S on H has its bounded inverse $(S^*S)^{-1}$ if and only if it is positive definite. Unless $S^*y \in \mathrm{Dom}\,((S^*S)^{-1})$, g_0 will not belong to H.

Remark 5. Let us consider the case that the operator A is symmetric and has compact resolvent and lower semibounded spectrum. From Remark 1 we have

$$U(t)g = \sum_{n=1}^{\infty} g_n \exp(\lambda_n t)\phi_n \quad \text{if } g \in H,$$

where $g_n = (g, \phi_n)_H$, $n = 1, 2, \cdots$. In this case we get

$$V(T)g = \int_0^T U(T-t)g\,dt = \sum_{n=1}^{\infty} g_n \frac{\exp(\lambda_n T)-1}{\lambda_n}\phi_n \quad \text{if } g \in H.$$

Moreover we obtain

$$\begin{aligned}(Sg)_k &= M\sum_{j=N}^{k-1} U(T)^{k-j-1}V(T)gf_j \\ &= M\sum_{j=N}^{k-1}\sum_{n=1}^{\infty} g_n \frac{\exp(\lambda_n T)-1}{\lambda_n}\exp((k-j-1)\lambda_n T)\phi_n f_j, \\ &\qquad\qquad k = N+1, N+2, \cdots.\end{aligned}$$

Let

$$s = \inf_{g \in H} \frac{(S^*Sg, g)_H}{(g, g)_H}$$

and then show $s = 0$. Taking $g = \phi_n$, we have

$$\frac{(S^*S\phi_n, \phi_n)}{(\phi_n, \phi_n)} = \sum_{k=N+1}^{\infty} \|(S\phi_n)_k\|^2$$

since $\|\phi_n\| = 1$. For some large n, $\lambda_n < 0$. Then we find

$$\begin{aligned}\|(S\phi_n)_k\| \leqq \frac{\exp(\lambda_n T)-1}{\lambda_n}\|M\phi_n\|(|f_{k-1}| + \exp(\lambda_n T)|f_{k-2}| + \cdots \\ + \exp((k-N-1)\lambda_n T)|f_N|).\end{aligned}$$

Using the Young's inequality we have

$$\sum_{k=N+1}^{\infty} \|(S\phi_n)_k\|^2 \leqq \left(\frac{\exp(\lambda_n T)-1}{\lambda_n}\right)^2 \|M\phi_n\|^2 \left(\sum_{k=0}^{\infty} \exp(k\lambda_n T)\right)^2 \sum_{k=N}^{\infty} f_k^2.$$

The right-hand side tends to 0 as $n \to \infty$. Thus S^*S is not positive definite.

Thus even if the system (4.1) and (4.2) is identifiable with $f_j = 0$, $0 \leqq j \leqq N-1$, the inverse $(S^*S)^{-1}$ is not bounded in general. The solution g_0 does not necessarily depend continuously on y and g_0 does not belong to H unless $S^*y \in \mathrm{Dom}\,((S^*S)^{-1})$.

From a numerical calculations point of view, the algorithms which seek g_0 from (4.9) are ill-conditioned in general.

It is not practical to use (4.9) to seek the input distribution functions from the measurement data. Thus we must consider a feasible approximation method which gives a constructive procedure to present approximately the true input distribution functions. We apply a feasible approximation method by regularization [9], [10].

Let us introduce a regularizing functional $J_\varepsilon(g)$ corresponding to $J(g)$

$$J_\varepsilon(g)=J(g)+\varepsilon\|g\|^2, \tag{4.10}$$

where ε is a positive regularizing parameter. Then the unique minimizing solution g_ε of $J_\varepsilon(g)$ is determined by

$$(S^*S+\varepsilon I)g_\varepsilon = S^*y, \tag{4.11}$$

that is,

$$g_\varepsilon = (S^*S+\varepsilon I)^{-1}S^*y. \tag{4.12}$$

Since the operator $G_\varepsilon = S^*S+\varepsilon I$ is positive-definite on H, its inverse G_ε^{-1} is bounded. The g_ε belongs to H and depends continuously on y (or z).

Here if we define $m=\{m_k\}_{k\in\sigma}$ by

$$m_k = M(U(T)^{k-N}u(NT)-U_p(T)^{k-N}\hat{u}(NT)) \quad \text{for } k\geqq N,$$

then we find from (4.4) and (4.5)

$$\begin{aligned} y_k &= z_k - MU_P(T)^{k-N}\hat{u}(NT) = M\bar{u}(kT)+m_k-MU_P(T)^{k-N}\hat{u}(NT)\\ &= (Sg^*)_k+m_k \quad \text{for } k\geqq N, \end{aligned}$$

where $g^*\in H$ is the true input distribution function. Thus m seems to be observation errors and the solutions g_0 and g_ε depend on m.

We can show the following theorem.

THEOREM 4. *Suppose that the system* (4.1) *and* (4.2) *is identifiable with* $f_j=0$, $0\leqq j\leqq N-1$. *Then we have*

$$\lim_{\varepsilon,\delta\to 0}\|g_\varepsilon - g^*\|_H = 0,$$

provided that $\delta/\sqrt{\varepsilon}\to 0$ *as* $\varepsilon\to 0$, *where* m *satisfies* $\|m\|^2_{l_2(\sigma;E)}\leqq\delta^2$.

Proof. We first show that

(i) $\lim_{\varepsilon\to 0}\|Gg_\varepsilon - Gg_0\|_H = 0$, where $G=S^*S$.

From (4.8) and (4.11) we have

$$\begin{aligned} 0 &= (G(Gg_0-S^*y), h)\\ &= (G(Gg_0-Gg_\varepsilon-\varepsilon g_\varepsilon), h)\\ &= (G(Gg_0-Gg_\varepsilon), h)-\varepsilon(Gg_\varepsilon, h) \quad \text{for any } h\in H. \end{aligned} \tag{4.13}$$

Taking $h=Gg_0-Gg_\varepsilon$ in (4.13), we have

$$(G(Gg_0-Gg_\varepsilon), Gg_0-Gg_\varepsilon)-\varepsilon(Gg_\varepsilon, Gg_0-Gg_\varepsilon)=0.$$

From this we get

$$(Gg_\varepsilon, Gg_\varepsilon)\leqq(Gg_\varepsilon, Gg_0), \tag{4.14}$$

which implies that $\|Gg_\varepsilon\|\leqq\|Gg_0\|$. Thus from every sequence of $\varepsilon\to 0$ we can extract

a subsequence η such that $Gg_\eta \to w$ weakly in H. As $\eta \to 0$, (4.13) becomes

$$(G(Gg_0 - w), h) = 0.$$

Here taking $h = Gg_0 - w$, we obtain

$$(G(Gg_0 - w), Gg_0 - w) = 0.$$

From the positiveness of G, we have $w = Gg_0$. Here $\{Gg_\eta\}$ is an arbitrary, weakly convergent subsequence and its weak limit Gg_0 does not depend on the subsequence. Thus the extraction of a subsequence is unnecessary and $Gg_\varepsilon \to Gg_0$ weakly in H.

Moreover from (4.14) we get

$$(Gg_\varepsilon - Gg_0, Gg_\varepsilon - Gg_0) \leqq -(Gg_0, Gg_\varepsilon - Gg_0),$$

which implies that

$$\lim_{\varepsilon \to 0} \|Gg_\varepsilon - Gg_0\|_H = 0.$$

Next let us note that the true input distribution function g^* satisfies

$$Gg^* = S^*(y - m).$$

Define g_ε^* by

$$G_\varepsilon g_\varepsilon^* = S^*(y - m). \tag{4.15}$$

Then both g^* and g_ε^* belong to H. Now we have

$$\|g_\varepsilon - g^*\| \leqq \|g_\varepsilon - g_\varepsilon^*\| + \|g_\varepsilon^* - g^*\|.$$

For the seond term on the right-hand side, we can show in a similar way

$$\lim_{\varepsilon \to 0} \|g_\varepsilon^* - g^*\| = 0,$$

replacing Gg_ε, Gg_0 and y by g_ε^*, g^* and $y - m$, respectively, in the proof of (i).

Next as for the first term, we have

$$(G_\varepsilon(g_\varepsilon - g_\varepsilon^*) - S^*m, h) = 0 \quad \text{for any } h \in H$$

from (4.11) and (4.15). This implies that the element $g_\varepsilon - g_\varepsilon^*$ realizes the lower bound of the functional

$$I(g) = \|m - Sg\|_{l_2(\sigma;E)}^2 + \varepsilon \|g\|^2, \qquad \varepsilon > 0.$$

Then

$$I(g_\varepsilon - g_\varepsilon^*) \leqq I(0) = \|m\|_{l_2(\sigma;E)}^2 \leqq \delta^2.$$

From this we obtain

$$\varepsilon \|g_\varepsilon - g_\varepsilon^*\|^2 \leqq \delta^2.$$

That is,

$$\|g_\varepsilon - g_\varepsilon^*\| \leqq \frac{\delta}{\sqrt{\varepsilon}}.$$

We have proved the theorem.

Theorem 4 shows that we can use g_ε as an approximate input distribution function. From a practical point of view, we cannot use an infinite number of input–output data. We must determine the input distribution function g from a finite number of input–output data. Therefore for L sufficiently large we can only determine $g_{L\varepsilon}$

$$g_{L\varepsilon} = (S_L^* S_L + \varepsilon I)^{-1} S_L^* y_L, \tag{4.16}$$

where

$$y_{L_k} = \begin{cases} y_k, & k = N+1, \cdots, N+L, \\ 0, & k = N+L+1, N+L+2, \cdots, \end{cases}$$
$$(S_L g)_k = \begin{cases} (Sg)_k, & k = N+1, \cdots, N+L, \\ 0, & k = N+L+1, N+L+2, \cdots. \end{cases} \tag{4.17}$$

The element $g_{L\varepsilon}$ is the minimizing solution of the functional

$$J_{L\varepsilon}(g) = \|y_L - S_L g\|^2_{l_2(\sigma_L, E)} + \varepsilon \|g\|^2, \tag{4.18}$$

where $\sigma_L = \{N+1, N+2, \cdots, N+L\}$.

Before showing that for fixed $\varepsilon > 0$ $g_{L\varepsilon}$ strongly converges to g_ε as L tends to ∞, we give the following lemmas.

LEMMA 6. *If* $r(U(T)) < 1$, *then* $\lim_{L\to\infty} \|S - S_L\| = 0$.

Proof. From (4.17) we find

$$(Sg - S_L g)_k = \begin{cases} 0, & k = N+1, \cdots, N+L, \\ (Sg)_k, & k = N+L+1, N+L+2, \cdots. \end{cases}$$

Let us decompose $(Sg)_k$ such that

$$(Sg)_k = (Sg)_k^{L+} + (Sg)_k^{L-}, \qquad k = N+L+1, N+L+2, \cdots,$$

where

$$(Sg)_k^{L+} = MV(T)gf_{k-1} + \cdots + MU(T)^{k-N-L-1}V(T)gf_{N+L},$$
$$(Sg)_k^{L-} = MU(T)^{k-N-L}V(T)gf_{N+L-1} + \cdots + MU(T)^{k-N-1}V(T)gf_N.$$

For $(Sg)_k^{L+}$ we have the following estimate

$$\sum_{k=N+L+1}^{\infty} \|(Sg)_k^{L+}\|_E^2 \leqq \left(\sum_{k=0}^{\infty} \|U(T)^k\|\right)^2 \left(\|M\|^2 \|V(T)\|^2 \sum_{k=N+L}^{\infty} f_k^2\right) \|g\|^2. \tag{4.19}$$

For $(Sg)_k^{L-}$ we obtain

$$\sum_{k=N+L+1}^{\infty} \|(Sg)_k^{L-}\|_E^2 = \sum_{k=N+L+1}^{\infty} \left\| \sum_{j=N}^{N+L-1} MU(T)^{k-j-1}V(T)gf_j \right\|^2$$

$$= \sum_{m=0}^{\infty} \left\| \sum_{j=N}^{N+L-1} MU(T)^{m+N+L-j}V(T)gf_j \right\|^2 \tag{4.20}$$

$$\leqq \|M\|^2 \|V(T)\|^2 \left(\sum_{m=0}^{\infty} \|U(T)^m\|^2\right) \left(\sum_{j=N}^{N+L-1} \|U(T)^{N+L-j} f_j\|\right)^2 \|g\|^2.$$

Consequently we get from (4.19) and (4.20)

$$\|Sg - S_L g\|^2_{l_2(\sigma;E)} = \sum_{k=N+L+1}^{\infty} \|(Sg)_k\|_E^2$$

$$\leqq 2 \sum_{k=N+L+1}^{\infty} \|(Sg)_k^{L+}\|_E^2 + 2 \sum_{k=N+L+1}^{\infty} \|(Sg)_k^{L-}\|_E^2$$

$$\leqq 2\|M\|^2 \|V(T)\|^2 \Big\{ \Big(\sum_{k=0}^{\infty} \|U(T)^k\| \Big)^2 \sum_{k=N+L}^{\infty} f_k^2$$

$$+ \Big(\sum_{k=0}^{\infty} \|U(T)^k\|^2 \Big) \Big(\sum_{k=N}^{N+L-1} \|U(T)^{N+L-k} f_k\| \Big)^2 \Big\} \|g\|^2.$$

Since $r(U(T)) < 1$, $\{f_k\}_{k\in\sigma} \in l_2(\sigma)$ and $\{\sum_{k=N}^{N+L-1} \|U(T)^{N+L-k} f_k\|\}_{L=1}^{\infty} \in l_2$, we get

$$\lim_{L\to\infty} \sum_{k=N+L}^{\infty} f_k^2 = 0 \quad \text{and} \quad \lim_{L\to\infty} \sum_{k=N}^{N+L-1} \|U(T)^{N+L-k} f_k\| = 0.$$

Thus we obtain $\lim_{L\to\infty} \|S - S_L\| = 0$.

LEMMA 7. *If $r(U(T)) < 1$, the operator S is a completely continuous operator from H to $l_2(\sigma; E)$.*

Proof. The operator S_L is completely continuous for each L, since it is a bounded linear operator whose range is a finite dimensional subspace of $l_2(\sigma; E)$. From Lemma 6, a sequence $\{S_L\}$ converges uniformly to the operator S. This implies that S is also a completely continuous operator.

By virtue of Lemmas 6 and 7 we can show the following theorem.

THEOREM 5. *If $r(U(T)) < 1$, then for fixed $\varepsilon > 0$, $g_{L\varepsilon}$ converges strongly to g_ε as L tends to $+\infty$.*

Proof. From the optimality we have

$$J_{L\varepsilon}(g_{L\varepsilon}) \leqq J_{L\varepsilon}(g) \quad \text{for any } g \in H. \tag{4.21}$$

Moreover from the definitions of $J_{L\varepsilon}(g)$ and $J_\varepsilon(g)$ we get

$$J_{L\varepsilon}(g) \leqq J_\varepsilon(g) \quad \text{for any } g \in H. \tag{4.22}$$

It follows from (4.21) and (4.22) that

$$J_{L\varepsilon}(g_{L\varepsilon}) \leqq J_\varepsilon(g_\varepsilon),$$

that is,

$$\|y_L - S_L g_{L\varepsilon}\|^2 + \varepsilon \|g_{L\varepsilon}\|^2 \leqq \|y - S g_\varepsilon\|^2 + \varepsilon \|g_\varepsilon\|^2. \tag{4.23}$$

This implies that

$$\|g_{L\varepsilon}\| \leqq \text{Const.}$$

Thus from every sequence of $L \to \infty$ we can extract a subsequence η such that $g_{\eta\varepsilon} \to w$ weakly in H. Since

$$|(S_\eta g_{\eta\varepsilon}, h) - (Sw, h)| = |((S_\eta - S) g_{\eta\varepsilon}, h) + (S(g_{\eta\varepsilon} - w), h)|$$

$$\leqq \|S_\eta - S\| \cdot \|g_{\eta\varepsilon}\| \cdot \|h\| + |(g_{\eta\varepsilon} - w, S^* h)|$$

and $\lim_{\eta\to\infty}\|S-S_\eta\|=0$, $S_\eta g_{\eta\varepsilon}$ weakly converges to Sw. As $\eta\to\infty$, (4.23) becomes

$$\|y-Sw\|^2+\varepsilon\|w\|^2\leqq\varliminf_{\eta\to\infty}(\|y_\eta-S_\eta g_{\eta\varepsilon}\|^2+\varepsilon\|g_{\eta\varepsilon}\|)$$
$$\leqq\|y-Sg_\varepsilon\|^2+\varepsilon\|g_\varepsilon\|^2.$$

Since the optimal solution of $J_\varepsilon(g)$ is unique, we get $w=g_\varepsilon$. Here $\{g_{\eta\varepsilon}\}$ is an arbitrary, weakly convergent subsequence and its weak limit g_ε does not depend on the subsequence. Thus the extraction of a subsequence is unnecessary and $g_{L\varepsilon}\to g_\varepsilon$ weakly in H. Then we have

$$\varliminf_{L\to\infty}\|g_{L\varepsilon}\|\geqq\|g_\varepsilon\|. \tag{4.24}$$

Moreover from (4.21) we obtain

$$\|y_L-S_Lg_{L\varepsilon}\|^2+\varepsilon\|g_{L\varepsilon}\|^2\leqq\|y_L-S_Lg_\varepsilon\|^2+\varepsilon\|g_\varepsilon\|^2. \tag{4.25}$$

It follows from (4.24) and (4.25) that for L sufficiently large

$$\|y_L-S_Lg_\varepsilon\|\geqq\|y_L-S_Lg_{L\varepsilon}\|.$$

We have

$$\|y_L-S_Lg_\varepsilon\|-\|y_L-S_Lg_{L\varepsilon}\|\leqq\|y_L-S_Lg_{L\varepsilon}\|+\|S_Lg_{L\varepsilon}-S_Lg_\varepsilon\|-\|y_L-S_Lg_{L\varepsilon}\|$$
$$\leqq\|S_Lg_{L\varepsilon}-Sg_{L\varepsilon}\|+\|Sg_{L\varepsilon}-Sg_\varepsilon\|+\|Sg_\varepsilon-S_Lg_\varepsilon\|$$
$$\leqq\|S_L-S\|\cdot\|g_{L\varepsilon}\|+\|Sg_{L\varepsilon}-Sg_\varepsilon\|+\|S-S_L\|\cdot\|g_\varepsilon\|.$$

By virtue of Lemmas 6 and 7 we get

$$\lim_{L\to\infty}\|y_L-S_Lg_{L\varepsilon}\|=\lim_{L\to\infty}\|y_L-S_Lg_\varepsilon\|=\|y-Sg_\varepsilon\|. \tag{4.26}$$

As $L\to\infty$, (4.25) becomes

$$\varepsilon\varlimsup_{L\to\infty}\|g_{L\varepsilon}\|^2\leqq\varepsilon\|g_\varepsilon\|^2,$$

which implies that

$$\|g_\varepsilon\|\geqq\varlimsup_{L\to\infty}\|g_{L\varepsilon}\|. \tag{4.27}$$

From (4.24) and (4.27) we obtain

$$\lim_{L\to\infty}\|g_{L\varepsilon}\|=\|g_\varepsilon\|,$$

which implies that $g_{L\varepsilon}\to g_\varepsilon$ strongly in H.

Theorem 4 and Theorem 5 show that we can use $g_{L\varepsilon}$ as an approximate input distribution function.

Before ending this section, we shall give a simple example to illustrate the presented theory.

Example 2. Let us consider the same system as discussed in Example 1:

$$\frac{\partial u(t,x)}{\partial t}=0.1\frac{\partial^2u(t,x)}{\partial x^2}-0.1u(t,x)+g\sum_{j=0}^{[t]}f_j(Y(t-jT)-Y(t-\overline{j+1}T)),$$

$$\frac{\partial u(t,0)}{\partial x}=\frac{\partial u(t,1)}{\partial x}=0,\qquad u(0,x)=u_0(x),$$

where $g \in L_2(0, 1)$ is the unknown input distribution function. The output of the system is given by

$$z_k = \int_0^1 h(x)u(kT, x)\,dx = \sum_{m=1}^{\infty} h_m u_m(kT), \qquad k = 0, 1, \cdots.$$

From Theorem 3 and Remark 4, the system is identifiable if $\{f\} \not\equiv 0$ and $h_m \neq 0$, $m = 1, 2, \cdots$. From Remark 5 we find

$$(Sg)_k = \sum_{j=N}^{k-1} \sum_{m=1}^{\infty} g_m h_m \frac{\exp(\lambda_m T) - 1}{\lambda_m} \exp((k-j-1)\lambda_m T) f_j$$

$$= \sum_{m=1}^{\infty} g_m h_m p_{mk}, \qquad k = N+1, N+2, \cdots,$$

where

$$p_{mk} = \sum_{j=N}^{k-1} \frac{\exp(\lambda_m T) - 1}{\lambda_m} \exp((k-j-1)\lambda_m T) f_j, \qquad m = 1, 2, \cdots.$$

The adjoint operator S^* is defined by $(S^*z, g)_H = (z, Sg)_{l_2(\sigma)}$, where

$$(S^*z, g) = \sum_{m=1}^{\infty} (S^*z, \phi_m) g_m,$$

$$(z, Sg)_{l_2(\sigma)} = \sum_{k=N+1}^{\infty} z_k (Sg)_k = \sum_{m=1}^{\infty} \sum_{k=N+1}^{\infty} h_m p_{mk} z_k g_m.$$

We obtain

$$(S^*z, \phi_m) = \sum_{k=N+1}^{\infty} h_m p_{mk} z_k, \qquad m = 1, 2, \cdots,$$

from which we get

$$S^*z = \sum_{m=1}^{\infty} \sum_{k=N+1}^{\infty} h_m p_{mk} z_k \phi_m \quad \text{if } z \in l_2(\sigma).$$

Moreover we have

$$S^*Sg = \sum_{m=1}^{\infty} \sum_{k=N+1}^{\infty} h_m p_{mk} \left(\sum_{j=1}^{\infty} g_j h_j p_{jk} \right) \phi_m$$

$$= \sum_{m=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=N+1}^{\infty} h_m h_j p_{mk} p_{jk} g_j \phi_m.$$

Therefore the equation

$$(S^*S + \varepsilon I) g_\varepsilon = S^*y$$

becomes

$$\sum_{m=1}^{\infty} \sum_{j=1}^{\infty} \sum_{k=N+1}^{\infty} h_m h_j p_{mk} p_{jk} g_{\varepsilon j} \phi_m + \varepsilon \sum_{m=1}^{\infty} g_{\varepsilon m} \phi_m = \sum_{m=1}^{\infty} \sum_{k=N+1}^{\infty} h_m p_{mk} y_k \phi_m.$$

Taking the inner product with ϕ_m and using the orthonormality of $\{\phi_m\}$, we obtain the infinite number of linear equations for $g_{\varepsilon m}$

$$\varepsilon g_{\varepsilon m} + \sum_{j=1}^{\infty} \sum_{k=N+1}^{\infty} h_m h_j p_{mk} p_{jk} g_{\varepsilon j} = \sum_{k=N+1}^{\infty} h_m p_{mk} y_k, \qquad m = 1, 2, \cdots$$

where

$$y_k = z_k - \sum_{j=1}^{n} h_j \exp((k-N)\lambda_j T)\hat{u}_j(NT).$$

In the case of numerical calculations, we may solve a set of linear equations for $g_{\varepsilon 1}, \cdots, g_{\varepsilon M}$,

$$\varepsilon g_{\varepsilon m} + \sum_{j=1}^{M} \sum_{k=N+1}^{L} h_m h_j p_{mk} p_{jk} g_{\varepsilon j} = \sum_{k=N+1}^{L} h_m p_{mk} y_k, \qquad m = 1, \cdots, M$$

for L sufficiently large ($L \geqq N+M$) and an appropriate small $\varepsilon > 0$.

5. Conclusion. In this paper we have investigated the problems estimating the system states and the input distribution functions from discrete-time input–output data for distributed parameter systems. First we have constructed finite dimensional discrete-time observers to estimate the system states at sampling periods in the case where the systems have unknown input sources. We have evaluated the estimation errors. Next we have considered the problem of determining input distribution functions using the estimated system states and discrete-time input–output data. Since the systems are never identifiable in a finite number of steps, the problem has been formulated as that of minimizing the functionals on the infinite time interval. This problem is not necessarily well posed even if the systems are identifiable. Thus we have presented a feasible approximation method by regularization and discussed the limit properties of the approximate solutions under the assumption that the systems are identifiable. Moreover we have shown that we can determine a practicable approximate input distribution function which is synthesized from a sufficiently large number of observations.

REFERENCES

[1] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences, 8, Springer-Verlag, Berlin-Heidelberg, 1978.

[2] N. DUNFORD AND T. SCHWARTZ, *Linear Operators, Part* II, Interscience, New York, 1963.

[3] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.

[4] T. KOBAYASHI, *Discrete-time observability for distributed parameter systems*, Internat. J. Control, 31 (1980), pp. 181–193.

[5] ———, *Parameter identifiability for distributed parameter systems of hyperbolic type*, Internat. J. Systems Sci., 11 (1980), pp. 247–259.

[6] T. KOBAYASHI AND S. HITOTSUYA, *Observers and parameter determination for distributed parameter systems*, Internat. J. Control, 33 (1981), pp. 31–50.

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

[8] A. P. SAGE, *Optimum Systems Control*, Prentice-Hall, Englewood Cliffs, NJ, 1968.

[9] A. N. TIKHONOV, *Solution of incorrectly formulated problems and the regularization method*, Sov. Math. Dokl., 4 (1963), pp. 1035–1038.

[10] A. N. TIKHONOV AND V. Y. ARSENIN, *Solutions of Ill-posed Problems*, V. H. Winston and Sons, 1977.

[11] K. YOSHIDA, *Functional Analysis*, Springer-Verlag, New York 1965.

[12] J. ZABCZYK, *Remarks on the control of discrete-time distributed parameter systems*, SIAM J. Control, 12 (1974), pp. 721–735.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0002 $01.25/0

ROBUST STABILIZATION OF UNCERTAIN SYSTEMS*

JACQUES L. WILLEMS† AND JAN C. WILLEMS‡

Abstract. In this paper we consider the systems described by

$$dx = Ax\,dt + Bu\,dt + \sum_i \sigma_i F_i x\,d\beta_i \quad \text{or} \quad \dot{x} = Ax + Bu + \sum_i B_i F_i(x, t) C_i x,$$

and we will derive conditions under which there exists a feedback control law $u = Kx$ such that the closed loop system is stable for all σ_i or (smooth) nonlinearities F_i. The nonlinearities F_i and the noisy gains $\sigma_i\,d\beta_i$ are unknown uncertainties in the system, and the problem considered is to obtain a control law which is robust against these uncertainties, as far as stability is concerned.

Key words. robustness, feedback stabilization, invariant subspaces, stochastic stabilizability

1. Introduction. Robustness is a very important feature of control system design; it deals with the question whether some relevant qualitative properties, such as stability, are preserved if unknown perturbations are present in the dynamic system. This property is also often called *structural stability*. Consequently, it is of interest to incorporate this property as a feature of control system synthesis.

We consider the following system:

$$\dot{x}(t) = Ax(t) + Bu(t) + \sum_{i \in \mathbf{l}} B_i F_i(x, t) C_i x(t). \tag{1}$$

In this equation the last terms represent nonlinear and/or time-varying unknown (deterministic) perturbations. In this paper we will be concerned with the question whether there exists a linear stationary feedback control law $u(t) = Kx(t)$, such that the dynamic system described by (1) remains stable for *all* $F_i(x, t)$ satisfying only a Lipschitz or some smoothness condition. A similar question is analysed for the stochastic system described by the Ito equation

$$dx(t) = Ax(t)\,dt + Bu(t)\,dt + \sum_{i \in \mathbf{l}} \sigma_i F_i x(t)\,d\beta_i(t) \tag{2}$$

where the processes β_i are standard Wiener processes. Intuitively (2) should be regarded as the equation

$$\dot{x}(t) = \left[A + \sum_i F_i f_i(t)\right] x(t) + Bu(t)$$

where the processes $f_i(t)$ are stationary white noise stochastic processes.

There has been some previous work on these stabilizability problems. In [1] conditions have been derived in terms of the solution of an algebraic Riccati equation. In § 2 of the present paper the same question will be reexamined; it is shown that concise stabilizability criteria can be developed by means of geometrical techniques using the concepts of (A, B)-invariant subspaces [2] and almost (A, B)-invariant subspaces [3], [4]. In § 3 the robust stabilization of the deterministic system (1) is

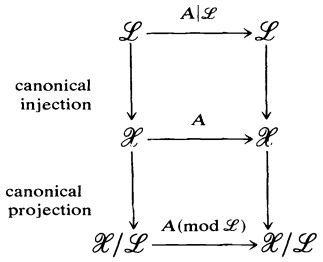
* Received by the editors July 11, 1981, and in revised form May 7, 1982.

† University of Gent, Grotesteenweg-Noord 2, 9710 Gent (Zwijnaarde), Belgium. The work of this author was supported in part by the Belgian National Research Council (FKFO grant).

‡ Mathematics Institute, P.O. Box 800, 9700 AV Groningen, the Netherlands. This work was done in part while this author was visiting the Laboratory for Information and Decision Systems of the Massachusetts Institute of Technology, supported in part by the U.S. Department of Energy under contract DOE-ET-A01-2295 To 50.

discussed; using techniques similar to those used for the Ito equation (2), criteria for robust stabilization are derived. It is shown that the model and also the results are more general than in Molander's thesis [5], which contains a rather general discussion of the robust stabilization question. A related reference is [6] where the importance of the problem considered here is argued. Finally, a recent special issue of the IEEE Transactions on Automatic Control (February (1981)) demonstrates a great deal of interest in robustness questions from the point of view of control system design.

A few words on the notation used in this paper: $\mathbb{R}$ denotes the reals, $\mathbb{C}$ the complex plane, $\mathbb{C}_g := \{s \in \mathbb{C} | \mathrm{Re}\,(s) < 0\}$, and $\bar{\mathbb{C}}_g := \{s \in \mathbb{C} | \mathrm{Re}\,(s) \leqq 0\}$. If q is a positive integer, then $\mathbf{q} := \{1, 2, \cdots, q\}$. Script capitals are used for vector spaces and subspaces. If $A : \mathscr{X} \to \mathscr{X}$ is linear and $A\mathscr{L} \subset \mathscr{L}$, then $A|\mathscr{L}$ and $A(\mathrm{mod}\,\mathscr{L})$ are the maps defined by the commutative diagram



$\langle \mathscr{K} | A \rangle$ is the largest A-invariant subspace in a given subspace $\mathscr{K}$. $\mathscr{X}_g(A)$ denotes the A-invariant subspace spanned by the eigenspaces of A corresponding to its eigenvalues in $\mathbb{C}_g$; $\mathscr{X}_{\bar{g}}(A)$ is similarly defined with respect to $\bar{\mathbb{C}}_g$. $\sigma(A)$ is the spectrum of A and $\sigma_g(A) := \sigma(A) \cap \mathbb{C}_g$. The kernel (null space) is denoted by Ker and the image (range space) by im.

Finally, for the linear system $\dot{x} = Ax + Bu$, $y = Cx$, with state space $\mathscr{X}$, we use $\langle A | \mathrm{im}\, B \rangle$ for the reachable subspace, i.e. the smallest A-invariant subspace containing im B, and $\langle \mathrm{Ker}\, C | A \rangle$ for the nonobservable subspace, i.e. the largest A-invariant subspace contained in Ker C. Finally we will be considering (almost) (A, B)-invariant and controllability subspaces [2], [3], [4] freely; the relevant facts and results are summarized in Appendix D. For a subspace $\mathscr{S}$ of $\mathscr{X}$, $\mathscr{V}^*(\mathscr{S})$, $\mathscr{V}_a^*(\mathscr{S})$, $\mathscr{V}_b^*(\mathscr{S})$ denote respectively the supremal (A, B)-invariant, $\mathscr{L}_\infty$-almost-(A, B)-invariant, and $\mathscr{L}_1$-almost-invariant subspace contained in $\mathscr{S}$, while $\mathscr{R}^*(\mathscr{S})$, $\mathscr{R}_a^*(\mathscr{S})$, $\mathscr{R}_b^*(\mathscr{S})$ denote the similarly defined (almost) controllability subspaces. The subspace $\mathscr{V}_g^*(\mathscr{S})$ is the supremal stabilizable (relative $\mathbb{C}_g$) subspace contained in $\mathscr{S}$, i.e.

$$\mathscr{V}_g^*(\mathscr{S}) = \sup \{\mathscr{V} \subset \mathscr{S} | \exists K \text{ such that } (A + BK)\mathscr{V} \subset \mathscr{V}, \sigma(A + BK) \subset \mathbb{C}_g\}.$$

$\mathscr{V}_{\bar{g}}^*(\mathscr{S})$ is similarly defined relative $\bar{\mathbb{C}}_g$.

2. Robust stabilization of stochastic systems.

2.1. Problem statement. Consider the system described by the Ito stochastic differential equation (2) where, without loss of generality, the Brownian motions β_i are assumed to be zero mean and independent:

$$E[d\beta_i(t)] = 0 \qquad \forall t, \quad \forall i \in \mathbf{l},$$

$$E[d\beta_i(t)^2] = dt \qquad \forall i \in \mathbf{l},$$

$$E[d\beta_i(t)d\beta_j(t)] = 0 \qquad \forall i, j \in \mathbf{l}, \quad i \neq j.$$

In other words, β_i is a standard Wiener process. In (2) $x \in \mathscr{X} = \mathbb{R}^n$ denotes the state, $u \in \mathscr{U} = \mathbb{R}^m$ denotes the control input. The constant matrices A, B, F_i have appropriate dimensions. The positive factors σ_i indicate the intensities of the disturbances. The symbol E denotes expectation.

We will consider the stabilizability of (2) by means of a time-invariant memoryless state feedback law

$$u(t) = Kx(t) \tag{3}$$

with K a constant matrix of appropriate dimension. Then (2) reduces to

$$dx(t) = (A + BK)x(t)\,dt + \sum_{i \in \mathbf{l}} \sigma_i F_i x(t)\,d\beta_i(t). \tag{4}$$

For this closed loop system, the mean square asymptotic stability property expressed by the definition below, will be analysed:

DEFINITION 1. System (4) is said to be *mean square asymptotically stable* if for all initial states $x(0)$

$$\lim_{t \to \infty} E[x(t)x(t)^T] = 0.$$

This leads to the following stabilizability definitions:

DEFINITION 2. System (2) is said to be *perfectly robustly stabilizable* if there exists a feedback control (3) such that (4) is mean square asymptotically stable for *all* noise intensities σ_i.

DEFINITION 3. System (2) is said to be *robustly stabilizable for all noise intensities* if for all bounds $\{s_1, \cdots, s_k\}$, there exists a feedback control (3) such that (4) is mean square asymptotically stable for all noise intensities satisfying

$$\sigma_i \leqq s_i \qquad (i \in \mathbf{l}).$$

The property expressed by Definition 3 is somewhat weaker than the property expressed by Definition 2 in that the feedback matrix K may depend on the bounds s_i; some entries of K may increase without bound as some of these bounds s_i tend to infinity.

2.2. Stability of uncontrolled systems with state-dependent noise. In order to derive stabilizability conditions for (2), we first discuss criteria for mean square asymptotic stability of the stochastic system described by the Ito differential equation

$$dx(t) = Ax(t)\,dt + \sum_{i \in \mathbf{l}} \sigma_i F_i x(t)\,d\beta_i(t). \tag{5}$$

This system is autonomous (in the sense that there are no exogenous inputs), but it contains a state-dependent noise term. The second moment matrix

$$M(t) := E[x(t)x(t)^T]$$

satisfies the matrix differential equation

$$\dot{M}(t) = AM(t) + M(t)A^T + \sum_{i \in \mathbf{l}} \sigma_i^2 F_i M(t) F_i^T \tag{6}$$

which evolves in the cone of nonnegative definite symmetric $(n \times n)$ matrices. The mean square stability properties of (5) hence depend on the eigenvalues of the linear mapping L_1 on the linear space of symmetric $(n \times n)$ matrices, defined by

$$L_1(M) := AM + MA^T + \sum_{i \in \mathbf{l}} \sigma_i^2 F_i M F_i^T. \tag{7}$$

The problem considered here is the asymptotic stability of (7) for all noise intensities σ_i. This may be resolved by introducing the subspaces $\mathcal{W}_j$, defined recursively by the following algorithm:

$$\mathcal{W}_0 := \{0\},$$

$$\mathcal{W}_j := \left\langle \bigcap_{i \in \mathbf{l}} F_i^{-1} \mathcal{W}_{j-1} \middle| A \right\rangle,$$

i.e. $\mathcal{W}_j$ is the maximal A-invariant subspace contained in $\cap_i F_i^{-1} \mathcal{W}_{j-1}$. It is easily seen by induction that the subspaces $\mathcal{W}_j$ are nested, i.e., $\mathcal{W}_{j+1} \supset \mathcal{W}_j$. Hence

$$\mathcal{W}_\infty := \lim_{j \to \infty} \mathcal{W}_j = \mathcal{W}_n$$

exists and satisfies

$$\mathcal{W}_\infty = \left\langle \bigcap_{i \in \mathbf{l}} F_i^{-1} \mathcal{W}_\infty \middle| A \right\rangle.$$

This limit is obtained monotonically in a finite number of steps.

THEOREM 1. *The following conditions are equivalent*:

(i) $\mathcal{W}_\infty = \mathcal{X}$ *and* $\sigma(A) \subset \mathbb{C}_g$.

(ii) *System* (5) *is mean square asymptotically stable for all* $\{\sigma_i\}$, $i \in \mathbf{l}$.

(iii) *In a suitable basis the matrices* A *and* F_i ($i \in \mathbf{l}$) *take the block triangular form*:

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1q} \\ 0 & A_{22} & \cdots & A_{2q} \\ 0 & 0 & \cdots & A_{3q} \\ & & \ddots & \\ 0 & 0 & \cdots & A_{qq} \end{bmatrix}, \qquad F_i = \begin{bmatrix} 0 & F_{i,12} & \cdots & F_{i,1q} \\ 0 & 0 & \cdots & F_{i,2q} \\ 0 & 0 & \cdots & F_{i,3q} \\ & & \ddots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

and $\sigma(A_{ii}) \subset \mathbb{C}_g$ *for* $i \in \mathbf{q}$.

(iv) *The Lie algebra generated by the set of matrices* $\{A, F_i; i \in \mathbf{l}\}$ (*i.e. the smallest Lie algebra containing this set*) *is solvable* [7]; *the matrices* F_i *are nilpotent, and* $\sigma(A) \subset \mathbb{C}_g$.

Proof. The equivalence of (ii), (iii), and (iv) has been shown in [1]. The elegant and computationally feasible geometrical condition (i) is proved in Appendix A. □

The geometrical criterion (i) turns out to be very well suited to attacking the stabilizability problem of system (2). This is the subject of the next section. The possibility of writing $\{A, F_i; i \in \mathbf{l}\}$ in block triangular form is related to the Jordan–Hölder theorem and has been studied in the context of constructing canonical forms for bilinear systems [8]. In fact, through condition (i) Theorem 1 yields a simple test for generalization of the question when a family of nilpotent matrices can be simultaneously triangularized. The solution of this problem is known as Engel's theorem [15]. It is concerned with a basic problem in the theory of Lie algebras, and it has implications in the theory of associative algebras and quivers.

The condition of the above theorem can be simplified if there is only one stochastic element ($l = 1$) with the corresponding F_1 of rank one: $F_1 = b_1 c_1$, where b_1 is a column vector and c_1 a row vector; in this case (5) becomes:

$$dx(t) = Ax(t)\,dt + \sigma_1 b_1 c_1 x(t)\,d\beta_1(t).$$

The condition for mean square asymptotic stability for all σ_1 is that the matrix A be Hurwitz (i.e. $\sigma(A) \subset \mathbb{C}_g$) and

$$\mathscr{W}_2 = \mathscr{X},$$

or equivalently

$$\text{im } b_1 \subset \langle \text{Ker } c_1 | A \rangle.$$

This condition is equivalent to

$$c_1 \exp(At) b_1 = 0 \quad \forall t \in \mathbb{R},$$

or

$$c_1 (Is - A)^{-1} b_1 = 0 \quad \forall s \in \mathbb{C}.$$

This decoupling condition [2] is an obvious sufficient condition also for $F_1 = B_1 C_1$ of any rank. However, if the rank of F_1 is larger than one, then the decoupling condition is in general much too strong.

2.3. Feedback stabilizability of stochastic systems. The results of § 2.2 will now be used to analyse the perfect robust stabilizability of (2). This system is perfectly robustly stabilizable if and only if there exists a matrix K such that the matrices $\{A + BK, F_i; i \in \mathbf{I}\}$ satisfy the conditions of Theorem 1. This condition can be made explicit by means of the concept of (A, B)-invariant subspaces and stabilizability subspaces (see Appendix D). To derive the criterion the following definition is required:

DEFINITION 4. Consider the subspace $\mathscr{V}^*_{g,\infty}$ defined by the following recursive algorithm:

$$\mathscr{V}^*_{g,0} := \{0\},$$

$$\mathscr{V}^*_{g,j} := \mathscr{V}^*_g \left(\bigcap_{i \in \mathbf{I}} F_i^{-1} \mathscr{V}^*_{g,j-1} \right),$$

$$\mathscr{V}^*_{g,\infty} := \lim_{j \to \infty} \mathscr{V}^*_{g,j} = \mathscr{V}^*_{g,n}.$$

As was the case for $\mathscr{W}_\infty$, this limit is attained monotonically in a finite number of steps.

THEOREM 2. *System* (2) *is perfectly robustly stabilizable if and only if* $\mathscr{V}^*_{g,\infty} = \mathscr{X}$.

Proof. (i) *The condition is necessary.* Suppose there exists a feedback matrix K such that the conditions of Theorem 1 are satisfied with respect to the system

$$dx(t) = (A + BK)x(t)\, dt + \sum_{i \in \mathbf{I}} \sigma_i F_i x(t)\, d\beta_i(t).$$

Then

$$\mathscr{W}_0 := \{0\}, \qquad \mathscr{W}_{j+1} := \left\langle \bigcap_{i \in \mathbf{I}} F_i^{-1} \mathscr{W}_j \middle| A + BK \right\rangle$$

yields $\mathscr{W}_\infty = \mathscr{X}$. Moreover, $\sigma(A + BK) \subset \mathbb{C}_g$. We claim that $\mathscr{V}^*_{g,j} \supset \mathscr{W}_j$. This is easily proved by induction. It is obviously true for $j = 0$. Moreover

$$\mathscr{V}^*_{g,j+1} = \mathscr{V}^*_g \left(\bigcap_{i \in \mathbf{I}} F_i^{-1} \mathscr{V}^*_{g,j} \right) \supset \mathscr{V}^*_g \left(\bigcap_{i \in \mathbf{I}} F_i^{-1} \mathscr{W}_j \right) \supset \left\langle \bigcap_{i \in \mathbf{I}} F_i^{-1} \mathscr{W}_j \middle| A + BK \right\rangle = \mathscr{W}_{j+1}$$

yields the inductive step. Hence $\mathscr{V}^*_{g,\infty} \supset \mathscr{W}_\infty = \mathscr{X}$, which proves the necessity of the condition.

(ii) *The condition is sufficient.* To prove the sufficiency of the condition by means of Theorem 1, we need to show that there exists a single feedback matrix K such that for all j the (A, B)-invariant subspaces $\mathcal{V}^*_{g,j}$ become $(A+BK)$-invariant subspaces with the properties required by Theorem 1, in particular stabilizability and inclusion of $\mathcal{V}^*_{g,j}$ in $\bigcap_i F_i^{-1}\mathcal{V}^*_{g,j-1}$. This is not trivial, since we have no guarantee that the (A, B)-invariant subspaces can be made $(A+BK)$-invariant by means of the *same* matrix K (independent of j). This feature is called *compatibility* of the (A, B)-invariant subspaces $\mathcal{V}^*_{g,j}$ (see Appendix D). In general, compatibility is difficult to analyse. It is not hard to show that the (A, B)-invariant subspaces $\mathcal{V}^*_{g,j}$ are compatible as (A, B)-invariant subspaces, because they are nested ($\mathcal{V}^*_{g,j} \subset \mathcal{V}^*_{g,j+1}$). However, here we have to prove in addition that they are also compatible with respect to the stabilizability and inclusion properties. Let the state space be partitioned as

$$\mathcal{X} = \mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \cdots \oplus \mathcal{X}_p$$

where p is the integer such that $\mathcal{V}^*_{g,p} = \mathcal{X}$, $\mathcal{V}^*_{g,p-1} \neq \mathcal{X}$, and where the subspaces $\mathcal{X}_j$ are chosen in such a way that for all $j \in \mathbf{p}$,

$$\mathcal{X}_1 \oplus \mathcal{X}_2 \oplus \cdots \oplus \mathcal{X}_j = \mathcal{V}^*_{g,j}. \tag{8}$$

If the conditions of the theorem hold, then for all $j \in \mathbf{p}$ there exists a feedback matrix K_j such that $\sigma(A+BK_j) \subset \mathbb{C}_g$, and

$$\mathcal{V}^*_{g,j} = \left\langle \bigcap_{i \in \mathbf{l}} F_i^{-1}\mathcal{V}^*_{g,j-1} \middle| A+BK_j \right\rangle.$$

Let K_j^i be defined by

$$K_j x = \sum_i K_j^i x_i$$

where x_i is the component of x in $\mathcal{X}_i$. Then we check that the feedback law

$$K^* x = K_1^1 x_1 + K_2^2 x_2 + \cdots + K_p^p x_p$$

makes the subspaces $\mathcal{V}^*_{g,j}$ $(A+BK^*)$-invariant, and such that $\mathcal{V}^*_{g,j} \subset \bigcap_{i \in \mathbf{l}} F_i^{-1}\mathcal{V}^*_{g,j-1}$. In a basis compatible with the above partitioning of the state space, $A^* := A+BK^*$ and F_i have hence the form:

$$A^* = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ 0 & A_{22} & \cdots & A_{2p} \\ & & \ddots & \\ 0 & 0 & \cdots & A_{pp} \end{bmatrix}, \qquad F_i = \begin{bmatrix} 0 & F_{i,12} & \cdots & F_{i,1p} \\ 0 & 0 & \cdots & F_{i,2p} \\ & & \ddots & \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Then there exists a transformation of the input and of the state, which does not change the structure of A^* and F_i (by redefining $\mathcal{X}_2, \cdots, \mathcal{X}_p$, such that (8) remains true), but which transforms the input matrix B into the form [9, pp. 543–544]:

$$B = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & B_p \end{bmatrix}.$$

Since $\mathcal{V}^*_{g,1}$ is a stabilizable (A, B)-invariant subspace, the pair (A_{11}, B_1) is stabilizable; there hence exists a partial feedback of the state $K_1 x_1$ such that $\sigma(A_{11}+B_1K_1) \subset \mathbb{C}_g$.

Also $\mathcal{V}^*_{g,2}$ is a stabilizable (A, B)-invariant subspace. Hence the pair

$$\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix}$$

is stabilizable. This, however, implies the stabilizability of the pair (A_{22}, B_2). Therefore a feedback $K_2 x_2$ exists such that $\sigma(A_{22}+B_2K_2) \subset \mathbb{C}_g$. Proceeding in this fashion all subspaces $\mathcal{V}^*_{g,j}$ are stabilized without altering the structure of A^* or F_i. □

From the definition of $\mathcal{V}^*_{g,\infty}$ it is immediately clear that an equivalent statement to the criterion of Theorem 2 is as follows:

COROLLARY 1. *System* (2) *is perfectly robustly stabilizable if and only if for some finite integer k*

$$\sum_{i \in \mathbf{l}} \operatorname{im} F_i \subset \mathcal{V}^*_{g,k}. \tag{9}$$

Proof. Condition (9) implies

$$\bigcap_{i \in \mathbf{l}} F_i^{-1} \mathcal{V}^*_{g,k} = \mathcal{X}$$

and hence

$$\mathcal{V}^*_{g,k+1} = \mathcal{X}. \qquad \square$$

Here also the condition can be simplified if there is only one stochastic element and the corresponding matrix F_1 has rank one:

$$dx(t) = [Ax(t) + Bu(t)]\, dt + \sigma_1 b_1 c_1 x(t)\, d\beta_1(t). \tag{10}$$

Then the stabilizability condition becomes

$$\mathcal{V}^*_{g,2} = \mathcal{X}$$

or

$$\operatorname{im} F_1 = \operatorname{im} b_1 \subset \mathcal{V}^*_{g,1} = \mathcal{V}^*_g(\operatorname{Ker} c_1).$$

This condition implies the existence of a feedback matrix K such that $A+BK$ is Hurwitz and $c_1(Is - A - BK)^{-1} b_1$ vanishes identically. The condition of Theorem 2 is then equivalent to the criterion for disturbance decoupling with stability from the disturbance input $\operatorname{im} F_1$ or $\operatorname{im} b_1$ to the output with $\operatorname{Ker} F_1$ or $\operatorname{Ker} c_1$. In general, however, the condition of Theorem 2 is much weaker than the disturbance decoupling requirement.

We notice that the criteria of Theorem 2 or Corollary 1 are also sufficient for feedback stabilizability in cases where:

(i) the stochastic disturbances are zero-mean but not necessarily white, provided they have finite second order moments which are uniformly bounded in time,

(ii) stabilizability with respect to other moments than the second moment is considered.

2.4. High-gain stabilizability of stochastic systems. In this section it is investigated to what extent the criterion of § 2.3 can be relaxed if only stabilizability of system (4) is required for all σ_i; this means that for any $\{\sigma_i\}$ a stabilizing feedback matrix K must exist, such that

$$\dot{M} = (A+BK)M + M(A+BK)' + \sum_{i \in \mathbf{l}} \sigma_i^2 F_i M F_i' \tag{11}$$

is asymptotically stable in the cone of nonnegative definite $(n \times n)$ matrices. Since the matrix K may depend on the noise intensities σ_i, some elements may go to infinity as the noise intensities increase without bound. Then there does not exist a single feedback matrix which stabilizes (4) in the mean square for all noise intensities. Considering the criterion derived in § 2.3, one might be tempted to conjecture that for this type of stabilizability the conditions of Theorem 2 may be relaxed by replacing

- (A, B)-invariant subspaces by almost (A, B)-invariant subspaces [4],
- $\mathbb{C}_g$ by $\bar{\mathbb{C}}_g$.

The notions of almost invariant subspaces have been introduced in [3] and further worked out in [4]. The relevant facts from that reference are summarized in Appendix D.

It is unlikely that the above conjecture is correct because of the high gains involved in the transfer function which results when the gains $\sigma_i \to \infty$. The criterion of Theorem 3 below is not as strong as the above conjecture, but nevertheless it yields a useful relaxation of the conditions of Theorem 2; indeed *in the last step* (A, B)-invariance may be replaced by almost (A, B)-invariance and $\mathbb{C}_g$ by $\bar{\mathbb{C}}_g$.

THEOREM 3. *Let the subspaces $\mathcal{V}^*_{g,j}$ be as defined in* § 2.3. *Let $\mathcal{R}^*_b(\mathcal{S})$ be as defined above. Then system* (2) *is robustly stabilizable for all noise intensities if the pair* (A, B) *is stabilizable and*

$$\sum_{i\in\mathbf{I}} \operatorname{im} F_i \subset \mathcal{V}^*_{\bar{g}}\Big(\bigcap_{i\in\mathbf{I}} F_i^{-1}\mathcal{V}^*_{g,\infty}\Big) + \mathcal{R}^*_b\Big(\bigcap_{i\in\mathbf{I}} F_i^{-1}\mathcal{V}^*_{g,\infty}\Big). \tag{12}$$

Proof. From the definition of the subspaces $\mathcal{V}^*_{g,j}$, it follows that there exists a constant feedback matrix K such that in an appropriate basis and with the control

$$u(t) = Kx(t) + v(t)$$

the system representation (2) takes the form

$$dx(t) = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,q+1} \\ 0 & A_{2,2} & \cdots & A_{2,q+1} \\ & & \ddots & \\ 0 & 0 & \cdots & A_{q+1,q+1} \end{bmatrix} x(t)\,dt + \begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_{q+1} \end{bmatrix} v(t)\,dt$$

$$+ \sum_{i\in\mathbf{I}} \sigma_i \begin{bmatrix} 0 & F_{i,1,2} & \cdots & F_{i,1,q} & F_{i,1,q+1} \\ 0 & 0 & \cdots & F_{i,2,q} & F_{i,2,q+1} \\ & & \ddots & & \\ 0 & 0 & \cdots & 0 & F_{i,q,q+1} \\ 0 & 0 & \cdots & 0 & F_{i,q+1,q+1} \end{bmatrix} x(t)\,d\beta_i(t)$$

with $\sigma(A_{i,i}) \subset \mathbb{C}_g$ for $i \in \mathbf{q}$, and with $\mathcal{V}^*_{g,\infty} = \mathcal{V}^*_{g,q}$. The conditions of the theorem imply that the pair $(A_{q+1,q+1}, B_{q+1})$ is stabilizable and that

$$\sum_{i\in\mathbf{I}} \operatorname{im} F_{i,q+1,q+1} \subset \mathcal{V}^*_{\bar{g}}\Big(\bigcap_{i\in\mathbf{I}} \operatorname{Ker} F_{i,q+1,q+1}\Big) + \mathcal{R}^*_b\Big(\bigcap_{i\in\mathbf{I}} \operatorname{Ker} F_{i,q+1,q+1}\Big)$$

where $\mathcal{R}^*_b$ and $\mathcal{V}^*_{\bar{g}}$ are taken relative to $(A_{q+1,q+1}, B_{q+1})$. Corollary B.2 and Lemma A.3 from the appendices then show that the reduced order system

$$dx_{q+1}(t) = A_{q+1,q+1}x_{q+1}(t)\,dt + B_{q+1}v(t)\,dt + \sum_{i\in\mathbf{I}} \sigma_i F_{i,q+1,q+1}x_{q+1}(t)\,d\beta_i(t) \tag{13}$$

is robustly stabilizable for all noise intensities. The remainder of the proof of the theorem now easily follows from the triangular structure of the system equation. □

Theorem 3 is particularly interesting in the special case considered in (10), where there is only one stochastic element and the corresponding matrix F_1 has rank one. Then the criterion for robust stabilizability can be derived from the criterion for perfect robust stabilizability from § 2.3 by just replacing (A, B)-invariance by almost (A, B)-invariance and $\mathbb{C}_g$ by $\bar{\mathbb{C}}_g$: system (10) is robustly stabilizable for all noise intensities if and only if (A, B) is stabilizable and

$$\operatorname{im} b_1 \subset \mathscr{V}_{\bar{g}}^*(\operatorname{Ker} c_1) + \mathscr{R}_b^*(\operatorname{Ker} c_1).$$

Suppose in addition that there is only one input: B is a column vector which is denoted by b. Then the perfect robust stabilizability and the high-gain stabilizability conditions for system (10) can be expressed in terms of the transfer function

$$F(s) := \frac{c_1(Is - A)^{-1} b_1}{c_1(Is - A)^{-1} b}.$$

We assume that (A, c_1) is a detectable pair [2]; this entails no loss of generality, since the stabilizability of the pair (A, b) implies that there exists a feedback vector k such that $(A + bk, c_1)$ is detectable. System (10) is perfectly robustly stabilizable if and only if (A, b) is stabilizable, $F(s)$ is strictly proper, and, after cancellation of common factors, $F(s)$ has no poles with nonnegative real parts. System (10) is robustly stabilizable for all noise intensities if and only if (A, b) is stabilizable and, after cancellation of common factors, $F(s)$ has no poles with positive real parts.

3. Robust stabilization of uncertain deterministic systems.

3.1. Problem formulation. In this section we consider the deterministic counterpart of the problem analysed in § 2. Here the question is: When can a system with an unknown nonlinear and/or time-varying element can be stabilized by means of a linear state feedback regulator? In § 1 we introduced the class of systems (1) which we have in mind. However, this equation may be written in the following form, which makes it more alike to the system considered in § 2:

$$\dot{x}(t) = Ax(t) + Bu(t) + \sum_{i \in \mathbf{l}} f_i(x(t), t) F_i x(t). \tag{14}$$

This formulation has (1) as a special case. To see this write the nonlinear term in (1) as

$$B_i F_i(x, t) C_i = \sum_{r,s} [F_i(x, t)]_{r,s} [B_i]_r [C_i]_s$$

where $[B_i]_r$ denotes the rth column, $[C_i]_s$ denotes the sth row, and $[F_i(x, t)]_{r,s}$ denotes the (r, s) entry of B_i, C_i and $F_i(x, t)$, respectively. The system formulation (14) also has as a special case the system

$$\dot{x}(t) = Ax(t) + Bu(t) + \sum_{i \in \mathbf{k}} \sum_{j \in \mathbf{l}} B_{ij} F_i(x(t), t) C_{ij} x(t), \tag{15}$$

which is perhaps the most logical starting point for the class of robustness problems considered here.

DEFINITION 5. We say that (14) is *perfectly robustly stabilizable* if there exists a feedback law (3) such that the null solution of

$$\dot{x}(t) = (A + BK)x(t) + \sum_{i \in \mathbf{l}} f_i(x(t), t) F_i x(t) \tag{16}$$

is asymptotically stable in the large for all bounded nonlinear and/or time-varying gains $f_i(x, t)$. We assume throughout that the gains $f_i(x, t)$ are sufficiently smooth, e.g. Lipschitz, such that the existence and the uniqueness of the solution of (1) is ensured. We say that (14) is *robustly stabilizable for all uncertain gains* if for any set $\{m_i; i \in \mathbf{l}\}$, there exists a control law (3) such that the null solution of (16) is asymptotically stable in the large for all $f_i(x, t)$ satisfying $|f_i(x, t)| < m_i$.

Note again that in the second formulation K may depend on the bounds m_i, while in the first formulation this is not possible. The results which will be obtained, actually imply the stabilizability of the system with the structure of (14) in which the nonlinearity f_i is any $\mathscr{L}_2$-input/output stable operator. The robust stabilizability problem for the related but more restricted class of systems

$$\dot{x}(t) = Ax(t) + Bu(t) + Gf(Hx, t) \tag{15'}$$

has been studied previously by Molander [5] in essentially the same setting. Without actually introducing almost invariant subspaces he does obtain results which are important special cases of ours. Specifically he shows that (15') is robustly stabilizable if the system (A, B, G, H) may be stably almost disturbance decoupled in the $\mathscr{L}_1$-sense. This result is a special case of our Theorem 6; it requires that

$$\operatorname{im} G \subset \mathscr{V}_g^*(\operatorname{Ker} H) + \mathscr{R}_b^*(\operatorname{Ker} H).$$

A similar result has been obtained independently in [4, Thm. 17].

3.2. Criterion for perfect robustness. In order to derive a criterion for robust stabilizability we first consider the uncontrolled system

$$\dot{x}(t) = Ax(t) + \sum_{i \in \mathbf{l}} f_i(x(t), t) F_i x(t) \tag{17}$$

and investigate when the null solution of this system is asymptotically stable in the large for all bounded functions $f_i(x, t)$. Our sufficient conditions are:

(i) the matrix A is Hurwitz;

(ii) the matrices F_i are nilpotent;

(iii) the matrices $\{A, F_i, i \in \mathbf{l}\}$ can be transformed to upper block triangular form by means of the same similarity transformation.

Expressed geometrically this yields:

THEOREM 4. *The null solution of* (17) *is asymptotically stable in the large for all bounded gains* $f_i(x, t)$, *if the matrix* A *is Hurwitz, and* $\mathscr{W}_\infty = \mathscr{X}$, *with* $\mathscr{W}_\infty$ *defined preceding Theorem* 1.

This result follows immediately from Theorem 1. By means of Theorem 4 and the ideas used in proving Theorem 2 from Theorem 1, we obtain:

THEOREM 5. *Let* $\mathscr{V}_{g,\infty}^*$ *be as in Definition* 4. *Then* (14) *is perfectly robustly stabilizable if* $\mathscr{V}_{g,\infty}^* = \mathscr{X}$.

The condition of Theorem 5 is of course equivalent to $\sum_{i \in \mathbf{l}} \operatorname{im} F_i \subset \mathscr{V}_{g,q}^*$ for some integer q. The criteria of Theorems 4 and 5 can be simplified in the cases

$$\dot{x}(t) = Ax(t) + Bu(t) + B_1 F_1(x(t), t) C_1 x(t), \tag{18}$$

in which case the criterion requires that the system

$$\dot{x}(t) = Ax(t) + Bu(t) + B_1 d(t), \qquad z(t) = C_1 x(t)$$

should be disturbance decouplable with internal stability by state feedback [2]. In the more general situation

$$\dot{x}(t) = Ax(t) + Bu(t) + \sum_{i \in \mathbf{l}} B_i F_i(x(t), t) C_i x(t),$$

the criterion requires that the system

$$\dot{x}(t) = Ax(t) + Bu(t) + \sum_{i \in \mathbf{l}} B_i d_i(t), \qquad z_i(t) = C_i x(t), \qquad i \in \mathbf{l},$$

should be strictly triangularly disturbance decouplable with internal stability in the sense that there should exist a feedback K with $\sigma(A+BK) \subset \mathbb{C}_g$ such that in the closed loop system

$$\dot{x}(t) = (A+BK)x(t) + \sum_{i \in \mathbf{l}} B_i d_i(t), \qquad z_i(t) = C_i x(t), \qquad i \in \mathbf{l},$$

there should exist a permutation of $\mathbf{l}$ such that the resulting transfer function $(d_1, d_2, \cdots, d_k) \mapsto (z_1, z_2, \cdots, z_k)$ is strictly upper block triangular.

3.3. Criterion for robustness for all uncertain gains. As in § 2.4, it is tempting to conjecture from Theorem 5 that robustness for all uncertain gains would be achievable under the conditions of Theorem 5, but with almost (A, B)-invariance replacing (A, B)-invariance and $\bar{\mathbb{C}}_g$ replacing $\mathbb{C}_g$. However, since the stabilizability condition in this case comes down to impulse response quenching in the $\mathscr{L}_1$-sense, it is not possible to replace $\mathbb{C}_g$ by $\bar{\mathbb{C}}_g$ (see the example at the end of Appendix C). Nevertheless it is possible to use almost (A, B)-invariant subspaces *in the last step* of the algorithm of Theorem 5.

THEOREM 6. *Let the subspace $\mathscr{V}^*_{g,\infty}$ be as defined in* § 2.3 *and let $\mathscr{R}^*_b(\cap_{i \in \mathbf{l}} F_i^{-1} \mathscr{V}^*_{g,\infty})$ be as defined preceding Theorem* 3. *Then* (14) *is robustly stabilizable for all uncertain gains if (A, B) is stabilizable and*

$$\sum_{i \in \mathbf{l}} \operatorname{im} F_i \subset \mathscr{V}^*_{g,\infty} + \mathscr{R}^*_b\Big(\bigcap_{i \in \mathbf{l}} F_i^{-1} \mathscr{V}^*_{g,\infty}\Big). \tag{19}$$

Proof. The proof of this theorem follows exactly the same route as the proof of Theorem 3 except where Lemma A.3 of Appendix A was used. Here instead Proposition C.1 of Appendix C yields the result. □

Note that condition (19) could equivalently be expressed as

$$\sum_{i \in \mathbf{l}} \operatorname{im} F_i \subset \mathscr{V}^*_g\Big(\bigcap_{i \in \mathbf{l}} F_i^{-1} \mathscr{V}^*_{g,\infty}\Big) + \mathscr{R}^*_b\Big(\bigcap_{i \in \mathbf{l}} F_i^{-1} \mathscr{V}^*_{g,\infty}\Big). \tag{20}$$

This shows more clearly the relationship between Theorems 3 and 6.

It is straightforward to specialize the result of Theorem 6 to the case where, as in equation (10) for the stochastic case, there is only one nonlinear term; the corresponding matrix F_1 has rank one, and there is only one input:

$$\dot{x}(t) = Ax(t) + bu(t) + f_1(x(t), t) b_1 c_1 x(t) \tag{21}$$

where the same notation as in (10) is used, and b is a column vector. Suppose (A, c_1) detectable and (A, b) stabilizable. This system is perfectly robustly stabilizable if the transfer function $F(s)$, defined in § 2.4, (i) is strictly proper and (ii), after cancellation of common factors, has only poles with negative real parts. It is robustly stabilizable for all uncertain gains if (ii) holds.

The conditions of Theorem 3 are in general not sufficient to guarantee robust stabilizability for all uncertain gains in the deterministic case. This distinction is an intrinsic one and may be illustrated by means of (21) and

$$dx(t) = Ax(t)\,dt + bu(t)\,dt + \sigma_1 b_1 c_1 x(t)\,d\beta_1(t). \tag{22}$$

Robust stabilizability of (22) for all noise intensities requires for any $\varepsilon > 0$ the existence of a feedback vector k such that $\sigma(A+bk) \subset \mathbb{C}_g$ and

$$\int_0^\infty w(t)^2\, dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} |H(j\omega)|^2\, d\omega \leqq \varepsilon$$

where

$$w : t \mapsto c_1 \exp[(A+bk)t]b_1,$$
$$H : s \mapsto c_1(Is - A - bk)^{-1}b_1.$$

On the other hand, robust stabilizability of (21) for all uncertain gains requires

$$\int_0^\infty |w(t)|\, dt \leqq \varepsilon \quad \text{or} \quad \sup_{\omega \in \mathbb{R}} |H(j\omega)| \leqq \varepsilon.$$

Take for example

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad b_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad c_1 = [0 \quad 1].$$

Then system (22) is stabilizable for any noise intensity. However, (21) is not perfectly robustly stabilizable. Even for linear time-invariant gains $f(x, t) = k$, there does not exist a feedback strategy which stabilizes the system at the same time for all gains satisfying $|k| < k_{\max}$ if $k_{\max} > 1$. This is in agreement with the above reasoning. Indeed, taking

$$u(t) = kx(t) = [-\alpha \quad -\beta]x(t)$$

yields the closed loop transfer function

$$H(s) = \frac{-\alpha}{s^2 + (1+\alpha+\beta)s + \alpha}.$$

The condition $\sigma(A+bk) \subset \mathbb{C}_g$ requires $\alpha > 0$, $1+\alpha+\beta > 0$. Now $|H(0)| = 1$ cannot be influenced by α and β, whereas

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} |H(j\omega)|^2\, d\omega = \frac{\alpha}{2(1+\alpha+\beta)}$$

can indeed be made arbitrarily small.

4. Discrete-time systems. A similar analysis can be performed on the stabilizability of the discrete-time stochastic system

$$x_{t+1} = Ax_t + Bu_t + \sum_{i \in \mathbf{I}} \sigma_i F_i x_t f_{it} \tag{23}$$

where the scalar processes f_{it} are zero mean uncorrelated normalized white noise processes, and with respect to the robustness of the nonlinear discrete-time deterministic system

$$x_{t+1} = Ax_t + Bu_t + \sum_{i \in \mathbf{I}} f_i(x_t, t)F_i x_t. \tag{24}$$

It follows from Appendix A that the criteria for perfect robust stabilizability of (23) and of (24) are exactly the same as in the continuous-time case, provided of course $\operatorname{Re}(s) < 0$ is replaced by $|z| < 1$. However for robust stabilizability of (23) for

all noise intensities or robust stabilizability of (24) for all uncertain gains, it is not possible to relax the conditions as much as in the continuous-time cases. For (24) in fact no relaxation has been obtained. For robust stabilizability for all noise intensities of (23) it is possible to replace the condition $|z|<1$ by $|z|\leqq 1$ in the last step.

The distinction between discrete-time and continuous-time systems can be seen as follows. The feedback strategy $u(t)=Kx(t)$ stabilizes the stochastic continuous-time system (2) if and only if the linear mapping

$$M \mapsto L_c(M) := (A+BK)M + M(A+BK)^T + \sum_{i\in \mathbf{l}} \sigma_i^2 F_i M F_i^T \tag{25}$$

has only eigenvalues with negative real parts. The feedback strategy $u_t = Kx_t$ stabilizes the stochastic discrete-time system (23) if and only if the linear mapping

$$M \mapsto L_d(M) := (A+BK)M(A+BK)^T + \sum_{i\in \mathbf{l}} \sigma_i^2 F_i M F_i^T \tag{26}$$

has only eigenvalues with magnitude smaller than 1. The eigenvalues of $L_d(M)$ are larger than the eigenvalues of the mappings

$$M \mapsto L_i(M) := \sigma_i^2 F_i M F_i^T.$$

The eigenvalues of $L_i(M)$ are $\sigma_i^2 \lambda_a(F_i)\lambda_b(F_i)$, where $\lambda_a(F_i)$ and $\lambda_b(F_i)$ are arbitrary eigenvalues of F_i. Hence the existence of a stabilizing feedback for all noise intensities requires that the matrices F_i have only zero eigenvalues. This is also true if the feedback matrix K is allowed to depend on the noise intensities $\{\sigma_i\}$, hence for the property of *robust stabilizability for all noise intensities*. A similar conclusion is not valid however for continuous-time systems.

5. Example. In this section the application of the criteria developed in §§ 2, 3, and 4, is illustrated on the example [2] of a second-order system with the data:

$$A = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad F_1 = b_1 c_1, \quad b_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad c_1 = [a \quad 1].$$

The continuous-time and discrete-time, stochastic and deterministic, cases will be examined:

$$dx(t) = Ax(t)\,dt + bu(t)\,dt + \sigma_1 F_1 x(t)\,d\beta_1(t), \tag{27}$$

$$\dot{x}(t) = Ax(t) + bu(t) + f_1(x(t), t)F_1 x(t), \tag{28}$$

$$x_{t+1} = Ax_t + bu_t + \sigma_1 F_1 x_t f_{1t}, \tag{29}$$

$$x_{t+1} = Ax_t + bu_t + f_1(x_t, t)F_1 x_t. \tag{30}$$

For the continuous-time case we obtain:

(i) $a<0$: $\quad \mathscr{V}^*_{g,\infty} = \{0\}, \quad \mathscr{V}^*_{\bar{g}}(F_1^{-1}\mathscr{V}^*_{g,\infty}) = \{0\}, \quad \mathscr{R}^*_b(F_1^{-1}\mathscr{V}^*_{g,\infty}) = \operatorname{im} b,$

(ii) $a=0$: $\quad \mathscr{V}^*_{g,\infty} = \{0\}, \quad \mathscr{V}^*_{\bar{g}}(F_1^{-1}\mathscr{V}^*_{g,\infty}) = \operatorname{im}\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathscr{R}^*_b(F_1^{-1}\mathscr{V}_{g,\infty}) = \operatorname{im} b,$

(iii) $a>0, \quad a\neq .5$: $\quad \mathscr{V}^*_{g,\infty} = \operatorname{im}\begin{bmatrix} 1 \\ -a \end{bmatrix}, \quad \mathscr{R}^*_b(F_1^{-1}\mathscr{V}^*_{g,\infty}) = \operatorname{im} b,$

(iv) $a=.5$: $\quad \mathscr{V}^*_{g,\infty} = \mathscr{X}.$

For the discrete-time case the results are:

(i) $a = .5$: $\mathcal{V}^*_{g,\infty} = \mathcal{X}$,

(ii) $|a| < 1$, $a \neq .5$: $\mathcal{V}^*_{g,\infty} = \operatorname{im}\begin{bmatrix} 1 \\ -a \end{bmatrix}$, $\mathcal{V}^*_{\bar{g}}(F_1^{-1}\mathcal{V}^*_{g,\infty}) = \mathcal{V}^*_{g,\infty}$,

(iii) $|a| = 1$: $\mathcal{V}^*_{g,\infty} = \{0\}$, $\mathcal{V}^*_{\bar{g}}(F_1^{-1}\mathcal{V}^*_{g,\infty}) = \operatorname{im}\begin{bmatrix} 1 \\ -a \end{bmatrix}$,

(iv) $|a| > 1$: $\mathcal{V}^*_{g,\infty} = \{0\}$, $\mathcal{V}^*_{\bar{g}}(F_1^{-1}\mathcal{V}^*_{g,\infty}) = \{0\}$.

Hence the stabilizability criteria are derived:

(i) The stochastic continuous-time system (27) is perfectly robustly stabilizable if $a = .5$. It is robustly stabilizable for all noise intensities if $a \geqq 0$.

(ii) The deterministic continuous-time system (28) is perfectly robustly stabilizable if $a = .5$. It is robustly stabilizable for all uncertain gains if $a > 0$.

(iii) The stochastic discrete-time system (29) is perfectly robustly stabilizable if $a = .5$. The same condition holds for robust stabilizability for all noise intensities.

(iv) The deterministic discrete-time system (30) is perfectly robustly stabilizable if $a = .5$. No relaxation of this condition is obtained for robust stabilizability for all uncertain gains.

Appendix A. The first part of this appendix is relevant to the proof of Theorem 1. We consider the following linear mappings in the space of $(n \times n)$ symmetric matrices:

$$M \mapsto L_1(M) := AM + MA^T + \sum_{i=1}^{l} \sigma_i^2 F_i M F_i^T, \tag{A.1}$$

$$M \mapsto L_2(M) := \sum_{i=1}^{l} \sigma_i^2 F_i \int_0^\infty \exp(A\tau) M \exp(A^T\tau)\, d\tau\, F_i^T, \tag{A.2}$$

$$M \mapsto L_3(M) := AMA^T + \sum_{i=1}^{l} \sigma_i^2 F_i M F_i^T, \tag{A.3}$$

$$M \mapsto L_4(M) := \sum_{i=1}^{l} \sigma_i^2 F_i \sum_{j=0}^{\infty} A^j M A^{Tj} F_i^T, \tag{A.4}$$

where L_2 is only defined if A is a Hurwitz matrix, i.e. $\sigma(A) \subset \mathbb{C}_g$, and where L_4 is only defined if A has only eigenvalues smaller than 1 in modulus. Since L_2 and L_4 map the cone of nonnegative definite matrices into itself, it follows that the largest eigenvalue of L_2 and L_4 is real and positive, and that it increases with increasing $\sigma_1, \sigma_2, \cdots, \sigma_l$.

LEMMA A.1. (i) *The linear mapping L_1 has all its eigenvalues in $\mathbb{C}_g$ if and only if the matrix A is a Hurwitz matrix and the mapping L_2 has only eigenvalues with modulus smaller than* 1.

(ii) *The linear mapping L_3 has all its eigenvalues inside the open unit disk of the complex plane if and only if all eigenvalues of the matrix A and of the mapping L_4 are smaller than* 1 *in modulus.*

Part (ii) follows from an earlier paper [10]; part (i) is proved in a similar fashion and is left to the reader. The lemma can also be obtained using the analysis of [11]. Lemma A.1 yields the following theorem:

THEOREM A.1. (i) *The mapping L_1 has all its eigenvalues in $\mathbb{C}_g$ for all $\{\sigma_i; i \in \mathbf{l}\}$ if and only if the eigenvalues of the mapping L_2 vanish for some nonzero values of $\sigma_1, \cdots, \sigma_l$. In this case all eigenvalues of L_2 vanish for all $\{\sigma_i; i \in \mathbf{l}\}$, i.e. L_2 is nilpotent; moreover, the eigenvalues of L_1 are independent of $\{\sigma_i; i \in \mathbf{l}\}$.*

(ii) *The mapping L_3 has all its eigenvalues in the open unit disk $\mathcal{D}_g := \{z \in \mathbb{C} | |z| < 1\}$ for all $\{\sigma_i; i \in \mathbf{l}\}$ if and only if the eigenvalues of the mapping L_4 vanish for some nonzero values of $\sigma_1, \cdots, \sigma_l$. In this case all eigenvalues of L_4 vanish for all $\{\sigma_i; i \in \mathbf{l}\}$, i.e., L_4 is nilpotent; moreover, the eigenvalues of L_3 are independent of $\{\sigma_i; i \in \mathbf{l}\}$.*

Let the subspaces $\mathcal{W}_j$ and $\mathcal{W}_\infty$ be defined as in § 2.2, preceding Theorem 1.

LEMMA A.2. *The following statements are equivalent*:

(i) $\mathcal{W}_\infty = \mathcal{X}$,

(ii) *L_2 is nilpotent,*

(iii) *L_4 is nilpotent.*

Proof. Only (i)$\Leftrightarrow$(ii) is proven; (i)$\Leftrightarrow$(iii) is completely similar. We need to prove that $L_2^m(M) = 0$ for all $M = M^T$ and $m \geqq n(n+1)/2$.

(i) *The condition is sufficient.* Let $x \in \mathcal{W}_j$. Compute $L_2(xx^T)$. Because of the definition of $\mathcal{W}_j$, we have

$$\exp(At)x \in \mathcal{W}_j$$

and

$$F_i \exp(At)x \in \mathcal{W}_{j-1} \qquad (\forall i \in \mathbf{l}).$$

Hence

$$L_2(xx^T) = \sum_k y_k y_k^T$$

with all $y_k \in \mathcal{W}_{j-1}$. Repeatedly applying L_2 yields

$$L_2^\alpha(xx^T) = 0$$

for $\alpha \geqq j$. This proves the sufficiency of the condition since any symmetric matrix can be expressed as the linear combination of dyads of the form xx^T.

(ii) *The condition is necessary.* If $\mathcal{W}_\infty \neq \mathcal{X}$, then

$$\mathcal{W}_\infty = \left\langle \bigcap_{i \in \mathbf{l}} F_i^{-1} \mathcal{W}_\infty \middle| A \right\rangle.$$

Let $x^* \in \mathcal{X}$ and $x^* \notin \mathcal{W}_\infty$. Consider $L_2(x^* x^{*T})$; from the above property of $\mathcal{W}_\infty$ it follows that $F_i \exp(At)x^* \in \mathcal{W}_\infty$ cannot be true for all i and all t. Hence

$$L_2(x^* x^{*T}) = \sum_k y_k y_k^T$$

where at least one of the vectors $y_k \notin \mathcal{W}_\infty$. Repeatedly applying L_2 yields that $L_2(x^* x^{*T})$ cannot vanish for any integer α. □

The second part of this appendix is relevant to the proof of Theorem 3.

LEMMA A.3. *For all $\{F_i; i \in \mathbf{l}\}$ and $\{K_i < \infty; i \in \mathbf{l}\}$ there exist bounds $\{\alpha_{ij} > 0; i, j \in \mathbf{l}\}$ such that system (5) is mean square asymptotically stable for all noise intensities $\{\sigma_i | \sigma_i \leqq K_i\}$ and all system matrices A such that $\sigma(A) \subset \mathbb{C}_g$ and*

$$\int_0^\infty \|F_i \exp(At) F_j\|^2 \, dt < \alpha_{ij} \qquad (i, j \in \mathbf{l}).$$

Proof. The largest eigenvector λ^* of the mapping L_2, defined by (A.2), corresponds to a nonnegative definite eigenvector M^* which, for nonzero λ^*, is of the form

$$M^* = \sum_{j \in \mathbf{I}} F_j N_j F_j^T.$$

Let M_j denote $F_j N_j F_j^T$. We have $N_j \geqq 0$, $M_j \geqq 0$, $j \in \mathbf{I}$. The eigenvalue equation

$$\lambda^* M^* = \sum_{\substack{i \in \mathbf{I} \\ j \in \mathbf{I}}} \sigma_i^2 F_i \int_0^\infty \exp(A\tau) F_j N_j F_j^T \exp(A^T\tau)\, d\tau\, F_i^T$$

leads to

$$\lambda^* M^* = \sum_{\substack{i \in \mathbf{I} \\ j \in \mathbf{I}}} \sigma_i^2 F_i \int_0^\infty \exp(A\tau) F_j F_j^+ F_j N_j F_j^T F_j^{+T} F_j^T \exp(A^T\tau)\, d\tau\, F_i^T$$

where F_j^+ denotes the generalized inverse [12, pp. 142–144] of the matrix F_j. This yields

$$\lambda^* M^* = \sum_{\substack{i \in \mathbf{I} \\ j \in \mathbf{I}}} \sigma_i^2 F_i \int_0^\infty \exp(A\tau) F_j F_j^+ M_j F_j^{+T} F_j^T \exp(A^T\tau)\, d\tau\, F_i^T$$

and

$$|\lambda| \leqq \|M^*\|^{-1} \sum_{\substack{i \in \mathbf{I} \\ j \in \mathbf{I}}} \sigma_i^2 F_j^{+\,2} \int_0^\infty \|F_i \exp(A\tau) F_j\|^2\, d\tau\, \|M_j\|.$$

The matrix M^* can be taken to be of unit norm; since $M^* = \sum M_j$ and since the matrices M_j are symmetric and nonnegative definite, then $\|M_j\| \leqq 1$, $j \in \mathbf{I}$. Hence the eigenvalues of the mapping L_2 are smaller than 1 in modulus if the constants α_{ij} are sufficiently small. □

Appendix B. In this appendix the following problem is investigated: let A, B, G, H, respectively, be $(n \times n)$, $(n \times m)$, $(n \times q)$, $(p \times n)$ matrices; we want to state conditions on these matrices such that for all $\varepsilon > 0$ there exists an $(m \times n)$ feedback matrix K such that

(i) $$\sigma(A + BK) \subset \mathbb{C}_g,$$

(ii) $$\int_0^\infty \|W_K(t)\|^2\, dt \leqq \varepsilon$$

where

$$W_K : t \in R^+ \mapsto H \exp[(A + BK)t] G.$$

This property is called *impulse response quenching in the $\mathscr{L}_2$-sense with internal asymptotic stability*. It is well known that a constant K exists such that (i) is true and $W_K = 0$ if and only if

$$\operatorname{im} G \subset \mathscr{V}_g^*(\operatorname{Ker} H).$$

If it is only required that (ii) hold, i.e., that W_K can be made arbitrarily small in the $\mathscr{L}_2$-sense, then one could expect two refinements:

(i) $\mathscr{V}_g^*(\operatorname{Ker} H)$ may be replaced by $\mathscr{V}_{\bar{g}}^*(\operatorname{Ker} H)$ since by a small feedback the eigenvalues can be shifted from the imaginary axis into the left half plane.

(ii) im G may be allowed to have a component in $\mathscr{R}_b^*(\operatorname{Ker} H)$, since in that space it is possible to make W_K arbitrarily small by high gain feedback [4].

The following result is indeed obtained.

THEOREM B.1. *Impulse response quenching in the $\mathscr{L}_2$-sense with internal asymptotic stability is possible if*

(i) (A, B) *is stabilizable (relative* $\mathbb{C}_g$),

(ii) $\operatorname{im} G \subset \mathscr{R}_b^*(\operatorname{Ker} H) + \mathscr{V}_{\bar{g}}^*(\operatorname{Ker} H)$.

The proof of this theorem proceeds via a number of propositions and lemmas:

PROPOSITION B.1. *Assume that (A, B) is stabilizable (relative $\mathbb{C}_g$) and that $x_0 \in \mathscr{V}_{\bar{g}}^*(\operatorname{Ker} H)$. Consider now*

$$J(x_0) := \inf \int_0^\infty \|y(t)\|^2 \, dt$$

subject to: $\dot{x} = Ax + Bu$, $y = Hx$, $x(0) = x_0$, $u \in \mathscr{L}_2(0, \infty)$, $x \in \mathscr{L}_2(0, \infty)$. *Then* $J(x_0) = 0$.

In order to prove this proposition, we start with a lemma.

LEMMA B.1. *Assume (A, B) controllable and $\sigma(A) \subset \{s \in \mathbb{C} | \operatorname{Re}(s) = 0\}$. Then*

$$\lim_{t_f \to \infty} W^{-1}(0, t_f) = 0$$

where

$$W(0, t_f) := \int_0^{t_f} \exp(-A\sigma) B B^T \exp(-A^T\sigma) \, d\sigma.$$

Proof. It suffices to prove that $a^T W(0, t_f) a = M_{t_f} \|a\|^2$ with $\lim_{t_f \to \infty} M_{t_f} = \infty$. By controllability of (A, B) there is a $\delta > 0$ such that

$$\int_0^1 \|B^T \exp(-A^T\sigma) a\|^2 \, d\sigma \geqq \delta \|a\|^2.$$

Now, since $\sigma(A) \subset \{s \in \mathbb{C} | \operatorname{Re}(s) = 0\}$, the solutions of $\dot{x} = -A^T x$ have the property that there exists $T > 1$ such that $\|x(T)\|^2 \geqq \|x(0)\|^2$ (to see this, assume A in Jordan form: if A is semisimple, it is immediate, otherwise it follows from some simple estimates). This yields

$$\int_0^{NT} \|B^T \exp(-A^T\sigma) a\|^2 \, d\sigma \geqq N\delta \|a\|^2.$$

This yields the desired growth of $W(0, t_f)$. □

LEMMA B.2. *Assume (A, B) controllable, x_0 given and $\sigma(A) \subset \{s \in \mathbb{C} | \operatorname{Re}(s) = 0\}$. Then, for all $\varepsilon > 0$, there exist $T > 0$ and $u \in \mathscr{L}_2(0, \infty)$ such that the solution of $\dot{x} = Ax + Bu$, $x(0) = x_0$, satisfies $x(T) = 0$ and $\|u\|_{\mathscr{L}_2(0,\infty)} \leqq \varepsilon$.*

Proof. Consider, for t_f fixed, $J(x_0) := \min \int_0^{t_f} \|u(t)\|^2 \, dt$ subject to $\dot{x} = Ax + Bu$, $x(0) = x_0$, $x(t_f) = 0$. It is well known (see [13, p. 137]) that $J(x_0) = x_0^T W^{-1}(0, t_f) x_0$. The result follows then from Lemma B.1. □

Proof of Proposition B.1. Since (A, B) is stabilizable (relative $\mathbb{C}_g$), there exists K such that

$$(A + BK)\mathscr{V}_{\bar{g}}^*(\operatorname{Ker} H) \subset \mathscr{V}_{\bar{g}}^*(\operatorname{Ker} H),$$

$$\sigma[(A + BK) | \mathscr{V}_{\bar{g}}^*(\operatorname{Ker} H)] \subset \mathbb{C}_g,$$

$$\sigma[(A + BK)(\operatorname{mod} \mathscr{V}_{\bar{g}}^*(\operatorname{Ker} H))] \subset \mathbb{C}_g.$$

By suitably choosing the basis, this yields

$$\mathscr{X} = \mathscr{X}_1 \oplus \mathscr{X}_2, \qquad A + BK = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$$

with

$$\sigma(A_1) \subset \{s \in \mathbb{C} | \operatorname{Re}(s) = 0\}, \quad \sigma(A_2) \subset \mathbb{C}_g, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad H = [0 \quad H_2].$$

Furthermore, since (A, B) is stabilizable, (A_1, B_1) will be controllable. From $x_0 \in \mathscr{V}_g^*(\operatorname{Ker} H)$ it follows that $H \exp[(A+BK)t]x_0$ vanishes for all t. Lemma B.2 implies that for all $\varepsilon > 0$, there exists u and $T > 0$ such that $\|u\|_{\mathscr{L}_2(0,\infty)} \leqq \varepsilon$, $x(0) = x_0$, and $x(T) \in \mathscr{X}_2$. This yields

$$y(t) = H_2 \int_0^t \exp[A_2(t-\tau)] B_2 u(\tau)\, d\tau$$

which, since $\sigma(A_2) \subset \mathbb{C}_g$, is the convolution of an $\mathscr{L}_1$-kernel $\{t \mapsto \exp(A_2 t)B\}$ with an arbitrary small $u \in \mathscr{L}_2(0, \infty)$. Hence y is arbitrarily small in the $\mathscr{L}_2$-norm. It is also immediate that the corresponding $x \in \mathscr{L}_2(0, \infty)$. This yields $J(x_0) = 0$, as desired. □

PROPOSITION B.2. *Assume* $x_0 \in \mathscr{R}_b^*(\operatorname{Ker} H)$. *Consider now* $J(x_0) := \inf \int_0^\infty \|y(t)\|^2\, dt$, *subject to* $\dot{x} = Ax + Bu$; $y = Hx$; $x(0) = x_0$, $u \in \mathscr{L}_2(0, \infty)$, $x \in \mathscr{L}_2(0, \infty)$. *Then* $J(x_0) = 0$.

Proof. That $\int_0^\infty \|y(t)\|^2\, dt$ vanishes without the constraints $u \in \mathscr{L}_2(0, \infty)$ and $x \in \mathscr{L}_2(0, \infty)$ follows immediately from [4, Thm. 10]. However, it is easily seen by examining the proof that the u and x used for showing that this infimum is zero are indeed $\mathscr{L}_2$-functions. This yields the proposition. □

Proof of Theorem B.1. Consider the least squares control for the system $\dot{x} = Ax + Bu$ with cost functional, with $\varepsilon > 0$

$$\int_0^\infty [\varepsilon(\|u\|^2 + \|x\|^2) + \|Hx\|^2]\, dt.$$

Let $J_\varepsilon(x_0)$ be the optimal cost with $x_0 = x(0)$ and $u = K_\varepsilon x$ the optimal control law. From Proposition B.1 it follows that $\lim_{\varepsilon \to 0} J_\varepsilon(x_0) = 0$ for $x_0 \in \mathscr{V}_g^*(\operatorname{Ker} H)$ and from Proposition B.2 this follows for $x_0 \in \mathscr{R}_b^*(\operatorname{Ker} H)$. Since (A, B) is stabilizable (relative $\mathbb{C}_g$), $u = K_\varepsilon x$ is an asymptotically stabilizing control law with $J_\varepsilon(x_0) \geqq \int_0^\infty \|Hx\|^2\, dt$ arbitrarily small for $\varepsilon \downarrow 0$ and $x_0 \in \mathscr{R}_b^*(\operatorname{Ker} H) + \mathscr{V}_g^*(\operatorname{Ker} H)$. This yields the theorem. □

COROLLARY B.1. *Simultaneous quenching of the impulse responses*

$$H_i \exp[(A+BK)t] G_j \qquad (i \in \mathbf{k}, j \in \mathbf{l})$$

in the $\mathscr{L}_2$*-sense, with internal asymptotic stability (by means of a common feedback matrix* K*) is possible if*

(i) (A, B) *is stabilizable (relative* $\mathbb{C}_g$*),*

(ii) $\sum_{j \in \mathbf{l}} \operatorname{im} G_j \subset \mathscr{R}_b^*(\bigcap_{i \in \mathbf{k}} \operatorname{Ker} H_i) + \mathscr{V}_g^*(\bigcap_{i \in \mathbf{k}} \operatorname{Ker} H_i)$.

This corollary is an immediate consequence of Theorem B.1. The next result follows directly from Corollary B.1 and Lemma A.3 in Appendix A.

COROLLARY B.2. *Consider the linear mapping*

$$M \to L_K(M) := \sum_{i \in \mathbf{l}} \sigma_i^2 F_i \int_0^\infty \exp[(A+BK)\tau] M \exp[(A+BK)^T \tau]\, d\tau\, F_i^T$$

in the space of $(n \times n)$ *symmetric matrices. Then for all* $\{\sigma_i\}$ *there exists a matrix* K *such that the eigenvalues of* $L_K(M)$ *are smaller than* 1 *in modulus if*

$$\sum_{i\in\mathbf{I}} \operatorname{im} F_i \subset \mathscr{R}_b^*\Big(\bigcap_{i\in\mathbf{I}} \operatorname{Ker} F_i\Big) + \mathscr{V}_g^*\Big(\bigcap_{i\in\mathbf{I}} \operatorname{Ker} F_i\Big).$$

Appendix C. In this appendix a question similar to that in Appendix B is considered, but now with respect to the $\mathscr{L}_1$-norm. With the same notations we say that *impulse response quenching in the* $\mathscr{L}_1$*-sense with internal asymptotic stability* is possible if for all $\varepsilon > 0$ there exists a feedback matrix K such that

(i) $$\sigma(A+BK) \subset \mathbb{C}_g,$$

(ii) $$\int_0^\infty \|W_K(t)\| dt \leqq \varepsilon.$$

The obtained condition is slightly stronger than the criterion of Theorem B.1; it is expressed by the following result:

THEOREM C.1. *Impulse response quenching in the* $\mathscr{L}_1$*-sense with internal asymptotic stability is possible if*

(i) (A, B) *is stabilizable (relative* $\mathbb{C}_g$*),*

(ii) $\operatorname{im} G \subset \mathscr{R}_b^*(\operatorname{Ker} H) + \mathscr{V}_g^*(\operatorname{Ker} H)$.

Proof. (i) It may be shown that there exists an (A, B)-invariant subspace $\mathscr{V}_1$ and a matrix K_1 such that $(A+BK_1)\mathscr{V}_1 \subset \mathscr{V}_1$, $\sigma((A+BK_1)|\mathscr{V}_1) \subset \mathbb{C}_g$, and

$$\mathscr{R}_b^*(\operatorname{Ker} H) + \mathscr{V}_g^*(\operatorname{Ker} H) = \mathscr{R}_b^*(\operatorname{Ker} H) \oplus \mathscr{V}_1.$$

(ii) By the results of [4, Thm. 12] there exists an (A, B)-invariant $\mathscr{R}_\varepsilon$ and a matrix K_ε such that $\mathscr{R}_\varepsilon \to_{\varepsilon\to 0} \mathscr{R}_b^*(\operatorname{Ker} H)$, $(A+BK_\varepsilon)\mathscr{R}_\varepsilon \subset \mathscr{R}_\varepsilon$, $\sigma((A+BK_\varepsilon)|\mathscr{R}_\varepsilon) \subset \mathbb{C}_g$, and

$$\int_0^\infty \|H \exp[(A+BK_\varepsilon)t]G'\| \, dt \leqq \varepsilon$$

where G': $\operatorname{im} G \cap \mathscr{R}_\varepsilon \to \mathscr{X}$ is the canonical injection.

(iii) Let $K : \mathscr{X} \to \mathscr{U}$ be defined by $K|\mathscr{V}_1 = K_1|\mathscr{V}_1$, $K|\mathscr{R}_\varepsilon = K_\varepsilon|\mathscr{R}_\varepsilon$ and $\sigma(A+BK) \subset \mathbb{C}_g$. The stabilizability of (A, B) guarantees the existence of such a K. Also

$$\int_0^\infty \|H \exp[(A+BK)t]G\| \, dt = \int_0^\infty \|H \exp[(A+BK_\varepsilon)t]G'\| \, dt \leqq \varepsilon$$

which yields Theorem C.1. □

Notice the difference between the conditions (ii) in Theorems B.1 and C.1. In the former case $\operatorname{im} G$ should lie in the *almost stabilizable almost* (A, B)*-invariant* subspace "contained" in $\operatorname{Ker} H$; in the latter case $\operatorname{im} G$ should be part of the *stabilizable almost* (A, B)*-invariant* subspace "contained" in $\operatorname{Ker} H$. It is not possible to replace condition (ii) of Theorem C.1 by the slightly weaker condition (ii) of Theorem B.1. This is illustrated by the following example:

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad H = [0 \quad 1].$$

Then

$$\mathscr{V}_g^*(\operatorname{Ker} H) = \operatorname{im}\begin{bmatrix}1\\0\end{bmatrix}, \quad \mathscr{V}_g^*(\operatorname{Ker} H) = \{0\}, \quad \mathscr{R}_b^*(\operatorname{Ker} H) = \begin{bmatrix}1\\1\end{bmatrix}.$$

This shows that condition (ii) of Theorem C.1 is not satisfied; it is shown in § 3.3 that impulse response quenching in the $\mathscr{L}_1$-sense with internal asymptotic stability is not possible. On the other hand, condition (ii) of Theorem B.1 holds, and impulse response quenching in the $\mathscr{L}_2$-sense is possible.

COROLLARY C.1. *Simultaneous quenching of the impulse responses*

$$H_i \exp[(A+BK)t]G_j \qquad (i \in \mathbf{k}, j \in \mathbf{l})$$

in the $\mathscr{L}_1$-sense with internal asymptotic stability (by means of a common feedback matrix K) is possible if

(i) *(A, B) is stabilizable (relative $\mathbb{C}_g$),*

(ii) $\sum_{j\in\mathbf{l}} \operatorname{im} G_j \subset \mathscr{R}_b^*(\bigcap_{i\in\mathbf{k}} \operatorname{Ker} H_i) + \mathscr{V}_g^*(\bigcap_{i\in\mathbf{k}} \operatorname{Ker} H_i)$.

The result of Corollary C.1 can be used to derive a condition for stabilizability of the nonlinear time-varying system

$$\dot{x}(t) = Ax(t) + Bu(t) + [F_1 \quad F_2 \quad \cdots \quad F_l]M(x(t), t)\begin{bmatrix}F_1\\F_2\\\vdots\\F_l\end{bmatrix}x(t) \tag{C.1}$$

with $x \in \mathbb{R}^n$. The matrices F_i are square $(n \times n)$ matrices. The gain matrix $M(x, t)$ is of dimension $(ln \times ln)$. Let a linear time-invariant feedback $u(t) = Kx(t)$ be applied to this system. Then, according to the small loop theorem [14], the system is $\mathscr{L}_p$-input-output-stable if

(i) $\sigma(A+BK) \subset \mathbb{C}_g$,

(ii) $\|M(x, t)\| < \alpha \quad \forall x, \quad \forall t$,

(iii) $\max_{i,j\in\mathbf{l}} \int_0^\infty \|F_i \exp[(A+BK)t]F_j\|\, dt < 1/\alpha l^2$.

Hence $F_i x(t) \in \mathscr{L}_2(0, \infty)$; the Hurwitz character of $A+BK$ then shows that the solution of (C.1) tends to zero as $t \to \infty$ for all initial conditions. Hence the null solution of (C.1) is asymptotically stable in the large. Sufficient conditions for the existence of a feedback matrix K satisfying (i) and (ii) can be derived from Corollary C.1. Consider now the special case that $M(x, t)$ is a block diagonal matrix

$$M(x) = \operatorname{diag}[f_1(x, t)F_1^+ \quad f_2(x, t)F_2^+ \quad \cdots \quad f_l(x, t)F_l^+]$$

when the functions f_i are scalar and F_i^+ denotes the generalized inverse [12, pp. 142–144] of the matrix F_i. Then (C.1) reduces to (14); the following result is hence obtained:

PROPOSITION C.1. *For any $\alpha > 0$ there exists a constant feedback matrix K such that the null solution of*

$$\dot{x}(t) = (A+BK)x(t) + \sum_{i\in\mathbf{l}} f_i(x(t), t)F_i x(t) \tag{C.2}$$

is asymptotically stable in the large for all nonlinear gains satisfying

$$|f_i(x, t)| < \alpha \qquad (i \in \mathbf{l}, \forall x, \forall t)$$

if (A, B) is stabilizable (relative $\mathbb{C}_g$), and if

$$\sum_{i\in\mathbf{l}} \operatorname{im} F_i \subset \mathscr{R}_b^*\Big(\bigcap_{i\in\mathbf{l}} \operatorname{Ker} F_i\Big) + \mathscr{V}_g^*\Big(\bigcap_{i\in\mathbf{l}} \operatorname{Ker} F_i\Big).$$

Appendix D. Following the suggestion of one of the reviewers we have collected in this appendix the relevant facts on (A, B)-invariant and almost (A, B)-invariant subspaces used in this paper. More details may be found in references [2], [3], [4].

Consider the system $\dot{x} = Ax + Bu$ with $x \in \mathscr{X} := \mathbb{R}^n$. A subspace $\mathscr{V} \subset \mathscr{X}$ is said to be an (A, B)-*invariant subspace* if there exists a matrix K such that $\mathscr{V}$ is $(A + BK)$-invariant (i.e. such that $(A+BK)\mathscr{V} \subset \mathscr{V}$). An equivalent property is that $\mathscr{V}$ satisfies

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

where $\mathscr{B} := \operatorname{im} B$.

Let $\mathbf{V}(\mathscr{S})$ denote the set of all (A, B)-invariant subspaces contained in a given subspace $\mathscr{S}$. Then this set is closed under subspace addition, i.e. $\mathscr{V}_1, \mathscr{V}_2 \in \mathbf{V}(\mathscr{S}) \Rightarrow \mathscr{V}_1 + \mathscr{V}_2 \in \mathbf{V}(\mathscr{S})$. Hence there exists a largest (A, B)-invariant subspace in $\mathscr{S}$, which is denoted by $\mathscr{V}^*(\mathscr{S})$. Systematic finite and linear algorithms are available [2, 3] to compute $\mathscr{V}^*(\mathscr{S})$. A related concept, denoted by $\mathscr{V}_g^*(\mathscr{S})$, is defined as follows

$$\mathscr{V}_g^*(\mathscr{S}) := \sup \{\mathscr{V} \in \mathbf{V}(\mathscr{S}) | \exists K \text{ such that } (A + BK)\mathscr{V} \subset \mathscr{V} \text{ and } \sigma(A + BK) \subset \mathbb{C}_g\}.$$

It is easily proven that this subspace is well defined. It is called the largest *stabilizability* subspace contained in $\mathscr{S}$ and is readily computed from $\mathscr{V}^*(\mathscr{S})$ [2]. Finally $\mathscr{V}_{\bar{g}}^*$ is similarly defined with $\bar{\mathbb{C}}_g$ replacing $\mathbb{C}_g$ in the definition.

Let $\mathscr{V}_i$, $i \in \mathbf{k}$, be a family of (A, B)-invariant subspaces. Then, by definition, there exist matrices K_i such that $(A + BK_i)\mathscr{V}_i \subset \mathscr{V}_i$. However, there is no guarantee that there exists a single K such that $(A + BK)\mathscr{V}_i \subset \mathscr{V}_i$ for all $i \in \mathbf{k}$. If this is the case, then the subspaces $\mathscr{V}_i$ are said to be *compatible* (A, B)-invariant subspaces. It is easy to prove that the subspaces $\mathscr{V}_i$ are compatible, for example, if they are nested ($\mathscr{V}_1 \subset \mathscr{V}_2 \subset \cdots \subset \mathscr{V}_k$), but in general compatibility is a difficult matter to verify.

A further generalization leads to controllability subspaces. Thus

$$\mathscr{R}^*(\mathscr{S}) := \sup \{\mathscr{V} \in \mathbf{V}(\mathscr{S}) | \text{for } K \text{ such that } (A + BK)\mathscr{V} \subset \mathscr{V},$$
$$\text{there holds } \langle A + BK | \mathscr{S} \cap \operatorname{im} B \rangle = \mathscr{V}\}.$$

Again, $\mathscr{R}^*(\mathscr{S})$ is well defined. For equivalent definitions and algorithms for computing $\mathscr{R}^*(\mathscr{S})$ we refer the reader to [2].

An interesting generalization of (A, B)-invariance is *almost* (A, B)-*invariance*. These notions have been introduced in [3] and further worked out in [4]. The largest almost (A, B)-invariant subspace contained in a given subspace $\mathscr{S}$ is the subspace of initial states in $\mathscr{S}$ for which there exists an input such that the resulting state trajectory is *almost* contained in $\mathscr{S}$. However, this depends on the topology chosen. In particular, we obtain a somewhat larger subspace if we measure "*almost being contained in*" in the $\mathscr{L}_p$-sense ($1 \leqq p < \infty$) rather than in the $\mathscr{L}_\infty$-sense. Similarly, the largest almost controllability subspace contained in $\mathscr{S}$ is the subspace of initial states in $\mathscr{S}$ which, by means of an input, may be transferred to any terminal state in that subspace, such that the resulting state trajectory is *almost* contained in $\mathscr{S}$. Let $\mathscr{V}_a^*(\mathscr{S})$ and $\mathscr{V}_b^*(\mathscr{S})$ denote respectively the supremal $\mathscr{L}_\infty$-almost-controllability and the $\mathscr{L}_p$-($1 \leqq p < \infty$)-almost-controllability subspace contained in $\mathscr{S}$. Similarly $\mathscr{R}_a^*(\mathscr{S})$ and $\mathscr{R}_b^*(\mathscr{S})$ denote respectively the supremal $\mathscr{L}_\infty$-almost-(A, B)-invariant and the $\mathscr{L}_p$-($1 \leqq p < \infty$)-almost-(A, B)-invariant subspace contained in $\mathscr{S}$.

In the present paper we use primarily $\mathscr{R}_b^*(\mathscr{S})$. We therefore define it formally:

$$\mathscr{R}_b^*(\mathscr{S}) := \sup \mathbf{R}_b(\mathscr{S})$$

where

$$\mathbf{R}_b(\mathcal{S}) := \{\mathcal{R}_b \subset \mathcal{X} | \forall x_0, x_1 \in \mathcal{R}_b \exists T > 0, \text{ such that for all } \varepsilon > 0,$$
$$\text{there exists } x \in \Sigma_x \text{ with the properties: (i) } x(0) = x_0,$$
$$\text{(ii) } x(T) = x_1, \text{ and (iii)} \|d(x(t), \mathcal{S})\|_{\mathcal{L}_1(0,T)} \leqq \varepsilon\}.$$

Here

$$d(x(t), \mathcal{S}) := \inf_{s \in \mathcal{S}} \|x(t) - s\|$$

and

$$\Sigma_x := \{x : \mathbb{R} \to \mathcal{X} | x \text{ is absolutely continuous and } \exists u : \mathbb{R} \to \mathcal{U},$$
$$\text{such that } \dot{x}(t) = Ax(t) + Bu(t) \text{ almost everywhere}\}.$$

This definition merely says that $\mathcal{R}_b^*(\mathcal{S})$ is the largest subspace of $\mathcal{X}$ in which any two states can be transferred to one another while keeping the $\mathcal{L}_1$-norm of the distance of the state trajectory to $\mathcal{S}$ arbitrarily small. This has an obvious interpretation in terms of $\mathcal{L}_1$-(almost) output nulling for the system $\dot{x} = Ax + Bu$, $z = Hx$, with $\mathcal{S} = \operatorname{Ker} H$. The subspace $\mathcal{R}_b^*(\mathcal{S})$ is readily computed by means of the following finite linear recursive algorithm:

$$\mathcal{S}_{k+1} = \operatorname{im} B + A(\mathcal{S} \cap \mathcal{S}_k),$$
$$\mathcal{S}_0 = \{0\}.$$

Then

$$\mathcal{R}_b^*(\mathcal{S}) = \mathcal{S}_\infty := \lim_{k \to \infty} \mathcal{S}_k$$

where this limit is obtained monotonically in at most Min [codim (im B), $1 + \dim(\mathcal{S})$] steps.

The subspace $\mathcal{V}_b^*(\mathcal{S})$ may be defined completely analogously as

$$\mathcal{V}_b^*(\mathcal{S}) := \sup \mathbf{V}_b(\mathcal{S})$$

where

$$\mathbf{V}_b(\mathcal{S}) = \{\mathcal{V}_b \subset \mathcal{X} | \forall x_0 \in \mathcal{V}_b \text{ and } \varepsilon > 0 \exists x \in \Sigma_x \text{ with the properties:}$$
$$\text{(i) } x(0) = x_0 \text{ and (ii) } \|d(x(t), \mathcal{S})\|_{\mathcal{L}_1(0,\infty)} \leqq \varepsilon\}.$$

In [4, Thm. 10] it is proven that $\mathcal{V}_b^*(\mathcal{S}) = \mathcal{V}^*(\mathcal{S}) + \mathcal{R}_b^*(\mathcal{S})$. Its main use in feedback system synthesis stems from the following result [4, Thm. 12].

THEOREM D.1. *Consider the finite dimensional linear system* $\dot{x} = Ax + Bu$, $z = Hx$. *Let* $1 \leqq p < \infty$. *Then for all* $\varepsilon > 0$ *there exists a matrix* K *such that*

$$\int_0^\infty \|H e^{(A+BK)t} G\| \, dt \leqq \varepsilon$$

if and only if

$$\operatorname{im} G \subset \mathcal{V}_b^*(\operatorname{Ker} H).$$

This theorem is also the basic tool for our results on robust stabilizability. However, a number of refinements were needed (Theorems B.1 and C.1). We note in closing that, because $1 \leqq p < \infty$, $\mathcal{R}_b^*(\mathcal{S})$ and $\mathcal{V}_b^*(\mathcal{S})$ need not be contained in $\mathcal{S}$. This fact is amply discussed in [4].

REFERENCES

[1] J. L. WILLEMS AND J. C. WILLEMS, *Feedback stabilizability for stochastic systems with state and control dependent noise*, Automatica, 12 (1976), pp. 277–283.

[2] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, 2nd ed., Springer-Verlag, New York, 1979.

[3] J. C. WILLEMS, *Almost A(mod B)-invariant subspaces*, Astérisque, 75–76 (1980), pp. 239–248.

[4] ———, *Almost invariant subspaces: An approach to high gain feedback design—Part* I: *Almost controlled invariant subspaces*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 235–252; *Part* II: *Almost conditionally invariant subspaces*, IEEE Trans. Automat. Control, AC-27 (1982), to appear.

[5] P. MOLANDER, *Stabilization of uncertain systems*, Report LUTFD2/(TFRT-1020)/I-111/(1979), Dept. Automatic Control, Lund Institute of Technology, Lund, Sweden, 1979.

[6] M. ATHANS, R. KU AND S. B. GERSHWIN, *The uncertainty threshold principle: Some fundamental limitations of optimal decision making under dynamic uncertainty*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 491–495.

[7] A. A. SAGLE AND R. E. WALDE, *Introduction to Lie Groups and Lie Algebras*, Academic Press, New York, 1973.

[8] 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.

[9] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[10] J. L. WILLEMS, *Mean square stability criteria for linear white noise stochastic systems*, Problems Control Inform. Theory, 2 (1973), pp. 199–217.

[11] J. C. WILLEMS AND G. C. BLANKENSHIP, *Frequency domain stability criteria for stochastic systems*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 292–299.

[12] B. NOBLE, *Applied Linear Algebra*, Prentice-Hall, Englewood Cliffs, NJ, 1969.

[13] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.

[14] C. A. DESOER AND M. VIDYASAGAR, *Feedback Systems: Input–Output Properties*, Academic Press, New York, 1975.

[15] J. E. HUMPHREYS, *Introduction to Lie Algebras and Representation Theory*, Springer-Verlag, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0003 $01.25/0

GEOMETRY OF THE ALGEBRAIC RICCATI EQUATION, PART I*

MARK A. SHAYMAN†

Abstract. We synthesize and generalize the two principal methods for classifying the set of real symmetric solutions of the algebraic Riccati equation (ARE) to obtain a result which combines the advantages of both existing methods. The geometric approach we take also clarifies the roles of controllability and observability in the theory of the ARE. In addition, we discuss the properties of a special subset of the solution set of the ARE, the *unmixed solutions*, and find that they exhibit many of the same properties as the extremal solutions, K^+ and K^-.

Key words. algebraic Riccati equation, linear quadratic control, Hamiltonian matrices

1. Introduction. In this paper we consider the algebraic Riccati equation (ARE) $-A'K-KA+KBB'K-Q=0$, where A, B and Q are real matrices of dimensions $n\times n$, $n\times m$ and $n\times n$, respectively, and $Q=Q'$. (We use prime to denote matrix transpose.) For the most part we shall be concerned with the set of real symmetric solutions of the ARE which we denote by Γ. In [19] (this issue, pp. 395–409), we discuss the topological properties of Γ. In contrast, this paper considers methods for describing Γ at the set-theoretic level.

There are two principal methods for classifying the set of solutions of the ARE. The first method relates the set of real symmetric solutions of the ARE to a certain class of n-dimensional invariant subspaces of the associated $2n\times 2n$ Hamiltonian matrix

$$H\equiv\begin{bmatrix} A & -BB' \\ -Q & -A' \end{bmatrix}.$$

For this reason we refer to it as the Hamiltonian matrix method. Its development is due to many people including A. G. J. MacFarlane [10], J. E. Potter [14], K. Mårtensson [11] and A. C. M. Van Swieten [22].

The second method of classifying the set of real symmetric solutions of the ARE was introduced by J. C. Willems [23] and was extended by W. Coppel [5]. If K^+ denotes the maximal element of Γ, then this approach relates Γ to the set of invariant subspaces of the matrix $A-BB'K^+$. We will refer to this approach as Willems' method.

The Hamiltonian matrix method and Willems' method are both very well known. However, to the best of our knowledge, the relationship between the two approaches has never been described. As we indicate in the next section, the two methods have different strengths. Thus, it is desirable to have a classification theorem which combines the two results. This is the principal motivation for this paper. We obtain a classification of the set of real symmetric solutions of the ARE which both unifies and extends the two existing methods. In the process of doing so, we also describe the geometric roles of controllability and observability in the theory of the ARE, and show that they are precisely dual. In addition, we describe the properties of a special subset of Γ consisting of what we call *unmixed* solutions and show that these solutions share several of the useful properties of the extremal solutions K^+ and K^-. This is particularly interesting because the concept of unmixed solutions extends to more general (even nonsquare) ARE's than the equation $-A'K-KA+KBB'K-Q=0$ of linear least squares stationary optimal control, whereas the concept of extremal solutions does not generalize.

* Received by the editors July 11, 1981, and in revised form May 10, 1982. This research was partially supported by the U.S. Army Office of Research under grant DAAG 29-79-C-0147.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

These "non-Hamiltonian" ARE's are useful in various engineering applications, such as in the theory of singularly perturbed systems [13].

2. Comparison of existing classifications. In this section we discuss the advantages and disadvantages of the two principal classification methods.

The following theorem describes the Hamiltonian matrix method for classifying Γ. The version of this result which we state is essentially that contained in [22].

THEOREM 1. *There is a one-to-one correspondence between the set of real solutions of the* ARE *and the set of n-dimensional H-invariant subspaces which are complementary to the n-dimensional subspace* $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$. (Sp *denotes the column span of a matrix.*) *This correspondence assigns the invariant subspace* $S(K)\equiv\operatorname{Sp}\begin{bmatrix}I\\K\end{bmatrix}$ *to the solution K. The matrix of the restriction of H to* $S(K)$ *with respect to the basis given by the columns of* $\begin{bmatrix}I\\K\end{bmatrix}$ *is* $A-BB'K$. *Furthermore, K is symmetric if and only if* $x'Jy=0$, *for all* $x, y\in S(K)$, *where J is the* $2n\times 2n$ *matrix* $\begin{bmatrix}0 & I\\ -I & 0\end{bmatrix}$.

Proof. If S is an n-dimensional subspace which is complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$, then there exists an $n\times n$ matrix K such that $S=\operatorname{Sp}\begin{bmatrix}I\\K\end{bmatrix}$. If S is also H-invariant, then there exists an $n\times n$ matrix R such that

$$\begin{bmatrix}A & -BB'\\ -Q & -A'\end{bmatrix}\begin{bmatrix}I\\K\end{bmatrix}=\begin{bmatrix}I\\K\end{bmatrix}R.$$

The top equation implies that $R=A-BB'K$. Substituting for R in the second equation then gives $-Q-A'K=K(A-BB'K)$, which shows that K is a solution of the ARE. Conversely, if K satisfies the ARE, then

$$\begin{bmatrix}A & -BB'\\ -Q & -A'\end{bmatrix}\begin{bmatrix}I\\K\end{bmatrix}=\begin{bmatrix}I\\K\end{bmatrix}(A-BB'K),$$

which shows that $S(K)$ is H-invariant and that $A-BB'K$ is the matrix of $H|S(K)$ with respect to the basis given by the columns of $\begin{bmatrix}I\\K\end{bmatrix}$. Finally note that

$$\begin{bmatrix}I & K'\end{bmatrix}\begin{bmatrix}0 & I\\ -I & 0\end{bmatrix}\begin{bmatrix}I\\K\end{bmatrix}=0$$

if and only if $K=K'$, which proves the last assertion. □

If S is a subspace of $\mathbb{R}^{2n}$ such that $x'Jy=0$, for all $x, y\in S$, then we call S a *Lagrangian* subspace. (The term "Lagrangian" is sometimes reserved for n-dimensional subspaces which satisfy the indicated condition, but we do not make this restriction.) Theorem 1 gives a one-to-one correspondence between Γ and the set of n-dimensional Lagrangian H-invariant subspaces which are complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$.

Before discussing the second principal classification, Willems' method, we consider the relationship between Theorem 1 and a result due to J. Rodriguez-Canabal [16], [17]. This result extends a method due to R. W. Bass and W. E. Roth [1], [4], [18] and is based on the factorization of the characteristic polynomial of H. Let $h(s)$ be the characteristic polynomial of H, and let K be a real symmetric solution of the ARE. Then Rodriguez-Canabal's theorem states that there exists a polynomial $g(s)$ such that $h(s)=(-1)^n g(s)g(-s)$ and $g(H)\begin{bmatrix}I\\K\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$. To see why this is true, let K be any real symmetric solution of the ARE. Then

$$\begin{bmatrix}I & 0\\K & I\end{bmatrix}^{-1}\begin{bmatrix}A & -BB'\\ -Q & -A'\end{bmatrix}\begin{bmatrix}I & 0\\K & I\end{bmatrix}=\begin{bmatrix}A-BB'K & -BB'\\ 0 & -(A-BB'K)'\end{bmatrix}.$$

This equation implies that $\operatorname{Sp}\begin{bmatrix}I\\K\end{bmatrix}$ is H-invariant and that $A-BB'K$ is the matrix of $H|\operatorname{Sp}\begin{bmatrix}I\\K\end{bmatrix}$ with respect to the basis given by the columns of $\begin{bmatrix}I\\K\end{bmatrix}$. It also shows that if

we let $g(s)$ be the characteristic polynomial of $A-BB'K$, then $h(s)=(-1)^n g(s)g(-s)$. Since $g(s)$ is the characteristic polynomial of $H|\operatorname{Sp}\begin{bmatrix}I\\K\end{bmatrix}$, $g(H)$ annihilates every vector in $\operatorname{Sp}\begin{bmatrix}I\\K\end{bmatrix}$. Thus, $g(H)\begin{bmatrix}I\\K\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$.

In contrast to Theorem 1, the Rodriguez-Canabal result is not a one-to-one correspondence. First, given a polynomial $g(s)$ satisfying $h(s)=(-1)^n g(s)g(-s)$, there might not be any real symmetric solution K of the ARE satisfying $g(H)\begin{bmatrix}I\\K\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$. From Theorem 1, we see that this will occur if and only if there is no n-dimensional Lagrangian H-invariant subspace which is complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$ and whose minimal annihilating polynomial divides $g(s)$. Secondly, given $g(s)$ satisfying $h(s)=(-1)^n g(s)g(-s)$, there can be more than one real symmetric solution K satisfying $g(H)\begin{bmatrix}I\\K\end{bmatrix}=\begin{bmatrix}0\\0\end{bmatrix}$. This occurs if and only if there is more than one n-dimensional Lagrangian H-invariant subspace which is complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$ and whose minimal annihilating polynomial divides $g(s)$. For examples of both possibilities (nonexistence and nonuniqueness) see [17].

It is well known [23] that if (A,B) is controllable and Γ is nonempty, then Γ contains a unique element K^+ (K^-) such that every eigenvalue of $A-BB'K^+$ ($A-BB'K^-$) has nonpositive (nonnegative) real part. K^+ and K^- have the additional property that if $K\in\Gamma$, then $K^-\leqq K\leqq K^+$ with respect to the usual partial ordering of symmetric matrices. For this reason, K^+ and K^- are called the *extremal* solutions of the ARE.

We introduce some notation. Let R be a linear operator on $\mathbb{R}^n$ with characteristic polynomial $p(s)=p^+(s)p^0(s)p^-(s)$, where the roots of $p^+(s)$ ($p^0(s)$) ($p^-(s)$) have negative (zero) (positive) real parts. Then $L^+(R)$ ($L^0(R)$) ($L^-(R)$) denotes the sum of the primary components of R which correspond to its left half-plane (imaginary axis) (right half-plane) eigenvalues—i.e., $L^+(R)\equiv\ker p^+(R)$, $L^0(R)\equiv\ker p^0(R)$, $L^-(R)\equiv\ker p^-(R)$.

Let $\Delta\equiv K^+-K^-$. It can be shown [5] that for any $K\in\Gamma$, $L^0(A-BB'K)=\ker\Delta$ and that $A-BB'K|\ker\Delta=A-BB'K^+|\ker\Delta$. Let $V_+\equiv L^+(A-BB'K^+)$, $V_0\equiv L^0(A-BB'K^+)$ ($=\ker\Delta$), and let $V_-\equiv L^-(A-BB'K^-)$. Note that $L^-(A-BB'K^+)=0=L^+(A-BB'K^-)$. Let $T\equiv A-BB'K^+|V_+$, and let S_T be the set of all invariant subspaces of T. The next theorem describes the second principal method for classifying the set of real symmetric solutions of the ARE. It is W. Coppel's extension [5] of a result of J. C. Willems [23].

THEOREM 2 (J. C. Willems, W. Coppel). *Let (A,B) be controllable and suppose that Γ is nonempty. If S is any $(A-BB'K^+)$-invariant subspace which is contained in V_+ (i.e., $S\in S_T$), then $S\oplus\Delta^{-1}(S^\perp)=\mathbb{R}^n$, where $S^\perp$ denotes the orthogonal complement of S in $\mathbb{R}^n$ and $\Delta^{-1}(S^\perp)$ is its inverse image. There is a bijection $\gamma:S_T\to\Gamma$ defined by $\phi(S)\equiv K^+P_S+K^-(I-P_S)$, where P_S is the projection onto S along $\Delta^{-1}(S^\perp)$. If $K=\phi(S)$, then $L^+(A-BB'K)=S$, $L^0(A-BB'K)=V_0$ and $L^-(A-BB'K)=\Delta^{-1}(S^\perp)\cap V_-$.*

Remark 1. If $\Delta>0$ (or equivalently, if H has no pure imaginary eigenvalues), then $V_+=\mathbb{R}^n$ so $T=A-BB'K^+$, and Theorem 2 reduces to the original theorem of Willems.

In our view, the Hamiltonian matrix method has two extremely attractive features. The first feature is that it easily generalizes to non-Hamiltonian Riccati equations. If P is $m\times p$ and B is the $(p+m)\times(p+m)$ partitioned matrix

$$\begin{bmatrix}B_{11} & B_{12}\\ B_{21} & B_{22}\end{bmatrix},$$

the set of solutions of the quadratic matrix equation $B_{21}+B_{22}P-PB_{11}-PB_{12}P=0$ is in one-to-one correspondence with the set of p-dimensional B-invariant subspaces

which are complementary to the m-dimensional subspace $\mathrm{Sp}\begin{bmatrix}0\\ I_m\end{bmatrix}$. The correspondence associates the solution P with the subspace $\mathrm{Sp}\begin{bmatrix}I\\ P\end{bmatrix}$. Of course, if B is not a Hamiltonian matrix, it is a misnomer to call this the Hamiltonian matrix method.

The second attractive feature of the Hamiltonian matrix method is that it permits us to make precise the idea of solutions at infinity. One of the phenomena of the Riccati differential equation which is difficult to understand is that of finite escape time. A standard mathematical technique for studying a differential equation whose solutions "blow up" is to compactify the phase space. The reason why this is useful is that every vector field on a compact manifold is complete [2], so finite escape times do not occur. For the Riccati differential $\dot{K} = -A'K - KA + KBB'K - Q$, the phase space is the vector space of real symmetric $n \times n$ matrices. The most convenient compactification of this space is the Lagrangian Grassmannian $\mathscr{L}$. It consists of all n-dimensional Lagrangian subspaces of $\mathbb{R}^{2n}$. The subset $\mathscr{L}_c$ of $\mathscr{L}$ consisting of those subspaces which are complementary to $\mathrm{Sp}\begin{bmatrix}0\\ I\end{bmatrix}$ is a chart of $\mathscr{L}$ which is open and dense.

R. Hermann [7] and C. Martin [12] have showed that every Riccati differential equation $\dot{K} = -A'K - KA + KBB'K - Q$ on the space of real symmetric $n \times n$ matrices is the expression in local coordinates of the restriction to $\mathscr{L}_c$ of a vector field on $\mathscr{L}$. The equilibrium points of the flow on $\mathscr{L}$ consist of those elements of $\mathscr{L}$ which are H-invariant. However, only those H-invariant subspaces which belong to $\mathscr{L}_c$ correspond to equilibrium points of the original Riccati differential equation on the space of $n \times n$ symmetric matrices. The points in $\mathscr{L}$ which are outside $\mathscr{L}_c$ can be viewed as the points at infinity in the Lagrangian Grassmannian. Thus, the n-dimensional Lagrangian H-invariant subspaces which are *not* complementary to $\mathrm{Sp}\begin{bmatrix}0\\ I\end{bmatrix}$ represent solutions of the ARE at infinity. In the next section, we show that the concept of solutions at infinity is necessary to understand the duality between controllability and observability in the theory of the ARE.

In our view, the major disadvantage of the Hamiltonian matrix method is that it fails to exploit the Hamiltonian symmetry of the equation $-A'K - KA + KBB'K - Q = 0$. In other words, it does not use the fact that H is a Hamiltonian matrix. (A Hamiltonian matrix is a $2n \times 2n$ matrix Z such that $JZ + Z'J = 0$.) This is precisely the reason why the Hamiltonian matrix method generalizes to non-Hamiltonian Riccati equations.

We have described elsewhere [19], [21] a simple parametrization for the set of all invariant subspaces of fixed dimension of an arbitrary finite dimensional linear operator. Using this result, we can describe all the n-dimensional H-invariant subspaces. However, it is not at all clear how the two additional conditions, Lagrangian and complementarity, determine a subset of the set of all n-dimensional H-invariant subspaces. In the next section, we show that the assumption of controllability makes the issue of complementarity trivial. Later we will see that it is possible to describe precisely which n-dimensional H-invariant subspaces are Lagrangian. To do this, we must use the Hamiltonian symmetry of the ARE.

Willems' method is quite different from the Hamiltonian matrix method. It relates the set of real symmetric solutions of the ARE to the set of *all* invariant subspaces of $T(= A - BB'K^{+} | V_{+})$. There are no additional requirements corresponding to the Lagrangian and complementarity conditions of the Hamiltonian matrix method. This is an attractive feature of Willems' method because it enables us to use results on the algebraic variety of invariant subspaces of a finite dimensional linear operator to describe the geometric structure of Γ. These results are detailed in [19], [21]. The disadvantages of Willems' method are that there is no obvious generalization to the non-Hamiltonian Riccati equation and there is no concept of a solution at infinity.

3. Controllability and observability.

PROPOSITION 1. *Let* $M \equiv \ker [B, AB, \cdots, A^{n-1}B]'$. *Let* $N \equiv \{[{}^{0}_{y}] \in \mathbb{R}^{2n} : y \in M\}$. *Then* N *is the largest subspace of* $\operatorname{Sp}[{}^{0}_{I}]$ *which is* H*-invariant.*

Proof. Let $y \in M$. Then

$$H\begin{bmatrix}0\\ y\end{bmatrix} = \begin{bmatrix}A & -BB'\\ -Q & -A'\end{bmatrix}\begin{bmatrix}0\\ y\end{bmatrix} = \begin{bmatrix}-BB'y\\ -A'y\end{bmatrix}.$$

Since $y \in M$, it follows that $B'y = 0$ and $A'y \in M$, so $H[{}^{0}_{y}] \in N$, showing that N is H-invariant. Let S be any H-invariant subspace which is contained in $\operatorname{Sp}[{}^{0}_{I}]$, and suppose that S is k-dimensional. Let $[{}^{0}_{Y}]$ be a basis matrix for S. The H-invariance of S implies that there exists a $k \times k$ matrix P such that

$$H\begin{bmatrix}0\\ Y\end{bmatrix} = \begin{bmatrix}-BB'Y\\ -A'Y\end{bmatrix} = \begin{bmatrix}0\\ Y\end{bmatrix}P = \begin{bmatrix}0\\ YP\end{bmatrix}.$$

This shows that $B'Y = 0$ and $A'Y = -YP$, which implies that $B'Y = 0$, $B'A'Y = 0, \cdots, B'(A')^{n-1}Y = 0$, so $\operatorname{Sp} Y \subseteq M$. Hence, $S \subseteq N$. □

COROLLARY. (A, B) *is controllable if and only if* $\operatorname{Sp}[{}^{0}_{I}]$ *contains no nontrivial* H*-invariant subspace.*

The next proposition is excerpted from the proof of a theorem by V. Kučera [8] which concerns stabilizability.

PROPOSITION 2 (Kučera). *Let* S *be an* n*-dimensional Lagrangian* H*-invariant subspace. Then* $S \cap \operatorname{Sp}[{}^{0}_{I}]$ *is* H*-invariant.*

Proof. Let $[{}^{X}_{Y}]$ be a basis matrix for S. Since S is Lagrangian, $[X'\ Y']J[{}^{X}_{Y}] = 0$, which implies that $Y'X = X'Y$. Since S is H-invariant, there exists an $n \times n$ matrix R such that

$$\begin{bmatrix}A & -BB'\\ -Q & -A'\end{bmatrix}\begin{bmatrix}X\\ Y\end{bmatrix} = \begin{bmatrix}X\\ Y\end{bmatrix}R.$$

In particular, this implies that $AX - BB'Y = XR$. Let $z \in \ker X$. Then $0 = Y'Xz = X'Yz$. Now, $z'Y'AXz - z'Y'BB'Yz = z'Y'XRz$, which implies that $B'Yz = 0$. Then $0 = AXz - BB'Yz = XRz$, so $Rz \in \ker X$. Thus $\ker X$ is R-invariant. Now,

$$S \cap \operatorname{Sp}\begin{bmatrix}0\\ I\end{bmatrix} = \left\{\begin{bmatrix}X\\ Y\end{bmatrix}z : z \in \ker X\right\}.$$

Since $H[{}^{X}_{Y}] = [{}^{X}_{Y}]R$, the R-invariance of $\ker X$ implies the H-invariance of $S \cap \operatorname{Sp}[{}^{0}_{I}]$. □

The following theorem is an immediate consequence of Propositions 1 and 2.

THEOREM 3. *Let* (A, B) *be controllable. Then every* n*-dimensional Lagrangian* H*-invariant subspace is complementary to* $\operatorname{Sp}[{}^{0}_{I}]$.

Remark 2. Theorem 3 shows that if (A, B) is controllable, then the ARE has no solutions at infinity. In other words, the compactification of the Riccati flow in the Lagrangian Grassmannian introduces no extraneous equilibrium points.

Remark 3. V. Kučera has proved [8] that if H has no pure imaginary eigenvalues and (A, B) is stabilizable, then the n-dimensional Lagrangian H-invariant subspace $L^{+}(H)$ is complementary to $\operatorname{Sp}[{}^{0}_{I}]$. ($L^{+}(H)$ is the sum of the primary components of H corresponding to its left half-plane eigenvalues.) By assuming that (A, B) is controllable rather than merely stabilizable, we obtain the much stronger result that

every n-dimensional Lagrangian H-invariant subspace is complementary to $\mathrm{Sp}\begin{bmatrix}0\\I\end{bmatrix}$. In the next section, we give a converse of Theorem 3.

Propositions 1 and 2 and Theorem 3 have interesting dual versions in the special case where $Q = C'C$ for some $p \times n$ matrix C. To emphasize the complete duality, we replace the submatrix $-BB'$ in H with an arbitrary symmetric matrix $-L$.

PROPOSITION 3. *Let* $\tilde{H} \equiv \begin{bmatrix} A & -L \\ -C'C & -A' \end{bmatrix}$. *Let* $\tilde{M} \equiv \ker [C', A'C', \cdots, (A')^{n-1}C']'$. *Let* $\tilde{N} \equiv \{\begin{bmatrix} x \\ 0 \end{bmatrix} \in \mathbb{R}^{2n} : x \in \tilde{M}\}$. *Then* $\tilde{N}$ *is the largest subspace of* $\mathrm{Sp}\begin{bmatrix}I\\0\end{bmatrix}$ *which is* $\tilde{H}$*-invariant.*

COROLLARY. (C, A) *is observable if and only if* $\mathrm{Sp}\begin{bmatrix}I\\0\end{bmatrix}$ *contains no nontrivial* $\tilde{H}$*-invariant subspace.*

PROPOSITION 4. *Let* S *be an* n*-dimensional Lagrangian* $\tilde{H}$*-invariant subspace. Then* $S \cap \mathrm{Sp}\begin{bmatrix}I\\0\end{bmatrix}$ *is* $\tilde{H}$*-invariant.*

The proofs of Proposition 3 and 4 are completely dual to the proofs of Propositions 1 and 2 and are therefore omitted. The following theorem is an immediate consequence of Propositions 3 and 4.

THEOREM 4. *Let* (C, A) *be observable. Then every* n*-dimensional Lagrangian* $\tilde{H}$*-invariant subspace is complementary to* $\mathrm{Sp}\begin{bmatrix}I\\0\end{bmatrix}$.

We have already noted that those elements of $\mathscr{L}$ which are not complementary to $\mathrm{Sp}\begin{bmatrix}0\\I\end{bmatrix}$ correspond to the points at infinity in $\mathscr{L}$. Similarly, it makes sense to view those elements of $\mathscr{L}$ which are not complementary to $\mathrm{Sp}\begin{bmatrix}I\\0\end{bmatrix}$ as the "points at zero" in the Lagrangian Grassmannian. The validity of this interpretation is especially clear for the subset $\mathscr{L}_c$ of $\mathscr{L}$. If $S \in \mathscr{L}_c$—i.e., S is complementary to $\mathrm{Sp}\begin{bmatrix}0\\I\end{bmatrix}$—then S can be expressed as $\mathrm{Sp}\begin{bmatrix}I\\K\end{bmatrix}$ for some symmetric $n \times n$ matrix K. Then the condition that S intersect $\mathrm{Sp}\begin{bmatrix}I\\0\end{bmatrix}$ is equivalent to K being singular. Thus, the elements of $\mathscr{L}_c$ which intersect $\mathrm{Sp}\begin{bmatrix}I\\0\end{bmatrix}$ are "points at zero" in the sense that they are the graphs of singular symmetric matrices.

Theorem 4 shows that if (C, A) is observable, then none of the equilibrium points of the Riccati flow in the Lagrangian Grassmannian are points at zero. In particular, if S is an equilibrium point of the form $\mathrm{Sp}\begin{bmatrix}I\\K\end{bmatrix}$, then K is nonsingular. Since the equilibrium points of the form $\mathrm{Sp}\begin{bmatrix}I\\K\end{bmatrix}$ are precisely the equilibrium points which correspond to solutions of the original (i.e., uncompactified) ARE, this shows that every solution of the ARE is nonsingular. We state this result as a theorem.

THEOREM 5. *Let* (C, A) *be observable. Then every real symmetric solution of the* ARE $-A'K - KA + KLK - C'C = 0$ *is nonsingular and hence defines a nondegenerate quadratic form.*

Remark 4. It is of course well known [3] that if (A, B, C) is a minimal triple, then the stabilizing solution K^+ of $-A'K - KA + KBB'K - C'C = 0$ exists and is positive definite. However, the result in Theorem 5, that observability implies that every solution is nonsingular, does not seem to be recognized.

Remark 5. K. Mårtensson [11] has proven several results which are related to those of this section. However, his results on the role of controllability and observability were derived under the assumption that A has distinct eigenvalues. A result which is equivalent to Theorem 3 has been discovered independently by P. Lancaster and L. Rodman and appears in a recent paper [9]. In the special case where L is negative semidefinite, Theorem 5 follows from a result of H. Wimmer [24]. However, his proof does not extend to the general case where L is only symmetric.

Remark 6. Comparing Theorems 3 and 4, we conclude that controllability and observability have precisely dual roles in the geometry of the ARE. Controllability guarantees that no n-dimensional Lagrangian H-invariant subspace intersects $\mathrm{Sp}\begin{bmatrix}0\\I\end{bmatrix}$, which means that the ARE has no solutions at infinity. Observability guarantees that no n-dimensional Lagrangian H-invariant subspace intersects $\mathrm{Sp}\begin{bmatrix}I\\0\end{bmatrix}$, which means that

the ARE has no solutions at zero. Note that this duality cannot be appreciated without having the precise meaning of points at infinity which is made possible by the compactification.

We close this section by considering the relationship between the controllability of (A, B) and the quadratic character of the ARE $-A'K - KA + KBB'K - Q = 0$. Suppose that the reachable subspace has codimension $r > 0$. By changing basis in $\mathbb{R}^n$, we may assume that

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{bmatrix}$$

with A_{11} $r \times r$, A_{22} $(n-r)\times(n-r)$ and that $B = \begin{bmatrix} 0 \\ B_2 \end{bmatrix}$ with B_2 $(n-r)\times m$ and that (A_{22}, B_2) is controllable. Let

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}' & Q_{22} \end{bmatrix} \quad \text{and} \quad K = \begin{bmatrix} K_{11} & K_{12} \\ K_{12}' & K_{22} \end{bmatrix}.$$

From the partitioned ARE which results, we obtain four matrix equations, but two of these are equivalent due to symmetry. Only one of these equations involves K_{11} and this equation is $-A_{11}'K_{11} - K_{11}A_{11} - A_{21}'K_{12}' - K_{12}A_{21} + K_{12}B_2B_2'K_{12}' - Q_{11} = 0$. This shows that the $\frac{1}{2}r(r+1)$ variables in the symmetric matrix K_{11} enter the ARE *linearly*. Hence, if (A, B) is not controllable, then the ARE is not genuinely quadratic.

We can also interpret this from a geometric viewpoint. In [19] we note that if (A, B) is controllable and Γ is nonempty, then Γ is compact. Consider the linear matrix equation $A_{11}'K_{11} + K_{11}A_{11} = 0$. This equation has nontrivial solutions if and only if A_{11} has zero as an eigenvalue or has a pair of eigenvalues λ_i and λ_j such that $\lambda_i + \lambda_j = 0$. If this is the case, then the solution set of the linear equation is a vector subspace of positive dimension. Since the eigenvalues of A_{11} are the uncontrollable eigenvalues of A, the following theorem is an immediate consequence.

THEOREM 6. *Suppose that Γ is nonempty and that (A, B) is not controllable. If 0 is an uncontrollable eigenvalue of A or there is a pair of uncontrollable eigenvalues λ_i, λ_j such that $\lambda_i + \lambda_j = 0$, then Γ is not compact.*

We can also interpret this result in the context of the Hamiltonian matrix method. It is not hard to show that the set of all n-dimensional Lagrangian H-invariant subspaces is compact. If (A, B) is controllable, then each of these subspaces is complementary to $\operatorname{Sp}\begin{bmatrix} 0 \\ I \end{bmatrix}$ and hence corresponds to a solution of the ARE. Thus, Γ is compact. On the other hand, if (A, B) is not controllable, some of the n-dimensional Lagrangian H-invariant subspaces may intersect $\operatorname{Sp}\begin{bmatrix} 0 \\ I \end{bmatrix}$. The remaining n-dimensional Lagrangian H-invariant subspaces need not form a compact set. Thus, in the absence of controllability, Γ need not be compact. Of course, the most extreme example which illustrates this is the case where $B = 0$. Then the ARE is a linear matrix equation, so if its solution set Γ contains more than one element, then Γ is an affine subspace and therefore not compact.

4. Unmixed solutions. In this section, we define a special subset of Γ which we call the *unmixed* solutions. We describe several of their properties which resemble the properties of the extremal solutions K^+ and K^-.

The first lemma is a tool which will be used to establish the existence of n-dimensional Lagrangian H-invariant subspaces. It appears in Van Swieten's thesis [22] and extends a result due to Potter [14] and Mårtensson [11].

LEMMA 1. *Let $\tilde{H}$ be a Hamiltonian matrix and let $p(s)$ be its characteristic polynomial. Suppose that $p(s) = p_1(-s)p_2(s)$ with $p_1(-s)$ and $p_2(s)$ relatively prime. Let $S_1 \equiv \ker p_1(\tilde{H})$ and let $S_2 \equiv \ker p_2(\tilde{H})$. Then $x_2'Jx_1 = 0$, for all $x_1 \in S_1$, $x_2 \in S_2$.*

Proof. Let X_1 and X_2 be basis matrices for S_1 and S_2. Since S_1 and S_2 are $\tilde{H}$-invariant, there exist square matrices R_1, R_2 (of the appropriate dimensions) such that $\tilde{H}X_1 = X_1R_1$ and $\tilde{H}X_2 = X_2R_2$. Suppose that $p_2(s)$ contains a positive power of an irreducible polynomial $g(s)$. Then $p_1(-s)$ does not contain $g(s)$ as a factor. It follows that the primary component of $\mathbb{R}^{2n}$ (relative to $\tilde{H}$) corresponding to $g(s)$ is completely contained in S_2. Since this is true for each irreducible factor in $p_2(s)$, S_2 is a sum of *whole* primary components of $\mathbb{R}^{2n}$ (relative to $\tilde{H}$). This implies that the characteristic polynomial of $\tilde{H}|S_2$ is $p_2(s)$, so $p_2(s) = \det(Is - R_2)$. Since $\tilde{H}$ is Hamiltonian, $p(s) = p(-s)$, so $p(s) = p_1(s)p_2(-s)$. Then the above argument applied to $p_1(s)$ shows that $p_1(s) = \det(Is - R_1)$.

Since $J\tilde{H} = -\tilde{H}'J$, it follows that $X_2'J\tilde{H}X_1 = -X_2\tilde{H}'JX_1$, which implies that $X_2'JX_1R_1 = -R_2'X_2'JX_1$. Letting $Z \equiv X_2'JX_1$ yields the equation $ZR_1 + R_2'Z = 0$. Since $p_1(s) = \det(Is - R_1)$ and $p_2(s) = \det(Is - R_2)$, R_1 and $-R_2'$ have no eigenvalues in common. This implies that the only solution is $Z = 0$. Hence, $X_2'JX_1 = 0$, which completes the proof. □

In the remainder of this section and in the next section, we make the assumption that H has no eigenvalues on the imaginary axis. In § 6, we describe what changes are necessary when H has pure imaginary eigenvalues.

Since H is a real Hamiltonian matrix, its spectrum is symmetrical with respect to both coordinate axes. Let $\mu_1, -\mu_1, \cdots, \mu_p, -\mu_p$ be the distinct real eigenvalues of H ($\mu_i > 0$), and let $\lambda_1, \bar{\lambda}_1, -\lambda_1, -\bar{\lambda}_1, \cdots, \lambda_q, \bar{\lambda}_q, -\lambda_q, -\bar{\lambda}_q$ be the distinct nonreal eigenvalues of H ($\operatorname{Re}\lambda_j > 0, \operatorname{Im}\lambda_j > 0$). Let E_i and E_{-i} be the generalized eigenspaces corresponding to μ_i and $-\mu_i$ respectively. Let F_j and F_{-j} be the primary components corresponding to the pairs of eigenvalues $\lambda_j, \bar{\lambda}_j$ and $-\lambda_j, -\bar{\lambda}_j$ respectively. Since H is a Hamiltonian matrix, $\dim E_i = \dim E_{-i}$ and $\dim F_j = \dim F_{-j}$ $(i = 1, \cdots, p; j = 1, \cdots, q)$.

Let Λ be the set of n-dimensional H-invariant subspaces of the form $E_{\pm 1} \oplus \cdots \oplus E_{\pm p} \oplus F_{\pm 1} \oplus \cdots \oplus F_{\pm q}$. In other words, an element of Λ is obtained by choosing E_1 or $E_{-1}, \cdots, E_p$ or E_{-p}, F_1 or $F_{-1}, \cdots, F_q$ or F_{-q} and summing the ones that are chosen. Λ clearly contains 2^{p+q} elements. If $L_1 \in \Lambda$, then there exists a unique subspace $L_2 \in \Lambda$ such that $H|L_1$ and $H|L_2$ have no eigenvalues in common. L_2 is obtained by making the opposite choices from those made in forming L_1. We say that L_1 and L_2 are a *pair of opposites in* Λ. Λ contains 2^{p+q-1} such pairs.

PROPOSITION 5. *Every element of* Λ *is Lagrangian.*

Proof. Let $L_1 \in \Lambda$. Let $p(s)$ be the characteristic polynomial of H, and let $p_1(s)$ be the characteristic polynomial of $H|L_1$. Then $p(s) = (-1)^n p_1(s)p_1(-s)$ with $p_1(s)$ and $p_1(-s)$ relatively prime. Applying Lemma 1 in the special case where $p_2(s) = p_1(s)$ gives $x'Jy = 0$, for all $x, y \in L_1$, so L_1 is Lagrangian. □

COROLLARY. *If* (A, B) *is controllable, then every element of* Λ *corresponds to a real symmetric solution of the* ARE.

Proof. Every element of Λ is an n-dimensional Lagrangian H-invariant subspace. If (A, B) is controllable, then every such subspace is complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$ and hence corresponds to a solution of the ARE. □

For the remainder of this section, we will assume that (A, B) is controllable unless otherwise stated. Let $L \in \Lambda$, and let K be the real symmetric solution of the ARE which corresponds to L—i.e., $L = \operatorname{Sp}\begin{bmatrix}I\\K\end{bmatrix}$. By Theorem 1, the matrix of $H|L$ with respect to the basis given by the columns of $\begin{bmatrix}I\\K\end{bmatrix}$ is $A - BB'K$. If $p(s)$ is the characteristic polynomial of H and $p_1(s)$ is the characteristic polynomial of $H|L$, then it follows from the definition of Λ that $p(s) = (-1)^n p_1(s)p_1(-s)$ with $p_1(s)$ and $p_1(-s)$ relatively prime. Since $p_1(s) = \det(Is - (A - BB'K))$, this means that if λ is an eigenvalue of H,

then either $A-BB'K$ has the eigenvalue λ with maximum multiplicity and the eigenvalue $-\lambda$ with zero multiplicity or it has the eigenvalue $-\lambda$ with maximum multiplicity and λ with zero multiplicity. For this reason, we call K an *unmixed* solution. There are exactly 2^{p+q} unmixed solutions. Note that if $L=E_{-1}\oplus\cdots\oplus E_{-p}\oplus F_{-1}\oplus\cdots\oplus F_{-q}$, then $K=K^+$, whereas if $L=E_1\oplus\cdots\oplus E_p\oplus F_1\oplus\cdots\oplus F_q$, then $K=K^-$, so the extremal solutions are unmixed.

The unmixed solutions are interesting because they resemble the extremal solutions in several respects. For one thing, they occur in pairs. If L_1 and L_2 are a pair of opposites in Λ and K_1 and K_2 are the corresponding elements of Γ, we say that K_1 and K_2 are a *pair of opposite unmixed solutions*. K_2 is the unique element of Γ with the property that $A-BB'K_1$ and $A-BB'K_2$ have no eigenvalues in common.

In [19] we show that the unmixed solutions are always isolated points in Γ. In fact, they are the only solutions which are guaranteed to be isolated no matter what Jordan block structure the matrix H has.

Another interesting property of the unmixed solutions is their analytic dependence on the coefficient matrices in the ARE. Let $\hat{\Sigma}$ be the open subset of $\mathbb{R}^{n^2}\times\mathbb{R}^{nm}\times\mathbb{R}^{pn}$ consisting of all minimal triples (A,B,C) with A $n\times n$, B $n\times m$ and C $p\times n$. Let $S(n)$ be the vector space of $n\times n$ symmetric matrices, and let $\psi:\hat{\Sigma}\to S(n)$ be the mapping which takes a minimal triple (A,B,C) to the unique symmetric positive definite solution K^+ of the ARE $-A'K-KA+KBB'K-C'C=0$. By using the implicit function theorem, D. Delchamps showed [6] that ψ is real-analytic. In other words, K^+ depends analytically on the matrices A, B and C. Using the same tool, we are able to prove the following more general result.

THEOREM 7. *Let P_0 be an $m\times p$ matrix which satisfies the equation $B_{21}^0+B_{22}^0P-PB_{11}^0-PB_{12}^0P=0$, where B_{11}^0, B_{12}^0, B_{21}^0, B_{22}^0 are real matrices of dimensions $p\times p$, $p\times m$, $m\times p$ and $m\times m$, respectively. Suppose that $B_{11}^0+B_{12}^0P_0$ and $B_{22}^0-P_0B_{12}^0$ have no eigenvalues in common. Then there exists a neighborhood U of $(B_{11}^0,B_{12}^0,B_{21}^0,B_{22}^0)$ in $\mathbb{R}^{p^2+pm+mp+m^2}$ and a unique mapping $\psi:U\to\mathbb{R}^{mp}$ such that*

$$\psi(B_{11}^0,B_{12}^0,B_{21}^0,B_{22}^0)=P_0$$

and

$$B_{21}+B_{22}\psi(B_{11},B_{12},B_{21},B_{22})-\psi(B_{11},B_{12},B_{21},B_{22})B_{11}$$
$$-\psi(B_{11},B_{12},B_{21},B_{22})B_{12}\psi(B_{11},B_{12},B_{21},B_{22})=0,$$
$$\forall(B_{11},B_{12},B_{21},B_{22})\in U.$$

Furthermore, the mapping ψ is real-analytic.

Proof. Define $\eta:\mathbb{R}^{p^2+pm+mp+m^2}\times\mathbb{R}^{mp}\to\mathbb{R}^{mp}$ by

$$\eta(B_{11},B_{12},B_{21},B_{22},P)\equiv B_{21}+B_{22}P-PB_{11}-PB_{12}P.$$

Let $M\in S(n)$. Then

$$\lim_{t\to 0}\frac{1}{t}[\eta(B_{11}^0,B_{12}^0,B_{21}^0,B_{22}^0,P_0+tM)-\eta(B_{11}^0,B_{12}^0,B_{21}^0,B_{22}^0,P_0)]$$
$$=(B_{22}^0-P_0B_{12}^0)M+M(-B_{11}^0-B_{12}^0P_0).$$

The eigenvalues of the linear map ϕ of $\mathbb{R}^{mp}$ into $\mathbb{R}^{mp}$ given by $M\overset{\phi}{\mapsto}(B_{22}^0-P_0B_{12}^0)M+M(-B_{11}^0-B_{12}^0P_0)$ are $\{\alpha_i+\beta_j\}$, where $\{\alpha_i\}_{i=1}^m$ and $\{\beta_j\}_{j=1}^p$ are the eigenvalues of $B_{22}^0-P_0B_{12}^0$ and $-B_{11}^0-B_{12}^0P_0$, respectively. Since $B_{11}^0+B_{12}^0P_0$ and $B_{22}^0-P_0B_{12}^0$ have no eigenvalues in common, ϕ is an isomorphism. The assertions of the theorem follow immediately from the implicit function theorem and the analyticity of η.

COROLLARY. *Suppose that* (A^0, B^0) *is controllable and that*

$$H^0 \equiv \begin{bmatrix} A^0 & -B^0B^{0\prime} \\ -Q^0 & -A^{0\prime} \end{bmatrix}$$

has no pure imaginary eigenvalues. Let K_0 *be an unmixed solution of the* ARE $-A^{0\prime}K - KA^0 + KB^0B^{0\prime}K - Q^0 = 0$. *Then there exists a neighborhood* U *of* (A^0, B^0, Q^0) *in* $\mathbb{R}^{n^2+nm} \times S(n)$ *and a unique mapping* $\psi: U \to S(n)$ *such that* $\psi(A^0, B^0, Q^0) = K_0$ *and* $-A'\psi(A, B, Q) - \psi(A, B, Q)A + \psi(A, B, Q)BB'\psi(A, B, Q) - Q = 0$, *for all* $(A, B, Q) \in U$. *Furthermore,* ψ *is real-analytic.*

Proof. Since L_0 is an unmixed solution, $A^0 - B^0B^{0\prime}K_0$ and $-(A^0 - B^0B^{0\prime}K_0)'$ have no eigenvalues in common. Applying Theorem 7 with $B^0_{11} = A^0$, $B^0_{12} = -B^0B^{0\prime}$, $B^0_{21} = -Q^0$ and $B^0_{22} = -A^{0\prime}$, the result follows immediately. □

Remark 7. If $Q^0 = C^{0\prime}C^0$ with (A^0, B^0, C^0) minimal, then H^0 has no pure imaginary eigenvalues, so the hypothesis of the preceding corollary is satisfied. Thus, if K^0 is an unmixed solution of $-A^{0\prime}K - KA^0 + KB^0B^{0\prime}K - C^{0\prime}C^0 = 0$, then K^0 depends analytically on the coefficient matrices in a neighborhood of (A^0, B^0, C^0). It is important to note that this is strictly a local result except in the case where K^0 is either K^+ or K^-. Since K^+ and K^- exist for every minimal triple $(A, B, C) \in \hat{\Sigma}$, K^+ and K^- are real-analytic functions defined on all of $\hat{\Sigma}$. These are the only unmixed solutions which exist globally since we can choose $(A, B, C) \in \hat{\Sigma}$ such that $\left[\begin{smallmatrix} A & -BB' \\ -C'C & -A' \end{smallmatrix}\right]$ has only two distinct eigenvalues, μ and $-\mu$. Then the only unmixed solutions of $-A'K - KA + KBB'K - C'C = 0$ are K^+ and K^-.

Using the results of this section, we can prove converses to the controllability and observability theorems in § 3. By Theorem 1, the real symmetric solutions of the ARE are in one-to-one correspondence with the n-dimensional Lagrangian H-invariant subspaces which are complementary to $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$. This is true whether or not (A, B) is controllable. If we assume that H has no pure imaginary eigenvalues (as we have in this section), then H always has n-dimensional Lagrangian invariant subspaces. In particular, every element of Λ is such a subspace. However, in the absence of controllability, there is no guarantee that any of these subspaces are complementary to $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$. Thus, the compactified Riccati flow always has equilibrium points, but without controllability some or all of these may be at infinity. In fact, if (A, B) is not controllable and if H has no pure imaginary eigenvalues, then the Riccati flow has at least 2^{p+q-1} equilibrium points at infinity. This follows as a corollary of the following theorem, which provides a converse to Theorem 3.

THEOREM 8. *Suppose that* H *has no pure imaginary eigenvalues and that* (A, B) *is not controllable. Let* L_1 *and* L_2 *be any pair of opposites in* Λ. *Then* L_1 *and* L_2 *are not both complementary to* $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$.

Proof. Let $M \equiv \ker [B, AB, \cdots, A^{n-1}B]'$. Let $N \equiv \{\left[\begin{smallmatrix} 0 \\ y \end{smallmatrix}\right] \in \mathbb{R}^{2n} : y \in M\}$. By Proposition 1, N is H-invariant. Since L_1 and L_2 are each sums of primary components of H such that $\mathbb{R}^{2n} = L_1 \oplus L_2$, it follows that $N = (N \cap L_1) \oplus (N \cap L_2)$. Since (A, B) is not controllable, N is not trivial, so $N \cap L_1$ and $N \cap L_2$ cannot both be zero. Since N is contained in $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$, L_1 and L_2 are not both complementary to $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$. □

COROLLARY. *If* H *has no pure imaginary eigenvalues and* (A, B) *is not controllable, then there are at least* 2^{p+q-1} n*-dimensional Lagrangian* H*-invariant subspaces which are not complementary to* $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$.

Proof. Since Λ contains 2^{p+q-1} pairs of opposites, the result follows immediately. □

Remark 8. There is another way to prove Theorem 8. Suppose that L_1 and L_2 are both complementary to $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$. Then there exist $K_1, K_2 \in \Gamma$ which correspond to

L_1, L_2 respectively. Since H has no pure imaginary eigenvalues and K_1 and K_2 are a pair of opposite unmixed solutions, $A-BB'K_1$ and $A-BB'K_2$ have no eigenvalues in common. But this means that every eigenvalue of A can be shifted by state feedback. By a well-known result [25, p. 52], (A,B) must be controllable.

Theorem 8 has a dual result regarding observability.

THEOREM 9. *Let* $\tilde{H} \equiv \left[\begin{smallmatrix} A & -L \\ -C'C & -A' \end{smallmatrix}\right]$. *Suppose that* $\tilde{H}$ *has no pure imaginary eigenvalues and that* (C, A) *is not observable. Let* L_1 *and* L_2 *be any pair of opposites in* Λ. *Then* L_1 *and* L_2 *are not both complementary to* $\operatorname{Sp}\left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$.

COROLLARY. *If* L_1 *and* L_2 *are complementary to* $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$ *and hence correspond to solutions* K_1 *and* K_2 *of the* ARE, *then* K_1 *and* K_2 *are not both invertible.*

Throughout this section, we have emphasized the similarity between the properties of the unmixed solutions of the ARE and those of the extremal solutions K^+ and K^-. We are interested in the unmixed solutions for more than just theoretical reasons. The existence of extremal solutions is a consequence of the Hamiltonian symmetry of the usual ARE $-A'K-KA+KBB'K-Q=0$. However, for certain engineering applications, one is interested in the more general ARE $B_{21}+B_{22}P-PB_{11}-PB_{12}P=0$, where the unknown matrix P is $m \times p$ with m and p possibly unequal. For example, this more general ARE is related to the problem of block-triangularizing linear systems, which is of interest in the study of singularly perturbed systems [13]. For this generalized ARE, there is no such thing as extremal solutions, and solutions do not occur naturally in pairs.

However, the concept of an unmixed solution does extend to the generalized ARE. As we noted in § 2, the "Hamiltonian" matrix method is still valid. Thus, there is a one-to-one correspondence between the set of solutions of the generalized ARE and the set of p-dimensional subspaces which are invariant with respect to

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

and complementary to the m-dimensional subspace $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I_m \end{smallmatrix}\right]$. The correspondence associates the solution P with the subspace $\operatorname{Sp}\left[\begin{smallmatrix} I_p \\ P \end{smallmatrix}\right]$. The solutions which correspond to the unmixed solutions of the usual ARE are those P for which $\operatorname{Sp}\left[\begin{smallmatrix} I_p \\ P \end{smallmatrix}\right]$ is a sum of whole primary components of B. This is equivalent to the condition that the restriction $B|\operatorname{Sp}\left[\begin{smallmatrix} I_p \\ P \end{smallmatrix}\right]$ has no eigenvalues in common with the induced map $\bar{B}$ on the quotient $R^{p+m}/\operatorname{Sp}\left[\begin{smallmatrix} I_p \\ P \end{smallmatrix}\right]$. In terms of matrices, this means that $B_{11}+B_{12}P$ and $B_{22}-PB_{12}$ have no eigenvalues in common. It follows from a result in our companion paper [19] that such a solution is always isolated. Also, it follows from Theorem 7 that such a solution depends analytically on the coefficient matrices. Thus, the "unmixed" solutions of the generalized ARE are the most attractive solutions.

5. Classification theorems. In § 2, we noted that the difficulty in using the Hamiltonian matrix method to understand the structure of Γ is due to the two additional conditions, complementarity and Lagrangian. In § 3, we showed that the assumption of controllability renders the question of complementarity trivial. If (A, B) is controllable, then every n-dimensional Lagrangian H-invariant subspace is complementary to $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$. Thus, the remaining problem is to identify those n-dimensional H-invariant subspaces which are Lagrangian.

In the preceding section we saw that the Potter–Mårtensson result (essentially Lemma 1) shows that certain n-dimensional H-invariant subspaces are Lagrangian, namely those which belong to Λ. However, Λ is always a finite set. On the other hand, it is easy to find matrices A, B and Q such that the corresponding ARE has uncountably

many real symmetric solutions. (A simple example is $K^2 - I_2 = 0$, where I_2 is a 2×2 identity matrix.) By Theorem 1, the corresponding Hamiltonian matrix H has uncountably many n-dimensional Lagrangian H-invariant subspaces. Thus, this result falls far short of describing all the n-dimensional Lagrangian H-invariant subspaces.

In this section, we solve this problem by incorporating some of the ideas behind Willems' approach into the framework of the Hamiltonian matrix method. Willems' result (Theorem 2) singles out the pair of extremal solutions K^+ and K^-. In our result, K^+ and K^- are replaced by an arbitrary pair of opposite unmixed solutions. This is yet another way in which the unmixed solutions resemble the extremal solutions. We will assume that (A, B) is controllable and that H has no pure imaginary eigenvalues. We discuss the case where H has pure imaginary eigenvalues in the next section.

Let L_1 and L_2 be any pair of opposites in Λ, and let K_1 and K_2 be the pair of opposite unmixed solutions of the ARE corresponding to L_1 and L_2 (i.e., $L_1 = \operatorname{Sp}\left[\begin{smallmatrix} I \\ K_1 \end{smallmatrix}\right]$, $L_2 = \operatorname{Sp}\left[\begin{smallmatrix} I \\ K_2 \end{smallmatrix}\right]$). Let $\Delta_{12} \equiv K_1 - K_2$, $A_1 \equiv A - BB'K_1$, $A_2 \equiv A - BB'K_2$. Using the fact that K_1 and K_2 are solutions of the ARE, it is easy to check that the following equation holds:

$$(*) \qquad A_1'\Delta_{12} + \Delta_{12}A_2 = 0.$$

LEMMA 2. *Δ_{12} is nonsingular.*

Proof. By $(*)$ it follows that $\ker \Delta_{12}$ is A_2-invariant. From the definition of Δ_{12}, A_1 and A_2 agree on $\ker \Delta_{12}$. Since K_1 and K_2 are a pair of opposite unmixed solutions, A_1 and A_2 have no eigenvalues in common. Hence $\ker \Delta_{12} = 0$. □

Suppose that S is an n-dimensional Lagrangian H-invariant subspace. Since L_1 is a sum of primary components of H and L_2 is the sum of the primary components not in L_1, it follows that $S = (S \cap L_1) \oplus (S \cap L_2)$. Thus, S determines an H-invariant subspace of L_1, namely $S \cap L_1$. The following fundamental result shows that the converse is true—that every H-invariant subspace of L_1 determines a unique n-dimensional Lagrangian H-invariant subspace.

PROPOSITION 6. *Let S_1 be an l-dimensional subspace of L_1. Then there exists a unique $(n-l)$-dimensional subspace S_2 of L_2 such that $S_1 \oplus S_2$ is Lagrangian. If $S_1 = \operatorname{Sp}\left[\begin{smallmatrix} I \\ K_1 \end{smallmatrix}\right] C$, where C is $n \times l$ full rank, then $S_2 = \operatorname{Sp}\left[\begin{smallmatrix} I \\ K_2 \end{smallmatrix}\right] D$, where D is any $n \times (n-l)$ full rank matrix which satisfies $C'\Delta_{12}D = 0$. Furthermore, S_1 is H-invariant if and only if S_2 is H-invariant.*

Proof. There is no loss of generality in taking S_1 to be of the form $\operatorname{Sp}\left[\begin{smallmatrix} I \\ K_1 \end{smallmatrix}\right] C$ with C $n \times l$ full rank. Let S_2 be an arbitrary $(n-l)$-dimensional subspace of L_2. Without loss of generality, let $S_2 \equiv \operatorname{Sp}\left[\begin{smallmatrix} I \\ K_2 \end{smallmatrix}\right] D$ with D $n \times (n-l)$ full rank. Then

$$\begin{bmatrix} C & D \\ K_1C & K_2D \end{bmatrix}' \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} \begin{bmatrix} C & D \\ K_1C & K_2D \end{bmatrix} = \begin{bmatrix} 0 & -C'\Delta_{12}D \\ D'\Delta_{12}C & 0 \end{bmatrix}.$$

Thus, $S_1 \oplus S_2$ is Lagrangian if and only if $C'\Delta_{12}D = 0$. Since Δ_{12} is nonsingular, the condition $C'\Delta_{12}D = 0$ determines $\operatorname{Sp} D$ uniquely. In fact, $\operatorname{Sp} D = \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$. Thus, S_2 is uniquely determined.

Now, it follows from Theorem 1 that $H(S_1) = \operatorname{Sp}\left[\begin{smallmatrix} I \\ K_1 \end{smallmatrix}\right] A_1C$ and $H(S_2) = \operatorname{Sp}\left[\begin{smallmatrix} I \\ K_2 \end{smallmatrix}\right] A_2D$. Hence, S_1 is H-invariant if and only if $\operatorname{Sp} C$ is A_1-invariant, and S_2 is H-invariant if and only if $\Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$ is A_2-invariant. However, it follows from $(*)$ that $\operatorname{Sp} C$ is A_1-invariant if and only if $\Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$ is A_2-invariant. Hence, S_1 is H-invariant if and only if S_2 is H-invariant. □

Since the controllability of (A, B) implies that every n-dimensional Lagrangian H-invariant subspace is complementary to $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$, the following theorem is an immediate consequence of Proposition 6.

THEOREM 10. *There exists a one-to-one correspondence between the set of H-invariant subspaces of* L_1 *and the set of n-dimensional Lagrangian H-invariant subspaces which are complementary to* $\operatorname{Sp}\begin{bmatrix}0\\ I\end{bmatrix}$. *This correspondence associates* $\operatorname{Sp}\begin{bmatrix}I\\ K_1\end{bmatrix}C$ *with* $\operatorname{Sp}\begin{bmatrix}C & D\\ K_1C & K_2D\end{bmatrix}$, *where C is a full rank matrix with* $\operatorname{Sp} C$ A_1*-invariant and D is any full rank matrix such that* $\operatorname{Sp} D = \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$.

Remark 9. Since $\operatorname{Sp}\begin{bmatrix}C & D\\ K_1C & K_2D\end{bmatrix}$ is complementary to $\operatorname{Sp}\begin{bmatrix}0\\ I\end{bmatrix}$, the $n \times n$ submatrix $[C, D]$ is always nonsingular. Since $\operatorname{Sp} D = \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$ this means that $\Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$ is complementary to $\operatorname{Sp} C$. Thus, this shows that if M is any A_1-invariant subspace of $\mathbb{R}^n$, then $\Delta_{12}^{-1}(M^{\perp})$ is complementary to M.

Theorem 1 gives a one-to-one correspondence between the set of real symmetric solutions of the ARE and the set of n-dimensional Lagrangian H-invariant subspaces which are complementary to $\operatorname{Sp}\begin{bmatrix}0\\ I\end{bmatrix}$. Composing this correspondence with that given in Theorem 10 yields the following result.

THEOREM 11. *There exists a one-to-one correspondence between the set of H-invariant subspaces of* L_1 *and the set* Γ *of real symmetric solutions of the* ARE. *This correspondence associates the subspace* $\operatorname{Sp}\begin{bmatrix}I\\ K_1\end{bmatrix}C$ *with the solution* $[K_1C, K_2D][C, D]^{-1}$, *where C is a full rank matrix with* $\operatorname{Sp} C$ A_1*-invariant and D is any full rank matrix such that* $\operatorname{Sp} D = \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$.

Proof. By Theorem 10, the H-invariant subspace of L_1 $\operatorname{Sp}\begin{bmatrix}I\\ K_1\end{bmatrix}C$ corresponds to the n-dimensional Lagrangian H-invariant subspace $\operatorname{Sp}\begin{bmatrix}C & D\\ K_1C & K_2D\end{bmatrix}$. Since $[C, D]$ is nonsingular, this subspace can be expressed as

$$\operatorname{Sp}\begin{bmatrix} I \\ [K_1C, K_2D][C, D]^{-1}\end{bmatrix}.$$

By Theorem 1, this subspace corresponds to the solution $[K_1C, K_2D][C, D]^{-1}$ of the ARE. □

Remark 10. Let $P_1 \in Gl(l, \mathbb{R})$ and let $P_2 \in Gl(n-l, \mathbb{R})$. Then $[K_1C, K_2D]$ $[C, D]^{-1} = [K_1CP_1, K_2DP_2]$ $[CP_1, DP_2]^{-1}$, which shows that the correspondence in Theorem 11 is well defined. In other words, the solution of the ARE depends only on $\operatorname{Sp} C$ and $\Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$ and not on the matrices C and D.

Theorem 11 gives a one-to-one correspondence between the H-invariant subspaces of L_1 and the real symmetric solutions of the ARE. Since the matrix of $H|L_1$ relative to the basis $\begin{bmatrix}I\\ K_1\end{bmatrix}$ is A_1, this can be viewed equivalently as a one-to-one correspondence between the A_1-invariant subspaces of $\mathbb{R}^n$ and the real symmetric solutions of the ARE. Let S_{A_1} be the set of A_1-invariant subspaces of $\mathbb{R}^n$, and let $M \in S_{A_1}$. By Remark 9, $\Delta_{12}^{-1}(M^{\perp})$ is complementary to M. Let P_M be the projection onto M along $\Delta_{12}^{-1}(M^{\perp})$. If C is such that $\operatorname{Sp} C = M$, then $\operatorname{Sp} C = \operatorname{Sp} P_M$, and if $C'\Delta_{12}D = 0$, then $\operatorname{Sp} D = \Delta_{12}^{-1}(M^{\perp}) = \operatorname{Sp}(I - P_M)$. Thus,

$$\operatorname{Sp}\begin{bmatrix} C & D \\ K_1C & K_2D\end{bmatrix} = \operatorname{Sp}\begin{bmatrix} P_M & I-P_M \\ K_1P_M & K_2(I-P_M)\end{bmatrix} = \operatorname{Sp}\begin{bmatrix} I \\ K_1P_M + K_2(I-P_M)\end{bmatrix}.$$

This shows that $[K_1C, K_2D][C, D]^{-1} = K_1P_M + K_2(I - P_M)$. This gives the next result.

THEOREM 12. *There exists a one-to-one correspondence between* S_{A_1} *and* Γ *which associates the solution* $K_1P_M + K_2(I - P_M)$ *with the* A_1*-invariant subspace M, where* P_M *is the projection onto M along* $\Delta_{12}^{-1}(M^{\perp})$.

Remark 11. If S_{A_1} and Γ are each topologized in the natural way, then the correspondence in Theorem 12 is actually a homeomorphism. The proof is analogous to the proof of a similar result in our companion paper [19].

Theorem 12 is a generalization of Willems' classification theorem (Theorem 2) because it shows that if L_1 and L_2 are any pair of opposites in Λ and K_1 and K_2 are the corresponding pair of opposite unmixed solutions of the ARE, then every real symmetric solution can be expressed as $K_1P_M+K_2(I-P_M)$. If $L_1=E_{-1}\oplus\cdots\oplus E_{-p}\oplus F_{-1}\oplus\cdots\oplus F_{-q}$ and $L_2=E_1\oplus\cdots\oplus E_p\oplus F_1\oplus\cdots\oplus F_q$, then $K_1=K^+$ and $K_2=K^-$, so Theorem 12 contains Willems' result as a special case. Willems uses the fact that $\Delta\equiv K^+-K^-$ is positive definite to show that if M is $A-BB'K^+$-invariant, then $\Delta^{-1}(M^\perp)$ is complementary to M. In the general case, $\Delta_{12}\equiv K_1-K_2$ is nondegenerate (i.e., nonsingular) but indefinite. Thus, Willems' proof does not generalize to give our result. It is Theorem 3 which is needed to show that in the presence of controllability, $\Delta_{12}^{-1}(M^\perp)$ is always complementary to M (for $M\in S_{A_1}$), even if Δ_{12} is indefinite.

6. Hamiltonian matrices with imaginary axis eigenvalues. In this section, we extend the results of §§ 4 and 5 to include the case where H has one or more eigenvalues on the imaginary axis. In § 4, we showed that if (A, B) is controllable and H has no pure imaginary eigenvalues, then the set Γ of real symmetric solutions of the ARE is nonempty. This is of course well known. However, if H has pure imaginary eigenvalues, the controllability of (A, B) is not sufficient to guarantee the existence of a real symmetric solution. For this, an additional hypothesis is necessary, which is often stated as a frequency domain condition [23] or as a condition on the elementary divisors of H corresponding to imaginary axis eigenvalues [9].

Suppose that (A, B) is controllable and that Γ is nonempty. As we noted in § 2, Γ contains unique solutions K^+ and K^- with the property that every eigenvalue of $A-BB'K^+$ has nonpositive real part and every eigenvalue of $A-BB'K^-$ has nonnegative real part. Furthermore, K^+ and K^- are extremal solutions in the sense that if $K\in\Gamma$, then $K^-\leqq K\leqq K^+$. Using the notation introduced in § 2, if R is a finite dimensional linear operator, $L^+(R)$ $(L^0(R))$ $(L^-(R))$ denotes the sum of the primary components of R corresponding to its left half-plane (pure imaginary) (right half-plane) eigenvalues.

The following result was noted by Willems [23]. A proof appears in Coppel's paper [5]. As in § 2, $\Delta\equiv K^+-K^-$.

PROPOSITION 7 (J. C. Willems, W. Coppel). *If* $K\in\Gamma$, $L^0(A-BB'K)=\ker\Delta$.

Let $r\equiv\dim\ker\Delta$, and let G be an $n\times r$ full rank matrix such that $\operatorname{Sp}G=\ker\Delta$. Let $N\equiv\operatorname{Sp}\left[\begin{smallmatrix}I\\K^+\end{smallmatrix}\right]\cap\operatorname{Sp}\left[\begin{smallmatrix}I\\K^-\end{smallmatrix}\right]$. If $K\in\Gamma$, the inequalities $K^-\leqq K\leqq K^+$ imply that K agrees with K^+ on $\ker\Delta$. It follows immediately that $N=\operatorname{Sp}\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right]G$ for all $K\in\Gamma$. Our next proposition gives a characterization of N.

PROPOSITION 8. *N is the unique Lagrangian H-invariant subspace of $L^0(H)$ which has dimension equal to $\frac{1}{2}\dim L^0(H)$.*

Proof. If $K\in\Gamma$ and $p(s)$ is the characteristic polynomial of H, it is easy to show that $p(s)=\det[Is-(A-BB'K)]\det[Is+(A-BB'K)']$. This implies that H has twice as many pure imaginary eigenvalues as does $A-BB'K$. Hence, $\dim L^0(H)=2\dim L^0(A-BB'K)=2r$, so $\dim N=r=\frac{1}{2}\dim L^0(H)$.

By Theorem 1, the matrix of $H|\operatorname{Sp}\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right]$ with respect to the basis given by the columns of $\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right]$ is $A-BB'K$. By Proposition 7, $L^0(A-BB'K)=\ker\Delta=\operatorname{Sp}G$. Hence, $\operatorname{Sp}\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right]\cap L^0(H)=L^0(H|\operatorname{Sp}\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right])=\operatorname{Sp}\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right]G=N$. Since $\operatorname{Sp}\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right]$ and $L^0(H)$ are H-invariant, so is N. Since $\operatorname{Sp}\left[\begin{smallmatrix}I\\K\end{smallmatrix}\right]$ is Lagrangian, the same is true of N—i.e., $x'Jy=0$, for all $x, y\in N$. Thus, N is a Lagrangian H-invariant subspace of $L^0(H)$ with $\dim N=\frac{1}{2}\dim L^0(H)$.

Suppose that $\tilde{N}$ is another r-dimensional Lagrangian H-invariant subspace of $L^0(H)$. Let $S\equiv L^+(H)\oplus\tilde{N}$. Then S is n-dimensional and is H-invariant since both

$L^+(H)$ and $\tilde{N}$ are H-invariant. To show it is also Lagrangian, suppose that x_1+x_2 and y_1+y_2 are in S with $x_1, y_1 \in L^+(H)$ and $x_2, y_2 \in \tilde{N}$. Then $(x_1+x_2)'J(y_1+y_2) = x_1'Jy_1+x_2'Jy_2+x_1'Jy_2+x_2'Jy_1 = x_1'Jy_2+x_2'Jy_1$ since $L^+(H)$ and $\tilde{N}$ are each Lagrangian. (That $L^+(H)$ is Lagrangian follows from Lemma 1.) However, it is an immediate consequence of Lemma 1 that if $u \in L^+(H)$ and $v \in L^0(H)$, then $u'Jv = 0$. Thus, $x_1'Jy_2 = 0$ and $x_2'Jy_1 = 0$, which shows that $L^+(H) \oplus \tilde{N}$ is Lagrangian. Since (A, B) is controllable, this n-dimensional Lagrangian H-invariant subspace is complementary to $\operatorname{Sp}\begin{bmatrix} 0 \\ I \end{bmatrix}$ and therefore corresponds to some $K \in \Gamma$. But since $H|(L^+(H) \oplus \tilde{N})$ has no right half-plane eigenvalues, the same is true of $A - BB'K$. By the uniqueness of K^+, we must have $K = K^+$, so

$$L^+(H) \oplus \tilde{N} = \operatorname{Sp}\begin{bmatrix} I \\ K^+ \end{bmatrix}.$$

But

$$\operatorname{Sp}\begin{bmatrix} I \\ K^+ \end{bmatrix} = \left[\operatorname{Sp}\begin{bmatrix} I \\ K^+ \end{bmatrix} \cap L^+(H)\right] \oplus \left[\operatorname{Sp}\begin{bmatrix} I \\ K^+ \end{bmatrix} \cap L^0(H)\right] = L^+(H) \oplus N,$$

which implies that $\tilde{N} = N$, proving uniqueness. □

Remark 12. The converse of Proposition 8 is also true. Since H is a Hamiltonian matrix, $\dim L^+(H) = \dim L^-(H)$, which implies that $\dim L^0(H)$ is even. Let $\dim L^0(H) = 2r$, and suppose that there exists an r-dimensional Lagrangian H-invariant subspace N of $L^0(H)$. Then $L^+(H) \oplus N$ is n-dimensional H-invariant, and by the same argument as in the proof of Proposition 8 it is Lagrangian. Thus, if (A, B) is controllable, $L^+(H) \oplus N$ is complementary to $\operatorname{Sp}\begin{bmatrix} 0 \\ I \end{bmatrix}$ and corresponds to a real symmetric solution K of the ARE. (In fact, $K = K^+$.) Hence, Γ is nonempty. Combining Proposition 8 and Remark 12 gives the following theorem.

THEOREM 13. *Suppose that* (A, B) *is controllable. Then the following two conditions are equivalent*:

(i) *There exists a real symmetric solution of the* ARE.

(ii) *There exists a Lagrangian H-invariant subspace* $N \subseteq L^0(H)$ *such that*

$$\dim N = \tfrac{1}{2} \dim L^0(H).$$

If (i) *and* (ii) *hold, the subspace N is unique.*

Suppose that (A, B) is controllable and that Γ is nonempty. Using Theorem 13, we can extend the classification results to the case where H may have pure imaginary eigenvalues. First we generalize the definition of Λ given in § 4. Let $\mu_1, -\mu_1, \cdots, \mu_p, -\mu_p$ be the distinct *nonzero* real eigenvalues of H ($\mu_i > 0$), and let $\lambda_1, \bar{\lambda}_1, -\lambda_1, -\bar{\lambda}_1, \cdots, \lambda_q, \bar{\lambda}_q, -\lambda_q, -\bar{\lambda}_q$ be the distinct nonreal, *nonimaginary* eigenvalues of H ($\operatorname{Re}\lambda_j > 0$, $\operatorname{Im}\lambda_j > 0$). (We do not include the pure imaginary eigenvalues of H on either list.) Let E_i, E_{-i}, F_j, F_{-j} be as defined in § 4. Let N be the subspace described in Theorem 13 and let $r = \dim N$.

We modify the definition of Λ given in § 4. Now we define Λ to be the set of n-dimensional Lagrangian H-invariant subspaces of the form $N \oplus E_{\pm 1} \oplus \cdots \oplus E_{\pm p} \oplus F_{\pm 1} \oplus \cdots \oplus F_{\pm q}$. As before, Λ contains 2^{p+q} elements. If $L_1 \in \Lambda$, then there exists a unique subspace $L_2 \in \Lambda$ such that $H|L_1$ and $H|L_2$ have only pure imaginary eigenvalues in common. As before, we say that L_1 and L_2 are a *pair of opposites* in Λ. Since (A, B) is controllable, each element L of Λ corresponds to a real symmetric solution K of the ARE. We call such a K an *unmixed* solution. If L_1 and L_2 are a pair of opposites in Λ and K_1 and K_2 are the corresponding elements of Γ, we say that K_1 and K_2 are

a *pair of opposite unmixed solutions* of the ARE. They have the property that $A-BB'K_1$ and $A-BB'K_2$ have only pure imaginary eigenvalues in common.

In [19], we show that even if H has pure imaginary eigenvalues, the unmixed solutions are isolated points in Γ. In § 4, we proved that if (A, B) is controllable and H has no pure imaginary eigenvalues, then the unmixed solutions depend analytically on the coefficient matrices (locally). This result does not extend to the case where H is allowed to have pure imaginary eigenvalues. In fact, even the extremal solutions fail to depend analytically on A, B and Q. A simple counterexample is given by L. Rodman [15] who considers the question of the analyticity of the extremal solutions in detail.

Next, we extend the classification results of § 5. Let L_1 and L_2 be a pair of opposites in Λ, and let K_1 and K_2 be the corresponding pair of opposite unmixed solutions of the ARE. In order to obtain a decomposition theorem, we need our pair of subspaces to be disjoint. Since $L_1 \cap L_2 = N$, we must modify either L_1 or L_2. There is no difference which one we modify, so we arbitrarily choose L_1. Let $\tilde{L}_1 \equiv L_1 \cap [L^+(H) \oplus L^-(H)]$. Then $\tilde{L}_1$ is of the form $E_{\pm 1} \oplus \cdots \oplus E_{\pm p} \oplus F_{\pm 1} \oplus \cdots \oplus F_{\pm q}$ while L_2 is of the form $N \oplus E_{\pm 1} \oplus \cdots \oplus E_{\pm p} \oplus F_{\pm 1} \oplus \cdots \oplus F_{\pm q}$, and it is understood that if E_i (F_j) is included in $\tilde{L}_1$, then E_{-i} (F_{-j}) is included in L_2, and conversely.

Suppose that S is an n-dimensional Lagrangian H-invariant subspace. Then $S = [S \cap L^0(H)] \oplus [S \cap E_1] \oplus [S \cap E_{-1}] \oplus \cdots \oplus [S \cap E_p] \oplus [S \cap E_{-p}] \oplus [S \cap F_1] \oplus [S \cap F_{-1}] \oplus \cdots \oplus [S \cap F_q] \oplus [S \cap F_{-q}]$. Since (A, B) is controllable, S corresponds to some $K \in \Gamma$. Since the matrix of $H|S$ with respect to the basis $\begin{bmatrix} I \\ K \end{bmatrix}$ is $A - BB'K$, Proposition 7 implies that

$$S \cap L^0(H) = L^0(H|S) = \operatorname{Sp} \begin{bmatrix} I \\ K \end{bmatrix} G = N.$$

It follows immediately that $S = [S \cap \tilde{L}_1] \oplus [S \cap L_2]$.

Remark 13. In the decomposition $S = [S \cap \tilde{L}_1] \oplus [S \cap L_2]$, L_1 and L_2 do not occur symmetrically. To obtain a symmetrical decomposition, we could modify L_2 as well as L_1. If we let $\tilde{L}_2 \equiv L_2 \cap [L^+(H) \oplus L^-(H)]$, then $S = N \oplus [S \cap \tilde{L}_1] \oplus [S \cap \tilde{L}_2]$ which has L_1 and L_2 occurring symmetrically. The reason why we use the asymmetric decomposition is that the result it gives contains the Willems–Coppel theorem (Theorem 2) as a special case. In the Willems–Coppel theorem, there is a corresponding asymmetry in the roles of K^+ and K^- due to the fact that $A - BB'K^+$ is restricted to V_+. If H has no pure imaginary eigenvalues, $\tilde{L}_1 = L_1$ and $\tilde{L}_2 = L_2$, so this issue does not arise.

The decomposition $S = [S \cap \tilde{L}_1] \oplus [S \cap L_2]$ shows that every n-dimensional Lagrangian H-invariant subspace S determines an H-invariant subspace of $\tilde{L}_1$, namely $S \cap \tilde{L}_1$. The converse is also true. First we need a preliminary result. As in § 5, we let $\Delta_{12} \equiv K_1 - K_2$, $A_1 \equiv A - BB'K_1$, $A_2 \equiv A - BB'K_2$.

LEMMA 3. $\ker \Delta_{12} = \ker \Delta$.

Proof. As in Lemma 2, $\ker \Delta_{12}$ is A_2-invariant, and A_1 and A_2 agree on this subspace. Since A_1 and A_2 have only pure imaginary eigenvalues in common $\ker \Delta_{12} \subseteq L^0(A_2) = \ker \Delta$. On the other hand, if $x \in \ker \Delta$, then $K^+x = K^-x$. Since $K^- \leqq K_1$, $K_2 \leqq K^+$, it follows that $K_1 x = K_2 x$, so $x \in \ker \Delta_{12}$. Hence, $\ker \Delta \subseteq \ker \Delta_{12}$. □

PROPOSITION 9. *Let S_1 be an l-dimensional subspace of $\tilde{L}_1$. Then there exists a unique $(n-l)$-dimensional subspace S_2 of L_2 such that $S_1 \oplus S_2$ is Lagrangian. If $S_1 = \operatorname{Sp} \begin{bmatrix} I \\ K_1 \end{bmatrix} C$, where C is an $n \times l$ full rank matrix such that $\operatorname{Sp} C \subseteq L^+(A_1) \oplus L^-(A_1)$, then $S_2 = \operatorname{Sp} \begin{bmatrix} I \\ K_2 \end{bmatrix} D$ where D is any $n \times (n-l)$ full rank matrix which satisfies $C' \Delta_{12} D = 0$. Furthermore, S_1 is H-invariant if and only if S_2 is H-invariant.*

Proof. Since S_1 is an l-dimensional subspace of $\tilde{L}_1 \subseteq L_1$, there exists an $n \times l$ full rank matrix C such that $S_1 = \operatorname{Sp}[{}^{I}_{K_1}]C$. Since A_1 is the matrix of $H|L_1$ with respect to the basis $[{}^{I}_{K_1}]$, $S_1 \subseteq \tilde{L}_1 = L_1 \cap [L^+(H) \oplus L^-(H)]$ if and only if $\operatorname{Sp} C \subseteq L^+(A_1) \oplus L^-(A_1)$. If S_2 is any $(n-l)$-dimensional subspace of L_2, then S_2 can be expressed as $S_2 = \operatorname{Sp}[{}^{I}_{K_2}]D$ for some $n \times (n-l)$ full rank matrix D. As in the proof of Proposition 6, $S_1 \oplus S_2$ is Lagrangian if and only if $C'\Delta_{12}D = 0$. Thus, to show that a unique S_2 exists such that $S_1 \oplus S_2$ is Lagrangian, we must show that given C $n \times l$ full rank, there exists an $n \times (n-l)$ full rank matrix D such that $C'\Delta_{12}D = 0$ and that this equation determines $\operatorname{Sp} D$ uniquely. This is equivalent to showing that given the l-dimensional subspace $\operatorname{Sp} C$ contained in $L^+(A_1) \oplus L^-(A_1)$, there exists a unique $(n-l)$-dimensional subspace, say M, of $\mathbb{R}^n$ such that $\Delta_{12}(M) \perp \operatorname{Sp} C$. But this is true if and only if $\dim \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp} = n - l$. (By $\Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$ we mean the inverse image of $(\operatorname{Sp} C)^{\perp}$ under Δ_{12}.)

Since Δ_{12} is not necessarily invertible, it is not obvious that $\dim \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp} = n - l$. Since $\dim \ker \Delta_{12} = r$, $\dim \Delta_{12}(\mathbb{R}^n) = n - r$. Also, $\dim (\operatorname{Sp} C)^{\perp} = n - l$. Thus, $\dim (\Delta_{12}(\mathbb{R}^n) \cap (\operatorname{Sp} C)^{\perp}) \geqq (n-r) + (n-l) - n = n - r - l$. Then $\dim \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp} = \dim \ker \Delta_{12} + \dim \Delta_{12}(\mathbb{R}^n) \cap (\operatorname{Sp} C)^{\perp} \geqq r + (n-r-l) = n - l$. Suppose that $\dim \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp} = m > n - l$, and let X be $n \times m$ full rank with $\operatorname{Sp} X = \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$. Since $\tilde{L}_1$ and L_2 are disjoint, the same is true of $\operatorname{Sp}[{}^{I}_{K_1}]C$ and $\operatorname{Sp}[{}^{I}_{K_2}]X$. It follows that $\operatorname{Sp}[{}^{C}_{K_1} \; {}^{X}_{K_2}]$ is a Lagrangian subspace of dimension $l + m > n$. However, it is obvious from the definition of a Lagrangian subspace that no such subspace can have dimension greater than n. Hence, $\dim \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp} = n - l$ as required. Thus, there is an $n \times (n-l)$ full rank matrix D such that $C'\Delta_{12}D = 0$. Furthermore, $\operatorname{Sp} D = \Delta_{12}^{-1}(\operatorname{Sp} C)^{\perp}$, so $\operatorname{Sp} D$ (and hence S_2) is uniquely determined. By the same argument as in the proof of Proposition 6, S_1 is H-invariant if and only if S_2 is H-invariant. □

Remark 14. In Proposition 9, the relationship between S_1 and S_2 is described in terms of the bases $[{}^{I}_{K_1}]$ and $[{}^{I}_{K_2}]$ for L_1 and L_2, respectively. In the case where S_1 (and hence S_2) is H-invariant, a useful basis-free description of S_2 (in terms of S_1) is possible. In fact, $S_2 = N \oplus ([J(S_1)]^{\perp} \cap \tilde{L}_2)$, where $\tilde{L}_2 = L_2 \cap (L^+(H) \oplus L^-(H))$. To prove this, it suffices to show that if we let $\hat{S}_2 = N \oplus ([J(S_1)]^{\perp} \cap \tilde{L}_2)$, then $S_1 \oplus \hat{S}_2$ is n-dimensional Lagrangian. Then the uniqueness part of Proposition 9 will imply that $\hat{S}_2 = S_2$.

Suppose that S_1 is an l-dimensional H-invariant subspace of $\tilde{L}_1$. Since $JH + H'J = 0$, it follows that $H = JH'J$. Thus, $H(S_1) \subset S_1$ implies that $JH'J(S_1) \subset S_1$. Since $J^2 = -I$, this gives $H'J(S_1) \subset J(S_1)$, which shows that $J(S_1)$ is H'-invariant. Thus $[J(S_1)]^{\perp}$ is H-invariant. Also, $\dim [J(S_1)]^{\perp} = 2n - l$. Since $[J(S_1)]^{\perp}$ is H-invariant, $[J(S_1)]^{\perp} = ([J(S_1)]^{\perp} \cap \tilde{L}_1) \oplus ([J(S_1)]^{\perp} \cap \tilde{L}_2) \oplus ([J(S_1)]^{\perp} \cap L^0(H))$. Since L_1 is Lagrangian and $S_1 \subset \tilde{L}_1 \subset L_1$, it follows that $J(S_1) \perp \tilde{L}_1$, so $[J(S_1)]^{\perp} \cap \tilde{L}_1 = \tilde{L}_1$. Since no eigenvalue of $H|L^0(H)$ is the negative of an eigenvalue of $H|\tilde{L}_1$, it follows from Lemma 1 that $J(\tilde{L}_1) \perp L^0(H)$. Thus, the fact that $S_1 \subset \tilde{L}_1$ implies that $[J(S_1)]^{\perp} \cap L^0(H) = L^0(H)$. Hence, $[J(S_1)]^{\perp} = \tilde{L}_1 \oplus ([J(S_1)]^{\perp} \cap \tilde{L}_2) \oplus L^0(H)$. Since $\dim L^0(H) = 2r$ and $\dim \tilde{L}_1 = n - r$, we conclude that $\dim ([J(S_1)]^{\perp} \cap \tilde{L}_2) = (2n-l) - (n-r) - 2r = n - l - r$. Thus, $\dim \hat{S}_2 = r + (n - l - r) = n - l$, so $\dim S_1 \oplus \hat{S}_2 = n$ as required. Since $\tilde{L}_1$, $\tilde{L}_2$ and N are each Lagrangian and $S_1 \subset \tilde{L}_1$, $[J(S_1)]^{\perp} \cap \tilde{L}_2 \subset \tilde{L}_2$, we have $J(S_1) \perp S_1$, $J([J(S_1)]^{\perp} \cap \tilde{L}_2) \perp [J(S_1)]^{\perp} \cap \tilde{L}_2$ and $J(N) \perp N$. We have already noted that $J(\tilde{L}_1) \perp L^0(H)$, so $J(S_1) \perp N$. Similarly, $J(\tilde{L}_2) \perp L^0(H)$, so $J([J(S_1)]^{\perp} \cap \tilde{L}_2) \perp N$. Also, we have the trivial fact that $J(S_1) \perp [J(S_1)]^{\perp} \cap \tilde{L}_2$. It follows immediately that $S_1 \oplus \hat{S}_2$ is Lagrangian, which completes the argument.

By using Proposition 9 in place of Proposition 6, we obtain the results which follow. They generalize Theorems 10, 11 and 12 to include the possibility that H has one or more pure imaginary eigenvalues.

THEOREM 14. *There exists a one-to-one correspondence between the set of H-invariant subspaces of* $\tilde{L}_1$ *and the set of n-dimensional Lagrangian H-invariant subspaces which are complementary to* $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$. *This correspondence associates* $\operatorname{Sp}\begin{bmatrix}I\\K_1\end{bmatrix}C$ *with* $\operatorname{Sp}\begin{bmatrix}C & D\\K_1C & K_2D\end{bmatrix}$, *where C is a full rank matrix such that* $\operatorname{Sp} C$ *is an* A_1*-invariant subspace which is contained in* $L^+(A_1)\oplus L^-(A_1)$ *and D is any full rank matrix such that* $\operatorname{Sp} D = \Delta_{12}^{-1}(\operatorname{Sp} C)^\perp$. *Equivalently, this correspondence associates the H-invariant subspace* $S_1 \subset \tilde{L}_1$ *with the n-dimensional Lagrangian H-invariant subspace* $S_1 \oplus ([J(S_1)]^\perp \cap \tilde{L}_2) \oplus N$, *which is complementary to* $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$.

Remark 15. Since (A, B) is controllable, $\operatorname{Sp}\begin{bmatrix}C & D\\K_1C & K_2D\end{bmatrix}$ is complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$, which means that the $n \times n$ submatrix $[C, D]$ is nonsingular. This shows that if M is any A_1-invariant subspace which is contained in $L^+(A_1)\oplus L^-(A_1)$, then $\Delta_{12}^{-1}(M^\perp)$ is complementary to M. Note that the additional requirement that $M \subseteq L^+(A_1)\oplus L^-(A_1)$ is trivial if H has no pure imaginary eigenvalues.

THEOREM 15. *There exists a one-to-one correspondence between the set of H-invariant subspaces of* $\tilde{L}_1$ *and the set* Γ *of real symmetric solutions of the* ARE. *This correspondence associates the subspace* $\operatorname{Sp}\begin{bmatrix}I\\K_1\end{bmatrix}C$ *with the solution* $[K_1C, K_2D][C, D]^{-1}$, *where C is a full rank matrix such that* $\operatorname{Sp} C$ *is an* A_1*-invariant subspace which is contained in* $L^+(A_1)\oplus L^-(A_1)$, *and D is any full rank matrix such that* $\operatorname{Sp} D = \Delta_{12}^{-1}(\operatorname{Sp} C)^\perp$.

THEOREM 16. *Let* $T_1 \equiv A_1|L^+(A_1)\oplus L^-(A_1)$. *There exists a one-to-one correspondence between* S_{T_1} *and* Γ *which associates the solution* $K_1P_M + K_2(I - P_M)$ *with the* T_1*-invariant subspace M, where* P_M *is the projection onto M along* $\Delta_{12}^{-1}(M^\perp)$.

Remark 16. Theorem 16 generalizes the Willems–Coppel classification theorem (Theorem 2). If $K_1 = K^+$ and $K_2 = K^-$, then this is precisely Theorem 2. Coppel's proof of Theorem 2 depends on the fact that $\Delta \equiv K^+ - K^-$ is nonnegative definite in order to show that $\Delta^{-1}(M^\perp)$ is complementary to M. Since Δ_{12} is generally indefinite, this argument does not extend to our situation. Instead we must use the controllability result (Theorem 3) as described in Remark 15.

7. Conclusion. In § 2, we noted that the principal disadvantage of the Hamiltonian matrix method is its failure to exploit the Hamiltonian symmetry which is present in the ARE. Consequently, it does not explain how the Lagrangian condition and the complementarity requirement cut out a subset from the set of all n-dimensional H-invariant subspaces. In this paper, we have addressed this question in some detail. In § 3 we showed that the controllability of (A, B) is sufficient to guarantee that every n-dimensional Lagrangian H-invariant subspace is complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$ and hence corresponds to a real symmetric solution of the ARE. If H has no pure imaginary eigenvalues, the controllability of (A, B) is a necessary condition as well.

In §§ 5 and 6, we have presented a simple classification for all the n-dimensional Lagrangian H-invariant subspaces which are complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$. This result (Theorem 14) describes a one-to-one correspondence between this set and the set of all invariant subspaces of a particular matrix. We have described elsewhere [19] a simple parametrization for the set of all invariant subspaces of an arbitrary finite-dimensional linear operator. Combining this parametrization with Theorem 14 gives a simple parametrization of the set of n-dimensional Lagrangian H-invariant subspaces which are complementary to $\operatorname{Sp}\begin{bmatrix}0\\I\end{bmatrix}$.

In order to derive Theorem 14, we have incorporated into the Hamiltonian framework some of the ideas behind the Willems–Coppel classification theorem. In doing so, we have obtained a generalization (Theorem 16) of the Willems–Coppel result. In our result, the pair of extremal solutions, K^+ and K^-, is replaced by an arbitrary pair of what we call opposite unmixed solutions, K_1 and K_2.

Our geometric approach gives a clear interpretation of the roles of controllability and observability in the theory of the ARE. The role of controllability is to guarantee that there are no solutions at infinity. Controllability also ensures that the ARE is genuinely quadratic and that the set Γ of real symmetric solutions is compact (with respect to the Euclidean topology). The role of observability is to guarantee that there are no solutions at zero—that every element of Γ is nonsingular and hence defines a nondegenerate quadratic form. When we make precise the concept of points at infinity and points at zero by compactifying the ARE in the Lagrangian Grassmannian, the geometric roles of controllability and observability are dual: controllability implies that every n-dimensional Lagrangian H-invariant subspace is complementary to $\operatorname{Sp}\left[\begin{smallmatrix}0\\I\end{smallmatrix}\right]$, whereas observability implies that every such subspace is complementary to $\operatorname{Sp}\left[\begin{smallmatrix}I\\0\end{smallmatrix}\right]$.

We have also studied the special subset of Γ consisting of the unmixed solutions. These solutions include the extremal solutions, K^+ and K^-, and share several of their properties. As we have already noted, they occur in pairs and each such pair gives a decomposition theorem which is analogous to the Willems–Coppel classification theorem. Each unmixed solution is isolated, and if H has no pure imaginary eigenvalues, then it depends analytically on the coefficient matrices. However, the concept of an unmixed solution extends to non-Hamiltonian (even nonsquare) ARE's, while the concept of extremal solutions does not.

The classification theorems in §§ 5 and 6 were derived for the set of real symmetric solutions of the usual (i.e., Hamiltonian) ARE $-A'K-KA+KBB'K-Q=0$ under the assumption that (A, B) is controllable. We have proved in § 4 that if (A, B) is not controllable and H has no pure imaginary eigenvalues, then no pair of opposite unmixed solutions exist. Hence, no Willems-type decomposition theorem is possible in the absence of controllability.

It is possible to obtain theorems resembling Theorems 10, 11 and 12 which classify the set of all n-dimensional H-invariant subspaces and the set of all real solutions (not necessarily symmetric) of the ARE, at least in the case where H has no pure imaginary eigenvalues [20]. However, these results are not so nice as those for the Lagrangian subspaces and symmetric solutions. This is due to the fact that the controllability of (A, B) does not guarantee that every n-dimensional H-invariant subspace is complementary to $\operatorname{Sp}\left[\begin{smallmatrix}0\\I\end{smallmatrix}\right]$. Also, in our view, there is no practical reason for wanting such classification theorems for the set of all real solutions. This is because the primary function of these theorems is to describe how the Lagrangian requirement determines a subset of the set of all n-dimensional H-invariant subspaces. If we do not care about the Lagrangian condition, this motivation is absent.

We have noted in § 4 that the concept of unmixed solutions extends to the generalized ARE $B_{21}+B_{22}P-PB_{11}-PB_{12}P=0$, where the unknown matrix P is $m\times p$. However, since this is not a Hamiltonian ARE, the unmixed solutions do not occur naturally in pairs. Thus, a Willems-type decomposition is not possible. In addition, there is no real motivation for such a decomposition since there is no Lagrangian condition with which to deal.

Acknowledgment. The results contained in this paper represent part of the author's doctoral thesis [20] completed under the direction of Professor R. W. Brockett of Harvard University.

REFERENCES

[1] R. W. BASS, *Machine solution of high order matrix Riccati equations*, Douglas Aircraft, Missiles and Space Systems Division, Santa Monica, CA, 1967.

[2] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.
[3] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, 1970.
[4] R. S. BUCY AND P. D. JOSEPH, *Filtering for Stochastic Processes with Applications to Guidance*, Interscience, New York, 1968.
[5] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.
[6] D. F. DELCHAMPS, *A note on the analyticity of the Riccati metric*, in Algebraic and Geometric Methods in Linear Systems Theory, C. I. Byrnes and C. F. Martin, eds, American Mathematical Society, Providence, RI, 1980.
[7] R. HERMANN, *Cartanian Geometry, Nonlinear Waves, and Control Theory, Part A*, Interdisciplinary Mathematics, Vol. XX, Math. Sci. Press, Brookline, MA, 1979.
[8] V. KUČERA, *A review of the matrix Riccati equation*, Kybernetika, 9 (1973), pp. 42–61.
[9] P. LANCASTER AND L. RODMAN, *Existence and uniqueness theorems for the algebraic Riccati equation*, Internat. J. Control, 32 (1980), pp. 285–309.
[10] A. G. J. MACFARLANE, *An eigenvector solution of the optimal linear regulator problem*, J. Electron. Contr., 14 (1963), pp. 496–501.
[11] K. MÅRTENSSON, *On the matrix Riccati equation*, Inform. Sci., 3 (1971), pp. 17–49.
[12] C. F. MARTIN, *Grassmannian manifolds, Riccati equations and feedback invariants of linear systems*, in Geometrical Methods for the Theory of Linear Systems, C. I. Byrnes and C. F. Martin, eds., Reidel, Dordrecht, 1980.
[13] N. NARASIMHAMURTHI AND F. F. WU, *On the Riccati equation arising from the study of singularly perturbed systems*, in Proc. Joint Automatic Control Conference, 1977, pp. 1244–1247.
[14] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496–501.
[15] L. RODMAN, *On extremal solutions of the algebraic Riccati equation*, in Algebraic and Geometric Methods in Linear Systems Theory, C. I. Byrnes and C. F. Martin, eds., American Mathematical Society, Providence, RI, 1980.
[16] J. RODRIGUEZ-CANABAL, *The geometry of the Riccati equation*, Ph.D. thesis, Univ. Southern California, Los Angeles, 1972.
[17] J. RODRIGUEZ-CANABAL, *The geometry of the Riccati equation*, Stochastics, 1 (1973), pp. 129–149.
[18] W. E. ROTH, *On the matrix equation* $X^2+AX+XB+C=0$, Proc. Amer. Math. Soc., 1 (1950), pp. 586–589.
[19] M. A. SHAYMAN, *Geometry of the algebraic Riccati equation, Part* II, this Journal, this issue, pp. 395–409.
[20] ———, *Varieties of invariant subspaces and the algebraic Riccati equation*, Ph.D. thesis, Harvard University, Cambridge, MA, 1980.
[21] ———, *On the variety of invariant subspaces of a finite-dimensional linear operator*, Trans. Amer. Math. Soc., to appear.
[22] A. C. M. VAN SWIETEN, *Qualitative behavior of dynamical games with feedback strategies*, Ph.D. thesis, University of Groningen, the Netherlands, 1977.
[23] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Contr., 16 (1971), pp. 621–634.
[24] H. K. WIMMER, *On the algebraic Riccati equation*, Bull. Austral. Math. Soc., 14 (1976), pp. 457–461.
[25] W. M. WONHAM, *Linear Multivariable Control*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL AND OTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0004 $01.25/0

GEOMETRY OF THE ALGEBRAIC RICCATI EQUATION, PART II*

MARK A. SHAYMAN†

Abstract. We prove that the set of real symmetric solutions of the algebraic Riccati equation is isomorphic to the algebraic variety of invariant subspaces of a related $n \times n$ matrix. By characterizing the structure of this variety, we obtain a detailed description of the geometric properties of the solution set of the algebraic Riccati equation.

Key words. algebraic Riccati equation, linear quadratic control, invariant subspaces

1. Introduction. By the algebraic Riccati equation (ARE), we mean the quadratic matrix equation $-A'K - KA + KBB'K - Q = 0$ where A, B and Q are given real matrices of dimensions $n \times n$, $n \times m$ and $n \times n$ respectively, and $Q = Q'$. The ARE is the algebraic equation satisfied by the equilibrium points of the Riccati differential equation, and is of critical importance in optimal control and filtering theory. For most applications, one is interested in the set of real symmetric solutions of the ARE. We use Γ to designate this set.

It is well known [13] that if (A, B) is controllable and Γ is nonempty, then Γ contains a unique element, K^+ (K^-), with the property that every eigenvalue of $A - BB'K^+$ ($A - BB'K^-$) has nonpositive (nonnegative) real part. It is the solution K^+ which is used to construct the optimal feedback control for the linear-quadratic stationary optimal control problem and for this reason a large portion of the research on the ARE is devoted to the design of algorithms to obtain K^+.

However, there are several reasons why it is important to understand the structure of all Γ. For one thing, knowledge of the entire solution set is useful for the development of numerical methods to find K^+. In addition, there are other applications where knowledge of all of Γ is of intrinsic importance. They include network synthesis realizations which employ a minimal number of resistors and the construction of all minimal square solutions to the spectral factorization problem with application to stochastic realization theory [4].

There are two approaches in the literature to the problem of characterizing Γ. J. C. Willems [13] proved that under mild assumptions, the elements of Γ are in one-to-one correspondence with the set of all invariant subspaces of the matrix $A^+ \equiv A - BB'K^+$. Using this bijection, Willems showed that every real symmetric solution can be expressed as a combination of K^+ and K^-. This result was generalized by W. Coppel [3].

The second approach to the characterization of Γ establishes a one-to-one correspondence between Γ and a certain subset of the set of all n-dimensional invariant subspaces of the $2n \times 2n$ Hamiltonian matrix $H \equiv \begin{bmatrix} A & -BB' \\ -Q & -A' \end{bmatrix}$. This method was developed and generalized through the work of several authors including J. Potter [6], K. Mårtensson [5] and A. C. M. Van Swieten [12]. A discussion of the relationship between Willems' approach and the Hamiltonian matrix method is included in our companion paper [9] (this issue, pp. 375–394).

Both existing results are set-theoretic. In each case the existence of a bijection is established. On the other hand, Γ has more structure than a set. At the very least, Γ has the topological structure it inherits as a subset of the Euclidean space of $n \times n$

* Received by the editors July 11, 1981, and in revised form May 10, 1982. This research was partially supported by the U.S. Army Office of Research under grant DAAG 29-79-C-0147.

† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

real symmetric matrices. Thus, it is appropriate to ask topological/geometric questions about Γ. For example, when is Γ finite? When does it contain continuous families of solutions? How many connected components are there? Do the connected components have a system-theoretic interpretation? Does Γ admit the structure of a differentiable manifold (with the Euclidean topology)? If so, what is its dimension? The answers to geometric questions of this type should prove useful to those interested in the applications mentioned above. In particular, they should aid in the design of numerical algorithms to solve the ARE.

Because the existing results are set-theoretic, the only geometric information they can supply is the cardinality of Γ. In § 2, we show that if (A, B) is controllable, then Γ is naturally identified with a projective variety. With this identification, Γ can itself be viewed as a projective variety. Then we prove that the bijection in Willems' theorem is an isomorphism of projective varieties. This enables us to obtain the geometric properties of Γ from those of S_{A^+}, the variety of invariant subspaces of A^+. This leaves us with the problem of describing the structure of the variety of invariant subspaces of an arbitrary linear operator on $\mathbb{R}^n$ or $\mathbb{C}^n$. This is a difficult problem unless the operator is semisimple (diagonalizable). However, the problem is tractible, and our results are outlined in § 3. For the proofs, the reader is referred to [11], [10]. In § 4, we use these results to describe the topological structure of Γ, and we give system-theoretic interpretations for some of the geometric properties. In § 5, we describe a parametrization for the set of invariant subspaces of an arbitrary linear operator on $\mathbb{R}^n$ or $\mathbb{C}^n$. We used this parametrization in [10] to derive several of the topological properties discussed in § 3. As far as we are aware, there are no published parametrizations of the set of invariant subspaces of a finite dimensional linear operator. If the operator is not semisimple, there is no obvious way of listing all of its invariant subspaces. However, without such a parametrization, it is not clear how to use either Willems' theorem or the Hamiltonian matrix method to generate all the solutions of the ARE. Since invariant subspaces occur throughout linear systems theory as well as elsewhere in applied mathematics, it is our hope that these results will prove useful in other applications besides the algebraic Riccati equation.

We close this section with an example which shows that Γ has interesting geometric structure even for a very simple ARE.

Example 1. Consider the ARE $K^2 - I = 0$. Then Γ consists of all symmetric matrices in $O(n)$, the group of $n \times n$ orthogonal matrices. If $X \in O(n)$ and $K \in \Gamma$, then $X'KX \in \Gamma$, so $O(n)$ acts on Γ by conjugation. K_1 and K_2 belong to the same orbit if and only if they have the same spectrum. Since the eigenvalues of a symmetric orthogonal matrix are ± 1, there are exactly $n+1$ orbits. Each orbit contains an element in the canonical form

$$\begin{bmatrix} I_p & 0 \\ 0 & -I_{n-p} \end{bmatrix} \qquad (p = 0, 1, \cdots, n).$$

Since the stabilizer of this matrix is $O(p) \times O(n-p)$, the corresponding orbit is the homogeneous space $O(n)/O(p) \times O(n-p)$, which is the Grassmann manifold $G^p(\mathbb{R}^n)$ of all p-dimensional subspaces of $\mathbb{R}^n$. (For a brief discussion of homogeneous spaces, see [1, pp. 164–171].) A simple argument using the fact that the eigenvalues of a matrix depend continuously on its entries shows that the connected components of Γ are precisely the $n+1$ orbits. Hence, Γ is the disjoint union $\bigsqcup_{p=0}^{n} G^p(\mathbb{R}^n)$.

2. Isomorphism theorem. Let $\mathcal{k}$ be any field, and let $\mathcal{k}[X_1, \cdots, X_n]$ be the ring of polynomials in n variables with coefficients in $\mathcal{k}$. Let $\{p_\alpha\}$ be any collection of polynomials in $\mathcal{k}[X_1, \cdots, X_n]$, and let $V = \{x \in \mathcal{k}^n : p_\alpha(x) = 0, \forall \alpha\}$. Then V is called

an (affine) *algebraic variety* in $\mathbb{k}^n$. A nonzero polynomial $q \in \mathbb{k}[X_1, \cdots, X_n]$ is said to be *homogeneous* if each term in q is of the same degree. If every polynomial in $\{p_\alpha\}$ is homogeneous, then V is called a *homogeneous variety*. In this case, it is easily seen that if $x \in V$, then every multiple of x belongs to V. Thus, V can be viewed as a collection of lines through the origin. To be precise, we define an equivalence relation on $V - \{0\}$ whereby $x \sim y$ if there exists a nonzero real number λ such that $y = \lambda x$. Then we can view V as the set of equivalence classes. When the homogeneous variety V is viewed in this way, it is called a *projective variety*.

In practice, the terms "algebraic variety" and "projective variety" are also used to describe objects which can be identified with algebraic or projective varieties by structure-preserving mappings. Let W be a finite-dimensional vector space over $\mathbb{k}(= \mathbb{R}$ or $\mathbb{C})$, and let $G^j(W)$ denote the set of all j-dimensional subspaces of W. Then the elements of $G^j(W)$ do not belong to a Euclidean space. However, the classical Plücker embedding identifies $G^j(W)$ with a projective variety, so $G^j(W)$ is itself regarded as a projective variety. For the basic concepts of algebraic geometry, we refer the reader to [2]. This volume also describes various applications of algebraic geometry to problems in systems theory.

The set Γ of real symmetric solutions of the algebraic Riccati equation is the zero set of a collection of quadratic polynomials defined on the $\frac{1}{2}n(n+1)$-dimensional vector space of real symmetric matrices. Thus, Γ is an algebraic variety over the field $\mathbb{R}$. In this section, we show that if (A, B) is controllable, then Γ can be identified with a certain projective variety in a natural way, and can therefore be itself regarded as a projective variety. We then prove that the one-to-one correspondence in the Willems–Coppel classification theorem is an isomorphism of projective varieties.

Let $G^n(\mathbb{R}^{2n})$ denote the Grassmann manifold of all n-dimensional subspaces of $\mathbb{R}^{2n}$. Let J be the $2n \times 2n$ matrix $\left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]$. If M is a subspace of $\mathbb{R}^{2n}$ (not necessarily n-dimensional), we say that M is Lagrangian if $x'Jy = 0$, $\forall x, y \in M$. Let $M(n)$ denote the vector space of all $n \times n$ real matrices. Define a mapping $\psi: M(n) \to G^n(\mathbb{R}^{2n})$ by $\psi(K) = \operatorname{Sp}\left[\begin{smallmatrix} I \\ K \end{smallmatrix}\right]$, the column space of the matrix $\left[\begin{smallmatrix} I \\ K \end{smallmatrix}\right]$. ψ embeds $M(n)$ in $G^n(\mathbb{R}^{2n})$ as an open and dense submanifold.

Let $H = \left[\begin{smallmatrix} A & -BB' \\ -Q & -A' \end{smallmatrix}\right]$ denote the $2n \times 2n$ Hamiltonian matrix corresponding to the ARE. Let $\tilde{\Gamma}$ denote the set of all elements of $G^n(\mathbb{R}^{2n})$ which are both Lagrangian and H-invariant. If (A, B) is controllable and Γ is nonempty, then it is shown in the companion paper [9] that ψ maps Γ onto $\tilde{\Gamma}$. Thus, we can use the embedding ψ to identify Γ with the subset $\tilde{\Gamma}$ of $G^n(\mathbb{R}^{2n})$. The assumption of controllability is essential. If (A, B) is not controllable, $\tilde{\Gamma}$ can (and generally does) contain elements which are not complementary to the n-dimensional subspace $\operatorname{Sp}\left[\begin{smallmatrix} 0 \\ I \end{smallmatrix}\right]$ and thus do not belong to the image of ψ.

Let $\mathscr{L}$ denote the subset of $G^n(\mathbb{R}^{2n})$ consisting of those elements which are Lagrangian, and let $S_H(n)$ denote the subset of $G^n(\mathbb{R}^{2n})$ consisting of those elements which are H-invariant. If $\mathbb{R}^{2n}$ is endowed with the standard inner product, then $\mathscr{L} = \{M \in G^n(\mathbb{R}^{2n}): [J(M)]^\perp = M\}$. Thus, $\mathscr{L}$ is the fixed point set of the mapping $M \mapsto [J(M)]^\perp$ on $G^n(\mathbb{R}^{2n})$. Since this is a regular mapping of the projective variety $G^n(\mathbb{R}^{2n})$, it follows immediately that $\mathscr{L}$ is a subvariety of $G^n(\mathbb{R}^{2n})$. It is also easily shown that $S_H(n)$ is a subvariety of $G^n(\mathbb{R}^{2n})$ [10]. Since $\tilde{\Gamma} = \mathscr{L} \cap S_H(n)$, it follows that $\tilde{\Gamma}$ is a subvariety of $G^n(\mathbb{R}^{2n})$. By using the embedding ψ to identify Γ with $\tilde{\Gamma}$, we can therefore view Γ as a subvariety of $G^n(\mathbb{R}^{2n})$.

We introduce some notation. Let R be a linear operator on $\mathbb{R}^k$ with characteristic polynomial $p(s)$. Express $p(s)$ as the product $p(s) = p^+(s)p^0(s)p^-(s)$, where the roots of $p^+(s)$ $(p^0(s))$ $(p^-(s))$ have negative (zero) (positive) real parts. Define $L^+(R) \equiv$

$\ker p^+(R)$, $L^0(R) \equiv \ker p^0(R)$, $L^-(R) \equiv \ker p^-(R)$. Thus, $L^+(R)$ $(L^0(R))$ $(L^-(R))$ is the sum of the primary components of R which correspond to its left half-plane (imaginary axis) (right half-plane) eigenvalues.

Suppose that a real symmetric solution of the ARE exists—i.e., Γ is nonempty. Then K^+ and K^- exist. It is well known [13] that if $K \in \Gamma$, then $K^- \leqq K \leqq K^+$ with respect to the usual partial ordering of symmetric matrices. Let $\Delta \equiv K^+ - K^-$. It can be shown [3] that for any $K \in \Gamma$, $L^0(A - BB'K) = \ker \Delta$. In fact, since $K^- \leqq K \leqq K^+$, it follows that $A - BB'K | \ker \Delta = A^+ | \ker \Delta$.

Let $V_+ \equiv L^+(A^+)$, $V_0 \equiv L^0(A^+)$ $(= \ker \Delta)$, and let $V_- \equiv L^-(A - BB'K^-)$. Note that $L^-(A^+) = 0$ and $L^+(A - BB'K^-) = 0$. Let $T \equiv A^+ | V_+$, and let S_T be the set of all invariant subspaces of T. The following result is Coppel's generalization of Willems' theorem [13], [3].

THEOREM 1 (J. C. Willems, W. Coppel). *Let (A, B) be controllable and suppose that Γ is nonempty. If S is any A^+-invariant subspace which is contained in V_+ (i.e., $S \in S_T$), then $S \oplus \Delta^{-1}(S^\perp) = \mathbb{R}^n$, where $S^\perp$ denotes the orthogonal complement of S in $\mathbb{R}^n$ and $\Delta^{-1}(S^\perp)$ is its inverse image. There is a bijection $\phi : S_T \to \Gamma$ defined by $\phi(S) \equiv K^+ P_S + K^-(I - P_S)$, where P_S is the projection onto S along $\Delta^{-1}(S^\perp)$. If $K = \phi(S)$, then $L^+(A - BB'K) = S$, $L^0(A - BB'K) = V_0$ and $L^-(A - BB'K) = \Delta^{-1}(S^\perp) \cap V_-$.*

If $\Delta > 0$, then $V_+ = \mathbb{R}^n$ so $T = A^+$, and Theorem 1 reduces to the original theorem of Willems.

We will assume that the hypotheses of Theorem 1 are satisfied. Under these conditions, it is shown in [9] that $L^0(H)$ contains a unique Lagrangian H-invariant subspace N with the property that $\dim N = (\frac{1}{2}) \dim L^0(H)$. It is also shown that there is a one-to-one correspondence between $\tilde{\Gamma}$ and the set of H-invariant subspaces of $L^+(H)$. Let H^+ denote the restriction of H to $L^+(H)$, and let S_{H^+} denote the set of all invariant subspaces (of all possible dimensions) of H^+. If we let $\delta : \tilde{\Gamma} \to S_{H^+}$ denote the bijection, then $\delta(M) = M \cap L^+(H)$ and $\delta^{-1}(M_1) = M_1 \oplus ([J(M_1)]^\perp \cap L^-(H)) \oplus N$. (See [9 Thm. 14].)

Let R be a linear operator on a finite dimensional real vector space V. Let $E_1, \cdots, E_p$ denote the primary components of R, and let R_i be the restriction of R to E_i. Let S_R denote the variety of all invariant subspaces of R, and let S_{R_i} denote the variety of all invariant subspaces of R_i. There is a natural bijection $\alpha : S_R \to S_{R_1} \times \cdots \times S_{R_p}$ defined by $\alpha(S) = (S \cap E_1, \cdots, S \cap E_p)$. The inverse of α is given by $\alpha^{-1}(S_1, \cdots, S_p) = S_1 \oplus \cdots \oplus S_p$. It is proven in [10] that α and α^{-1} are regular mappings, and thus α is an isomorphism of projective varieties. Now, let E, F and G each be sums of some of the E_i's such that $V = E \oplus F \oplus G$. Let R_E, R_F and R_G denote the restrictions of R to E, F and G respectively. There is a natural bijection $\beta : S_R \to S_{R_E} \times S_{R_F} \times S_{R_G}$ given by $\beta(S) = (S \cap E, S \cap F, S \cap G)$. By essentially the same argument used for α, it follows that β is an isomorphism of projective varieties.

We return to the Riccati equation. Recall that $H^+ = H | L^+(H)$. Let $H^- = H | L^-(H)$ and let $H^0 = H | L^0(H)$. Then the mapping $\gamma : S_H \to S_{H^+} \times S_{H^-} \times S_{H^0}$ with $\gamma(M) = (M \cap L^+(H), M \cap L^-(H), M \cap L^0(H))$ is an isomorphism (of projective varieties). The mapping δ is the restriction of γ to $S_H(n) \cap \mathscr{L}$ followed by projection on the first factor. Thus, δ is a regular mapping. The mapping δ^{-1} is the mapping $M_1 \mapsto (M_1, [J(M_1)]^\perp \cap L^-(H), N)$ followed by γ^{-1}. The mapping $M_1 \mapsto [J(M_1)]^\perp \cap L^-(H)$ consists of the regular mapping $M_1 \mapsto [J(M_1)]^\perp$ followed by γ which is then followed by projection on the second factor. It is therefore a regular mapping. We conclude that δ^{-1} is a regular mapping. We have:

LEMMA 1. *Suppose that (A, B) is controllable and that Γ is nonempty. Then the mapping $\delta : \tilde{\Gamma} \to S_{H^+}$ with $\delta(M) = M \cap L^+(H)$ is an isomorphism of projective varieties.*

We can now prove that Γ is isomorphic to S_T.

THEOREM 2. *Let (A, B) be controllable and suppose Γ is nonempty. Then the mapping $\phi: S_T \to \Gamma$ defined by $\phi(S) = K^+P_S + K^-(I - P_S)$ is an isomorphism of projective varieties.*

Proof. The structure of Γ as a projective variety is that which it inherits from $\tilde{\Gamma}$ via the embedding ψ. Thus $\phi: S_T \to \Gamma$ is an isomorphism if and only if $\psi \circ \phi: S_T \to \tilde{\Gamma}$ is an isomorphism. Therefore, by Lemma 1, it suffices to show that $\delta \circ \psi \circ \phi: S_T \to S_{H^+}$ is an isomorphism of projective varieties. We have

$$(*) \qquad \begin{aligned} \psi(\phi(S)) &= \operatorname{Sp}\begin{bmatrix} I \\ K^+P_S + K^-(I - P_S) \end{bmatrix} \\ &= \operatorname{Sp}\begin{bmatrix} I \\ K^+ \end{bmatrix} P_S + \operatorname{Sp}\begin{bmatrix} I \\ K^- \end{bmatrix}(I - P_S). \end{aligned}$$

It is shown in [9] that $\operatorname{Sp}[{}^{I}_{K^+}] = L^+(H) \oplus N$. Thus, the mapping $x \overset{\mu}{\mapsto} [{}^{I}_{K^+}]x$ is an isomorphism of $\mathbb{R}^n$ onto $L^+(H) \oplus N$. It is well known that A^+ is the matrix of the restriction of H to $\operatorname{Sp}[{}^{I}_{K^+}]$ relative to the basis given by the columns of $[{}^{I}_{K^+}]$. (See, e.g., [9].) Thus, $H[{}^{I}_{K^+}] = [{}^{I}_{K^+}]A^+$. This implies that μ maps $L^+(A^+)(= V_+)$ onto $L^+(H)$. It is also clear that if S is an A^+-invariant subspace of V_+ (i.e., $S \in S_T$), then the image of S under μ is an H-invariant subspace of $L^+(H)$ (i.e., $\mu(S) \in S_{H^+}$). Thus, μ induces an isomorphism, call it $\bar{\mu}$, of S_T onto S_{H^+}, where $\bar{\mu}(S)$ is the image of S under μ.

Since μ maps V_+ onto $L^+(H)$, the subspace $\operatorname{Sp}[{}^{I}_{K^+}]P_S$ in (*) is contained in $L^+(H)$. From [9], we know that $\operatorname{Sp}[{}^{I}_{K^-}] = L^-(H) \oplus N$. Hence, the subspace $\operatorname{Sp}[{}^{I}_{K^-}](I - P_S)$ is contained in $L^-(H) \oplus N$. (Thus, the sum in (*) is a direct sum.) This means that $\delta(\psi(\phi(S))) = \psi(\phi(S)) \cap L^+(H) = \operatorname{Sp}[{}^{I}_{K^+}]P_S$. But this means that $\delta(\psi(\phi(S)))$ is the image of S under μ. Thus, $\delta \circ \psi \circ \phi = \bar{\mu}$, showing that $\delta \circ \psi \circ \phi$ is an isomorphism. □

It is shown in [9] that the Willems–Coppel theorem holds with K^+, K^- replaced by other pairs of solutions (called *opposite unmixed solutions*). By an argument completely analogous to the one given above, it is easily seen that the corresponding bijection is again an isomorphism of projective varieties.

3. The variety of invariant subspaces. In this section, we describe the geometric structure of the variety of invariant subspaces of an arbitrary finite dimensional linear operator. We omit the proofs since they appear elsewhere [11], [10]. Let V be an n-dimensional vector space over the field $\mathcal{F}$ of real or complex numbers. Let $A \in \operatorname{Hom}(V)$—i.e., a linear mapping of V into itself. S_A denotes the set of all invariant subspaces of A, and $S_A(k)$ denotes those elements of S_A which are k-dimensional. $S_A(k)$ is topologized as a subset of $G^k(V)$, the Grassmann manifold of all k-dimensional subspaces of V, and the topologies on $\{S_A(k)\}_{k=0}^n$ generate a topology on S_A which makes S_A the topological disjoint union $\bigsqcup_{k=0}^n S_A(k)$.

THEOREM 3. *$S_A(k)$ is a compact subvariety of $G^k(V)$.*

THEOREM 4a. *Let $\mathcal{F} \equiv \mathbb{C}$. Let $F_1, \cdots, F_q$ be the generalized eigenspaces of A. Let A_j be the restriction of A to F_j, and let S_{A_j} be the variety of invariant subspaces of A_j, $j = 1, \cdots, q$. Define $\eta: S_A \to S_{A_1} \times \cdots \times S_{A_q}$ by $\eta(S) \equiv (S \cap F_1, \cdots, S \cap F_q)$. Then η is an isomorphism of projective varieties.*

THEOREM 4b. *Let $\mathcal{F} = \mathbb{R}$. Let B be the complexification of A. Let $E_1, \cdots, E_p$ be the generalized eigenspaces of A corresponding to its real eigenvalues, and let $F_1, \bar{F}_1, \cdots, F_q, \bar{F}_q$ be the generalized eigenspaces (in conjugate pairs) of B corresponding to the nonreal eigenvalues of A. Let A_i be the restriction of A to E_i $(i = 1, \cdots, p)$, and let B_j be the restriction of B to F_j $(j = 1, \cdots, q)$. Define $\eta: S_A \to S_{A_1} \times \cdots \times S_{A_p} \times S_{B_1} \times$*

$\cdots \times S_{B_q}$ *by* $\eta(S) \equiv (S \cap E_1, \cdots, S \cap E_p, S^+ \cap F_1, \cdots, S^+ \cap F_q)$, *where* S^+ *is the complexification of the subspace* S. *Then* η *is an isomorphism of real projective varieties.*

It is important to note that B_j is an operator on the complex vector space F_j, so $S_{B_j}(k)$ is a subvariety of the complex Grassmann manifold $G^k(F_j)$.

The importance of Theorems 4a and 4b is that they express the variety of invariant subspaces of A as the product of varieties of invariant subspaces of operators which have only one distinct eigenvalue. Thus, to characterize S_A for every operator A, it suffices to characterize S_A for every real operator and every complex operator possessing a single eigenvalue. However, for such an operator A, there exists $\lambda \in \mathcal{F}$ such that $A - \lambda I$ is nilpotent. Since $S_{A-\lambda I} = S_A$, there is no loss of generality in taking A to be nilpotent. *The remainder of the results in this section are stated for nilpotent* A.

Let A be a nilpotent operator on V, and let $(m_1, \cdots, m_r)$ be the partition of n corresponding to the Jordan block structure of A $(m_1 \geqq \cdots \geqq m_r \geqq 1)$. Let k be fixed, and let $(p_1, \cdots, p_l)$ be a partition of k $(p_1 \geqq \cdots \geqq p_l \geqq 1)$. We say that $(p_1, \cdots, p_l)$ is a *partition of* k *compatible with the block structure of* A if $l \leqq r$ and $p_j \leqq m_j$, $j = 1, \cdots, l$. Let $S_A(k; p_1, \cdots, p_l)$ be the subset of $S_A(k)$ consisting of those elements which have cyclic structure $(p_1, \cdots, p_l)$. It is easy to show that $S_A(k; p_1, \cdots, p_l)$ is nonempty if and only if $(p_1, \cdots, p_l)$ is compatible with the block structure of A. It is trivial to show that there is always at least one partition of k which is compatible with the block structure of A, so $S_A(k)$ is nonempty.

There are two special cases where the structure of $S_A(k)$ is readily apparent and well known. These are described in Theorem 5.

THEOREM 5.

(a) *If* A *is semisimple (diagonalizable), then* $S_A(k) = G^k(V)$.

(b) *If* A *is cyclic, then* $S_A(k)$ *consists of exactly one point.*

The next result shows that S_A is finite if and only if A is cyclic.

THEOREM 6. *If* A *is cyclic, then* S_A *consists of exactly* $n+1$ *points. Otherwise,* S_A *contains a connected component which is a projective space of positive dimension.*

Theorem 5 describes the structure of S_A in the cases where A is semisimple or cyclic. These are the extreme cases. In the semisimple case, A has n 1×1 blocks, whereas in the cyclic case A has one $n \times n$ block. A class of nilpotent operators which includes both the semisimple and the cyclic cases is the set of operators with block structure of the form $(m_1, \cdots, m_r) = (m_1, 1, \cdots, 1)$. (Note that $n = m_1 + r - 1$.) These are the operators which have at most one block of size greater than 1×1. Note that if $r = n$ and $m_1 = 1$, then A is semisimple, while if $r = 1$ and $m_1 = n$, then A is cyclic. Also, if $(m_1, \cdots, m_r) = (m_1, 1, \cdots, 1)$, then the only partitions of k which are compatible with the block structure of A are of the form $(p_1, \cdots, p_l) = (p_1, 1, \cdots, 1)$, where $p_1 \leqq m_1$ and $l \leqq r$. The following result describes completely the geometry of $S_A(k)$ for an operator A with block structure $(m_1, 1, \cdots, 1)$.

THEOREM 7. *Let* A *have block structure* $(m_1, \cdots, m_r) = (m_1, 1, \cdots, 1)$, *and let* $(p_1, \cdots, p_l) = (p_1, 1, \cdots, 1)$ *be a partition of* k *which is compatible with the block structure of* A. *Then* (1) *If* $p_1 = 1$ *or* $l = r$, $S_A(k; p_1, 1, \cdots, 1) \approx G^l(\mathcal{F}^r)$. (2) *If* $p_1 > 1$ *and* $l < r$, *then*

$$S_A(k; p_1, 1, \cdots, 1) \approx G^l(\mathcal{F}^r) - G^l(\mathcal{F}^{r-1}), \qquad \overline{S_A(k; p_1, 1, \cdots, 1)} \approx G^l(\mathcal{F}^r),$$

and

$$\overline{S_A(k; p_1, 1, \cdots, 1)} - S_A(k; p_1, 1, \cdots, 1) \subseteq S_A(k; \underbrace{p_1 - 1, 1, \cdots, 1}_{(l+1)\ terms}).$$

(An overbar indicates closure and $\approx$ indicates isomorphism.)

Example 2. Let $n=6$ and suppose that the block structure of A is $(m_1, m_2, m_3, m_4) \equiv (3, 1, 1, 1)$. Let $k=3$. There are three partitions of k which are compatible with the block structure of A: (3), (2, 1), (1, 1, 1). By Theorem 7, $S_A(3;3) \approx G^1(\mathcal{F}^4) - G^1(\mathcal{F}^3)$, $S_A(3;2,1) \approx G^2(\mathcal{F}^4) - G^2(\mathcal{F}^3)$ and $S_A(3;1,1,1) \approx G^3(\mathcal{F}^4)$. Furthermore, $\overline{S_A(3;3)} \approx G^1(\mathcal{F}^4)$, $\overline{S_A(3;3)} - S_A(3;3) \subseteq S_A(3;2,1)$, $\overline{S_A(3;2,1)} \approx G^2(\mathcal{F}^4)$ and $\overline{S_A(3;2,1)} - S_A(3;2,1) \subseteq S_A(3;1,1,1)$. Thus, $S_A(3;3)$ has the topological structure of $G^1(\mathcal{F}^4)$ except that it is missing a piece which has the structure of $G^1(\mathcal{F}^3)$. This missing piece is a subset of $S_A(3;2,1)$. $S_A(3;2,1)$ has the topological structure of $G^2(\mathcal{F}^4)$ except that it is missing a piece which has the structure of $G^2(\mathcal{F}^3)$. This missing piece is a subset of $S_A(3;1,1,1)$. Finally, $S_A(3;1,1,1)$ is isomorphic to $G^3(\mathcal{F}^4)$. Thus, $S_A(3)$ consists of three Grassmann manifolds, $G^1(\mathcal{F}^4)$, $G^2(\mathcal{F}^4)$, $G^3(\mathcal{F}^4)$ which are not disjoint. The $G^1(\mathcal{F}^4)$ intersects the $G^2(\mathcal{F}^4)$ along a common submanifold which has the structure of $G^1(\mathcal{F}^3)$. The $G^2(\mathcal{F}^4)$ intersects the $G^3(\mathcal{F}^4)$ along a $G^2(\mathcal{F}^3)$.

Remark 1. Theorem 7 shows that if A has block structure $(m_1, 1, \cdots, 1)$, then $S_A(k)$ consists of a finite sequence of Grassmann manifolds which are joined to each other along Grassmannian submanifolds. In general, the Grassmann manifolds have different dimensions, and at the points of the joining submanifolds, $S_A(k)$ is not locally Euclidean. Thus, the connected components of S_A need not be manifolds.

The class of operators with block structure $(m_1, 1, \cdots, 1)$ appears to be the largest class of nilpotent operators for which the structure of S_A can be explicitly described in terms of familiar manifolds. The next result deals with the connectivity of S_A for an arbitrary nilpotent A.

THEOREM 8. *$S_A(k)$ is path-connected.*

COROLLARY. *S_A consists of exactly $n+1$ connected components, namely $S_A(0), \cdots, S_A(n)$.*

From Theorem 7, we know that $S_A(k)$ need not be a manifold. It is natural to ask whether $S_A(k)$ is a union of manifolds. The following theorem shows that this is indeed the case. In fact, the subset of $S_A(k)$ consisting of those elements which have a given cyclic structure is a regular submanifold of $G^k(V)$. Let $(p_1, \cdots, p_l)$ be a partition of k which is compatible with the block structure of A. Let $(c_1, \cdots, c_s)$ be the partition of n which is conjugate to $(m_1, \cdots, m_r)$, and let $(q_1, \cdots, q_d)$ be the partition of k which is conjugate to $(p_1, \cdots, p_l)$.

THEOREM 9. *$S_A(k; p_1, \cdots, p_l)$ is a connected regular submanifold of $G^k(V)$ of dimension $\sum_{i=1}^{d} q_i(c_i - q_i)$ over $\mathcal{F}$.*

Remark 2. In order to prove Theorem 9, we have constructed charts (local parametrizations) for $S_A(k; p_1, \cdots, p_l)$. These charts give a simple parametrization for $S_A(k; p_1, \cdots, p_l)$ which is useful even in applications where the topology is not important—i.e., where only a convenient method for enumerating the invariant subspaces is needed. We discuss this in detail in § 5.

Using Theorems 8 and 9, we can identify all the isolated points of S_A. Since this result does not appear in either [11] or [10], we give a detailed proof.

THEOREM 10. *If A is cyclic, each connected component of S_A contains only one point, so there are $n+1$ isolated points in S_A. If A is not cyclic, the only connected components of S_A which consist of single points are $S_A(0)$ and $S_A(n)$, so there are exactly 2 isolated points in S_A.*

Proof. The case where A is cyclic is treated in Theorem 5, so suppose that A is not cyclic. The connected components $S_A(0)$ and $S_A(n)$ trivially consist of one point each, so we need to show that if $0<k<n$, then $S_A(k)$ contains more than one point. Suppose that $0<k<n$ and that $S_A(k)$ contains only one point. It is easy to show that there is always at least one partition of k which is compatible with the block structure

of A. Let $p \equiv (p_1, \cdots, p_l)$ be such a partition. Since $S_A(k)$ contains only one point, p must be the only partition of k which is compatible with the block structure of A and $S_A(k; p_1, \cdots, p_l)$ must contain only one point.

Let $(m_1, \cdots, m_r)$ be the partition of n corresponding to the block structure of A. Since A is not cyclic, $r > 1$. Let $(c_1, \cdots, c_s)$ and $(q_1, \cdots, q_d)$ be the conjugate partitions of $(m_1, \cdots, m_r)$ and $(p_1, \cdots, p_l)$ respectively. By Theorem 9, $S_A(k; p_1, \cdots, p_l)$ is an $\mathcal{F}$-analytic manifold of dimension $\sum_{i=1}^{d} q_i(c_i - q_i)$. Since $S_A(k; p_1, \cdots, p_l)$ contains only one point, it follows that $q_i = c_i$, $i = 1, \cdots, d$. If $d = s$, then $k = q_1 + \cdots + q_d = c_1 + \cdots + c_s = n$ which is contrary to assumption. Since $k > 0$, it follows that $0 < d < s$. A consequence of the definition of conjugate partition is that $s = m_1$, $d = p_1$, $r = c_1$ and $l = q_1$. Thus, $p_1 < m_1$. Since $c_1 = r > 1$, we have $l = q_1 = c_1 > 1$. Since $p_1 < m_1$ and $l > 1$, it follows that $(p_1 + 1, p_2, \cdots, p_l - 1)$ is a new partition of k which is compatible with the block structure of A, a contradiction. (If $p_l = 1$, then in our notation the new partition would be written $(p_1 + 1, \cdots, p_{l-1})$.) □

4. Geometry of the solution set. In this section, we use the isomorphism theorem (Theorem 2) and the results on the variety of invariant subspaces to characterize the geometric structure of the solution set of the ARE. As before, Γ denotes the space of real symmetric solutions with the Euclidean topology. We assume that (A, B) is controllable and that Γ is nonempty. T is the restriction of $A^+ (\equiv A - BB'K^+)$ to $V_+ (\equiv L^+(A^+))$. By Theorem 2, Γ is isomorphic to S_T.

Our first result follows immediately from Theorem 3.

THEOREM 11. *Γ is compact.*

Remark 3. The compactness of the solution set is a consequence of the fact that the ARE is a quadratic equation. This is in sharp contrast with the situation for a linear matrix equation. If a linear equation has multiple solutions, the solution set is an affine subspace of positive dimension and is therefore never compact. Our assumption that (A, B) is controllable is necessary to guarantee that the ARE is genuinely quadratic. If (A, B) is not controllable, then Γ need not be compact. This is discussed in more detail in our companion article [9]. The compactness of Γ (when (A, B) is controllable) also follows directly from the existence of maximal and minimal solutions. Since $K^- \leqq K \leqq K^+$, $\forall K \in \Gamma$, it follows that Γ is bounded as a subset of the vector space of real symmetric matrices. Since Γ is closed, this implies that Γ is compact.

We let m denote the dimension of V_+. Let R be the complexification of the operator T. Let $E_1, \cdots, E_p$ be the generalized eigenspaces of T corresponding to its real eigenvalues, and let $F_1, \bar{F}_1, \cdots, F_q, \bar{F}_q$ be the generalized eigenspaces (in conjugate pairs) of R corresponding to the nonreal eigenvalues of T. Let m_i and n_j denote the dimensions of E_i and F_j, respectively. Let T_i be the restriction of T to E_i $(i = 1, \cdots, p)$, and let R_j be the restriction of R to F_j $(j = 1, \cdots, q)$. By Theorem 4b, the mapping $\eta : S_T \to S_{T_1} \times \cdots \times S_{T_p} \times S_{R_1} \times \cdots \times S_{R_q}$ defined by $\eta(S) \equiv (S \cap E_1, \cdots, S \cap E_p, S^+ \cap F_1, \cdots, S^+ \cap F_q)$ is an isomorphism. (S^+ denotes the complexification of the subspace S.)

Recall that for an arbitrary linear operator P on a vector space V, $S_P = S_{P - \lambda I}$. Since T_i and R_j differ from nilpotent operators on E_i and F_j by multiples of I, this means that as far as the structure of S_{T_i} and S_{R_j} is concerned, we can treat T_i and R_j as though they were nilpotent. Using the isomorphism η and the theorems for nilpotent operators from § 3, we obtain the results which follow.

THEOREM 12. *Γ has finite cardinality if and only if T is cyclic. In this case, Γ contains exactly $(m_1 + 1) \cdots (m_p + 1)\ (n_1 + 1) \cdots (n_q + 1)$ elements. If T is not cyclic, then Γ contains a connected component which is a product of real and/or complex projective spaces, at least one of which has positive dimension.*

Proof. If T is cyclic, then the same is true of T_i $(i = 1, \cdots, p)$ and R_j $(j = 1, \cdots, q)$. From Theorem 6, S_{T_i} and S_{R_j} are finite and contain $m_i + 1$ and $n_j + 1$ points, respectively. Since η is a bijection, S_T (and hence Γ) is finite and contains exactly $(m_1 + 1) \cdots (m_p + 1)(n_1 + 1) \cdots (n_q + 1)$ points. Conversely, if T is not cyclic, then at least one of the T_i or R_j is not cyclic. If T_i is not cyclic, Theorem 6 states that S_{T_i} contains a connected component which is a real projective space of positive dimension, while if R_j is not cyclic, then S_{R_j} contains a connected component which is a complex projective space of positive dimension. If T_i or R_j is cyclic, then its connected components are trivially zero-dimensional projective spaces. Since η is an isomorphism, it follows that S_T (and hence Γ) contains a connected component which is a product of projective spaces, at least one of which has positive dimension. □

Remark 4. An interesting consequence of Theorem 12 is that the solution set of the ARE is never countably infinite. It is either finite or it contains at least one continuous family of solutions.

It is appropriate to make some historical comments regarding Theorem 12. The conclusion that Γ is finite if and only if T is cyclic is a purely set-theoretic result and as such follows immediately from the Willems–Coppel bijection. This was noted by Willems in [13]. On the other hand, the last conclusion in Theorem 12 is a topological result and thus depends on our isomorphism theorem (Theorem 2).

We also compare Theorem 12 to some conclusions made by Rodriguez-Canabal [7], [8]. Let $h(s)$ denote the characteristic polynomial of the Hamiltonian matrix $H = \left[\begin{smallmatrix} A & -BB' \\ -Q & -A' \end{smallmatrix}\right]$. Each polynomial $g(s)$ satisfying $h(s) = (-1)^n g(s)g(-s)$ is called a factorization of $h(s)$. The basic result in [7], [8] is that if $K \in \Gamma$, then there exists a factorization $g(s)$ such that $g(H)\left[\begin{smallmatrix} I \\ K \end{smallmatrix}\right] = \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right]$. In fact, $g(s)$ is the characteristic polynomial of $A - BB'K$. (See our companion paper [9] for further discussion.) Rodriguez-Canabal observed that Γ can be infinite if there exists a factorization $g(s)$ which coincides with the minimal polynomial of H. This observation is not equivalent to the first statement in Theorem 12. For example, consider the ARE

$$K^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

It is easily seen that this ARE has infinitely many real symmetric solutions. In fact,

$$K^+ = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Since $A = 0$ and $B = I$, $A^+ = -K^+$. Thus, A^+ is not cyclic, so Theorem 12 implies that Γ is infinite. However, it is easily checked that the minimal polynomial of H is $(s+1)(s+2)(s-1)(s-2)$. Since this polynomial has degree 4, it cannot coincide with any factorization $g(s)$ of $h(s)$.

Rodriguez-Canabal also gives some formulas for the number of real symmetric solutions. However, as is noted in [9], there is not generally a one-to-one correspondence between the set of factorizations of $h(s)$ and the set Γ of real symmetric solutions. Thus, the formulas in [7], [8] are not valid without additional restrictions which are not mentioned. We discuss this further in Remark 6 below.

The next result gives a complete description of Γ in the case where T is semisimple (diagonalizable).

THEOREM 13. *If T is semisimple, then Γ is isomorphic to a disjoint union of products of real and/or complex Grassmann manifolds. Specifically,*

$$\Gamma \approx \bigsqcup_{k_1=0}^{m_1} \cdots \bigsqcup_{k_p=0}^{m_p} \bigsqcup_{l_1=0}^{n_1} \cdots \bigsqcup_{l_q=0}^{n_q} G^{k_1}(\mathbb{R}^{m_1}) \times \cdots \times G^{k_p}(\mathbb{R}^{m_p}) \times G^{l_1}(\mathbb{C}^{n_1}) \times \cdots \times G^{l_q}(\mathbb{C}^{n_q}).$$

Proof. T is semisimple if and only if T_i is semisimple ($i = 1, \cdots, p$) and R_j is semisimple ($j = 1, \cdots, q$). If T_i is semisimple, then by Theorem 5,

$$S_{T_i} = \bigsqcup_{k_i=0}^{m_i} G^{k_i}(E_i) \approx \bigsqcup_{k_i=0}^{m_i} G^{k_i}(\mathbb{R}^{m_i}).$$

If R_j is semisimple, then

$$S_{R_j} = \bigsqcup_{l_j=0}^{n_j} G^{l_j}(F_j) \approx \bigsqcup_{l_j=0}^{n_j} G^{l_j}(\mathbb{C}^{n_j}).$$

The result follows immediately. □

Remark 5. Using Theorem 7, we can completely describe the geometric structure of Γ if the Jordan canonical form of T contains at most one nontrivial block for each eigenvalue. In this case, each of the operators T_i ($i = 1, \cdots, p$) and R_j ($j = 1, \cdots, q$) satisfy the assumptions of the theorem, so each connected component of S_T (and hence of Γ) is a product of "joined" Grassmannians. The connected components are not generally manifolds.

The next theorem describes the connected components of Γ.

THEOREM 14. *Γ has exactly $(m_1+1)\cdots(m_p+1)(n_1+1)\cdots(n_q+1)$ connected components. If $K_1, K_2 \in \Gamma$, then K_1 and K_2 belong to the same connected component if and only if $A-BB'K_1$ and $A-BB'K_2$ have the same set of eigenvalues.*

Proof. It follows from Theorem 8 and the isomorphism η that S_T (and hence Γ) has exactly $(m_1+1)\cdots(m_p+1)(n_1+1)\cdots(n_q+1)$ connected components. It also follows that if $S_1, S_2 \in S_T$, then S_1 and S_2 belong to the same connected component of S_T if and only if $\dim S_1 \cap E_i = \dim S_2 \cap E_i$ ($i = 1, \cdots, p$) and $\dim S_1^+ \cap F_j = \dim S_2^+ \cap F_j$ ($j = 1, \cdots, q$). But this is true if and only if the restrictions $T|S_1$ and $T|S_2$ have the same set of eigenvalues. Let $K_1, K_2 \in \Gamma$ be the solutions of the ARE which correspond to S_1 and S_2. By Theorem 1, $L^+(A-BB'K_1) = S_1$ and $L^+(A-BB'K_2) = S_2$. It also follows from Theorem 1 that $A-BB'K_1$ agrees with T on S_1 and $A-BB'K_2$ agrees with T on S_2. This implies that the left half-plane eigenvalues of $A-BB'K_1$ are the eigenvalues of $T|S_1$, and the left half-plane eigenvalues of $A-BB'K_2$ are the eigenvalues of $T|S_2$. Since K_1 and K_2 belong to the same connected component of Γ if and only if S_1 and S_2 belong to the same connected component of S_T, we see that K_1 and K_2 belong to the same connected component of Γ if and only if $A-BB'K_1$ and $A-BB'K_2$ have the same left half-plane eigenvalues. In § 2 we noted that for any $K \in \Gamma$, $L^0(A-BB'K) = V_0$ (=ker Δ) and $A-BB'K$ agrees with A^+ on V_0. In particular, this implies that $A-BB'K_1$ and $A-BB'K_2$ have the same imaginary axis eigenvalues. Now, let H be the $2n \times 2n$ Hamiltonian matrix $\left[\begin{smallmatrix} A & -BB' \\ -Q & -A' \end{smallmatrix}\right]$. It is easy to show that for any $K \in \Gamma$, H is similar to $\left[\begin{smallmatrix} A-BB'K & -BB' \\ 0 & -(A-BB'K)' \end{smallmatrix}\right]$. Thus, $\det(Is-H) = \det(Is-(A-BB'K))\det(Is+(A-BB'K))$ which shows that the left half-plane eigenvalues of $A-BB'K$ uniquely determine the right half-plane eigenvalues of $A-BB'K$. We conclude that K_1 and K_2 belong to the same connected component of Γ if and only if $A-BB'K_1$ and $A-BB'K_2$ have the same set of eigenvalues. □

Remark 6. Theorem 14 shows that two solutions of the ARE belong to the same connected component of Γ if and only if they yield the same closed loop characteristic polynomial. As mentioned above in reference to the work of Rodriguez-Canabal and in the preceding proof, each $K \in \Gamma$ gives rise to a factorization $h(s) = (-1)^n g(s)g(-s)$ of the characteristic polynomial of H. In fact, $g(s) = \det(Is - (A - BB'K))$. It follows immediately from Theorem 14 that if (A, B) is controllable, then two solutions $K_1, K_2 \in \Gamma$ give rise to the same factorization of $h(s)$ if and only if K_1 and K_2 belong to the same connected component of Γ. It is easy to see that there are exactly $(m_1+1)\cdots(m_p+1)(n_1+1)\cdots(n_q+1)$ distinct factorizations of $h(s)$. Since this equals the number of connected components, we conclude that the factorizations of $h(s)$ are in one-to-one correspondence not with the set Γ but with the connected components of Γ. This is true provided that (A, B) is controllable and Γ is nonempty. In the special case where T is cyclic, each connected component consists of one point, so there is a one-to-one correspondence between the factorizations of $h(s)$ and the set Γ. Hence if Γ is nonempty but finite and (A, B) is controllable, then the cardinality of Γ can be obtained by counting factorizations. This provides the theoretical justification for the counting procedure of Rodriguez-Canabal. We emphasize that this procedure fails if (A, B) is not controllable. Geometrically, this is because some n-dimensional Lagrangian H-invariant subspaces may intersect $\operatorname{Sp}\left[\begin{smallmatrix}0\\I\end{smallmatrix}\right]$ and hence fail to correspond to any solutions of the ARE. (See [9] for further discussion.)

For applications, it is useful to know which solutions are isolated. Let δ_i be equal to $m_i + 1$ if T_i is cyclic and equal to 2 otherwise $(i = 1, \cdots, p)$. Let ε_j be equal to $n_j + 1$ if R_j is cyclic and equal to 2 otherwise $(j = 1, \cdots, q)$.

THEOREM 15. *Γ contains exactly $\delta_1\delta_2\cdots\delta_p\varepsilon_1\varepsilon_2\cdots\varepsilon_q$ isolated solutions. In particular, Γ contains at least 2^{p+q} and at most $(m_1+1)\cdots(m_p+1)(n_1+1)\cdots(n_q+1)$ isolated solutions.*

Proof. This is an immediate consequence of Theorem 10 and the isomorphism η. □

Remark 7. Using Theorem 10 and the isomorphism η, we can actually identify those solutions which are isolated. Let $K \in \Gamma$ and let $S \in S_T$ be the invariant subspace which corresponds to K. Then K is isolated in Γ if and only if $S \cap E_i = 0$ or $S \cap E_i = E_i$ if T_i is not cyclic $(i = 1, \cdots, p)$ and $S^+ \cap F_j = 0$ or $S^+ \cap F_j = F_j$ if R_j is not cyclic $(j = 1, \cdots, q)$. In particular, if $S \cap E_i = 0$ or $S \cap E_i = E_i$ $(i = 1, \cdots, p)$ and $S^+ \cap F_j = 0$ or $S^+ \cap F_j = F_j$ $(j = 1, \cdots, q)$, then K is isolated regardless of which (if any) of the T_i and R_j are cyclic. The solutions satisfying this condition are the 2^{p+q} isolated solutions whose existence is guaranteed by Theorem 15.

In the proof of Theorem 14 we noted that if $K \in \Gamma$ and H is the $2n \times 2n$ Hamiltonian matrix $\left[\begin{smallmatrix}A & -BB'\\ -Q & -A'\end{smallmatrix}\right]$, then $\det(Is - H) = \det(Is - (A - BB'K))\det(Is + (A - BB'K))$. If $p_K(s) \equiv \det(Is - (A - BB'K))$, then $\det(Is - H) = p_K(s)(-1)^n p_K(-s)$. The 2^{p+q} solutions described above are precisely those solutions K for which $p_K(s)$ and $p_K(-s)$ have only pure imaginary roots in common. If λ is an eigenvalue of H with nonzero real part, and K is such a solution, then either $A - BB'K$ has the eigenvalue λ with maximum multiplicity and the eigenvalue $-\lambda$ with zero multiplicity or it has $-\lambda$ with maximum multiplicity and λ with zero multiplicity. For this reason, we call such a K an *unmixed* solution. In particular, K^+ and K^- are such solutions. The unmixed solutions are interesting because they share some of the useful properties of the extreme solutions K^+ and K^-. This is discussed in detail in our companion paper [9].

If the T_i $(i = 1, \cdots, p)$ and R_j $(j = 1, \cdots, q)$ are neither cyclic nor semisimple, then the connected components of S_T (and hence of Γ) are not generally manifolds.

(See Remark 5 for a class of examples.) However, we will show that each connected component is a (set-theoretic) disjoint union of products of analytic manifolds.

Let $K_1, K_2 \in \Gamma$, and let $S_1, S_2 \in S_T$ be the subspaces which correspond to K_1 and K_2 respectively. In other words, $\phi(S_1) = K_1$ and $\phi(S_2) = K_2$. Then K_1 and K_2 belong to the same connected component of Γ if and only if $\dim S_1 \cap E_i = \dim S_2 \cap E_i$ $(i = 1, \cdots, p)$ and $\dim S_1^+ \cap F_j = \dim S_2^+ \cap F_j$ $(j = 1, \cdots, q)$ – i.e., if and only if $T|S_1$ and $T|S_2$ have the same set of eigenvalues. (In particular, $\dim S_1 = \dim S_2$.) Let k_i and l_j be integers such that $0 \leqq k_i \leqq m_i$ and $0 \leqq l_j \leqq n_j$ $(i = 1, \cdots, p,\ j = 1, \cdots, q)$. Let $\Gamma(k_1, \cdots, k_p, l_1, \cdots, l_q)$ denote the connected component of Γ defined by $\{K \in \Gamma$: $\dim \phi^{-1}(K) \cap E_i = k_i \quad (i = 1, \cdots, p), \quad \dim [\phi^{-1}(K)]^+ \cap F_j = l_j \quad (j = 1, \cdots, q)\}$. ($[\phi^{-1}(K)]^+$ is the complexification of the subspace $\phi^{-1}(K)$.)

Let $\boldsymbol{\alpha}^i \equiv (\alpha_1^i, \cdots, \alpha_{r_i}^i)$ be the partition of m_i corresponding to the block structure of T_i $(i = 1, \cdots, p)$, and let $\boldsymbol{\beta}^j \equiv (\beta_1^j, \cdots, \beta_{s_j}^j)$ be the partition of n_j corresponding to the block structure of R_j $(j = 1, \cdots, q)$. Let $\mathbf{a}^i \equiv (a_1^i, \cdots, a_{u_i}^i)$ be a partition of k_i which is compatible with the block structure of T_i, and let $\mathbf{b}^j \equiv (b_1^j, \cdots, b_{v_j}^j)$ be a partition of l_j which is compatible with the block structure of R_j. Let $\boldsymbol{\gamma}^i \equiv (\gamma_1^i, \cdots, \gamma_{r_i'}^i)$, $\boldsymbol{\delta}^j \equiv (\delta_1^j, \cdots, \delta_{s_j'}^j)$, $\mathbf{e}^i \equiv (e_1^i, \cdots, e_{d_i}^i)$ and $\mathbf{f}^j \equiv f_1^j, \cdots, f_{g_j}^j)$ denote the conjugate partitions of $\boldsymbol{\alpha}^i$, $\boldsymbol{\beta}^j$, $\mathbf{a}^i$ and $\mathbf{b}^j$ respectively. Let $\Gamma((k_1; \mathbf{a}^1), \cdots, (k^p; \mathbf{a}^p), (l_1; \mathbf{b}^1), \cdots, (l_q; \mathbf{b}^q))$ denote the subset of $\Gamma(k_1, \cdots, k_p, l_1, \cdots, l_q)$ defined by $\{K \in \Gamma$: $\phi^{-1}(K) \cap E_i$ has cyclic structure $\mathbf{a}^i$ $(i = 1, \cdots, p)$, $[\phi^{-1}(K)]^+ \cap F_j$ has cyclic structure $\mathbf{b}^j$ $(j = 1, \cdots, q)\}$. By Theorem 4b, it follows that the image of $\Gamma((k_1; \mathbf{a}^1), \cdots, (k_p; \mathbf{a}^p), (l_1; \mathbf{b}^1), \cdots, (l_q; \mathbf{b}^q))$ under ϕ^{-1} is isomorphic to the product $S_{T_1}(k_1; \mathbf{a}^1) \times \cdots \times S_{T_p}(k_p; \mathbf{a}^p) \times S_{R_1}(l_1; \mathbf{b}^1) \times \cdots \times S_{R_q}(l_q; \mathbf{b}^q)$. By Theorem 9, $S_{T_i}(k_i, \mathbf{a}^i)$ is a real-analytic manifold of dimension $\sum_{\nu=1}^{d_i} e_\nu^i(\gamma_\nu^i - e_\nu^i)$. By the same theorem, $S_{R_j}(l_j; \mathbf{b}^j)$ is a complex-analytic manifold of complex dimension $\sum_{\nu=1}^{g_j} f_\nu^j(\delta_\nu^j - f_\nu^j)$ and therefore a real-analytic manifold of twice this dimension. The next theorem follows immediately.

THEOREM 16. *$\Gamma((k_1; \mathbf{a}^1), \cdots, (k_p; \mathbf{a}^p), (l_1; \mathbf{b}^1), \cdots, (l_q; \mathbf{b}^q))$ is a real-analytic manifold of dimension*

$$\sum_{i=1}^{p} \sum_{\nu=1}^{d_i} e_\nu^i(\gamma_\nu^i - e_\nu^i) + 2 \sum_{j=1}^{q} \sum_{\nu=1}^{g_j} (\delta_\nu^j - f_\nu^j).$$

Note that a connected component of Γ corresponds to the set of all $K \in \Gamma$ for which $T|\phi^{-1}(K)$ has a given set of eigenvalues. By Theorem 16, we see that a manifold is obtained by fixing not only the spectrum of $T|\phi^{-1}(K)$ but also its Jordan canonical form.

Since $\Gamma((k_1; \mathbf{a}^1), \cdots, (k_p; \mathbf{a}^p), (l_1; \mathbf{b}^1), \cdots, (l_q; \mathbf{b}^q))$ is a manifold, each of its points has a neighborhood which is homeomorphic to $\mathbb{R}^r$, where r is the dimension given in Theorem 16. In other words, the manifold can be locally parametrized by vectors of r real numbers. Such a parametrization is called a *chart*. By the isomorphism theorem (Theorem 2), the problem of constructing charts for the subsets $\Gamma((k_1, \mathbf{a}^1), \cdots, (k_p; \mathbf{a}^p), (l_1; \mathbf{b}^1), \cdots, (l_q; \mathbf{b}^q))$ of Γ is equivalent to the problem of constructing charts for the images of these subsets under ϕ^{-1}. Thus, the problem becomes that of parametrizing (subsets of) the variety of invariant subspaces of a finite-dimensional linear operator. A solution to this problem is given in the final section.

5. Parametrizing the variety of invariant subspaces. In this section, we describe a solution to the problem of parametrizing the set of all invariant subspaces of a finite dimensional linear operator. This parametrization provides the charts for the subset $\Gamma((k_1, \mathbf{a}^1), \cdots, (k_p; \mathbf{a}^p), (l_1; \mathbf{b}^1), \cdots, (l_q; \mathbf{b}^q))$ of Γ. However, it is of interest for other

reasons as well. To use either Willems' theorem (Theorem 1) or the Hamiltonian matrix method to generate all the solutions of the algebraic Riccati equation, some method of listing the invariant subspaces is essential. In fact, a parametrization for the set of invariant subspaces would be useful in many problems in systems theory and applied mathematics. The lack of such a parametrization has apparently led many authors to restrict their results to operators which have distinct eigenvalues or which are diagonalizable.

As in § 3, V is an n-dimensional vector space over the field $\mathcal{F}$ of real or complex numbers and $A \in \mathrm{Hom}\,(V)$. By choosing an appropriate basis, we can assume that A is a matrix in lower Jordan canonical form. Of course, this may require complexification. Using the bijection η described in Theorems 4a and 4b, we see that there is no loss of generality in assuming that A has exactly one distinct eigenvalue, say λ. Since A and $A-\lambda I$ have the same invariant subspaces, we lose nothing by assuming that A is nilpotent.

Let A be a nilpotent matrix in lower Jordan canonical form which acts on $\mathcal{F}^n$. Let $(m_1, \cdots, m_r)$ be the block structure of A. Fix k such that $0 \leqq k \leqq n$, and let $(p_1, \cdots, p_l)$ be a partition of k which is compatible with the block structure of A. Let $S \in S_A(k; p_1, \cdots, p_l)$. Since S has cyclic structure $(p_1, \cdots, p_l)$, it has an ordered basis of the form $\{v_1, Av_1, \cdots, A^{p_1-1}v_1, v_2, Av_2, \cdots, A^{p_2-1}v_2, \cdots, v_l, Av_l, \cdots, A^{p_l-1}v_l\}$, with $A^{p_i}v_i = 0$, $i = 1, \cdots, l$. Conversely, if S is any subspace which has a basis of this form, then $S \in S_A(k; p_1, \cdots, p_l)$. We call such an ordered basis a *cyclic basis* for S.

Let $S \in S_A(k; p_1, \cdots, p_l)$, and let B be an $n \times k$ rank k matrix whose columns form a cyclic basis for S. Partition the rows of B according to the partition $(m_1, \cdots, m_r)$ of n, and partition the columns according to the partition $(p_1, \cdots, p_l)$ of k. Then B consists of rl blocks, and the ijth block, B_{ij}, is $m_i \times p_j$. It is easy to verify that B_{ij} has the following structure: (i) B_{ij} is constant along diagonals; (ii) if the diagonals of B_{ij} are numbered starting with the *lower* left-hand corner and if a_t is the constant value of the entries on the tth diagonal ($t = 1, \cdots, m_i + p_j - 1$), then $a_t = 0$ for $t > \min(m_i, p_j)$. A matrix with properties (i) and (ii) will be called *regular lower triangular* (RLT). A partitioned matrix whose blocks are RLT matrices will be called *block regular lower triangular* (BRLT). Thus, if B is a matrix whose columns form a cyclic basis for a subspace $S \in S_A(k; p_1, \cdots, p_l)$, then B is an $n \times k$ rank k matrix which is BRLT when the rows are partitioned according to $(m_1, \cdots, m_r)$ and the columns are partitioned according to $(p_1, \cdots, p_l)$. Conversely, if B is such a matrix, then Sp B (the column span of B) is an element of $S_A(k; p_1, \cdots, p_l)$. We let $\mathcal{B}(k; p_1, \cdots, p_l)$ denote the set of all such matrices. Note that the row partition $(m_1, \cdots, m_r)$ corresponding to the block structure of A is suppressed.

Example 3. Let $(m_1, m_2, m_3) = (4, 3, 1)$. Let $k = 5$ and let $(p_1, p_2) = (3, 2)$. Then the elements of $\mathcal{B}(5; 3, 2)$ are the full rank matrices of the form

$$B = \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 0 & 0 \\ x_2 & 0 & 0 & 0 & 0 \\ x_3 & x_2 & 0 & y_3 & 0 \\ x_4 & x_3 & x_2 & y_4 & y_3 \\ \hline x_5 & 0 & 0 & 0 & 0 \\ x_6 & x_5 & 0 & y_6 & 0 \\ x_7 & x_6 & x_5 & y_7 & y_6 \\ \hline x_8 & 0 & 0 & y_8 & 0 \end{array}\right].$$

The problem with using $\mathcal{B}(k; p_1, \cdots, p_l)$ to parametrize $S_A(k; p_1, \cdots, p_l)$ is that the correspondence is not one-to-one. In general, many matrices in $\mathcal{B}(k; p_1, \cdots, p_l)$ span the same element of $S_A(k; p_1, \cdots, p_l)$. To obtain a one-to-one correspondence, we define an equivalence relation on $\mathcal{B}(k; p_1, \cdots, p_l)$. If $B_1, B_2 \in \mathcal{B}(k; p_1, \cdots, p_l)$, then $B_1 \sim B_2$ if and only if $\operatorname{Sp} B_1 = \operatorname{Sp} B_2$ – i.e., if and only if B_1 and B_2 correspond to the same element of $S_A(k; p_1, \cdots, p_l)$. In order to parametrize $S_A(k; p_1, \cdots, p_l)$ by elements of $\mathcal{B}(k; p_1, \cdots, p_l)$, we must choose a representative from each equivalence class. To do this, we define canonical forms on $\mathcal{B}(k; p_1, \cdots, p_l)$ with respect to the equivalence relation $\sim$.

Let $\gamma = (\gamma_1, \cdots, \gamma_l)$ be a multi-index of length l such that $\gamma_1, \cdots, \gamma_l$ are distinct integers between 1 and r, not necessarily in increasing order. We say that γ is a *compatible* multi-index if $m_{\gamma_j} \geqq p_j$, $j = 1, \cdots, l$, and we let $C(k; p_1, \cdots, p_l)$ denote the set of compatible multi-indices corresponding to the partition $(p_1, \cdots, p_l)$ of k. Since $(p_1, \cdots, p_l)$ is compatible with the block structure of A, $m_j \geqq p_j$, $j = 1, \cdots, l$, so $\gamma = (1, 2, \cdots, l)$ is a compatible multi-index. Hence, $C(k; p_1, \cdots, p_l)$ is nonempty.

For each $\gamma \in C(k; p_1, \cdots, p_l)$ and $B \in \mathcal{B}(k; p_1, \cdots, p_l)$, let $M_\gamma(B)$ be the $k \times k$ submatrix obtained by taking the last p_j rows from the γ_jth block of rows, $j = 1, \cdots, l$. The following result is proved in [10].

PROPOSITION 1. *Let $B \in \mathcal{B}(k; p_1, \cdots, p_l)$. Then there exists $\gamma \in C(k; p_1, \cdots, p_l)$ such that* $\det M_\gamma(B) \neq 0$.

Example 4. As in the previous example, let $(m_1, m_2, m_3) = (4, 3, 1)$, $k = 5$, $(p_1, p_2) = (3, 2)$. Then $C(5; 3, 2) = \{(1, 2), (2, 1)\}$. If B is the matrix in Example 3, then

$$M_{(1,2)}(B) = \left[\begin{array}{ccc|cc} x_2 & 0 & 0 & 0 & 0 \\ x_3 & x_2 & 0 & y_3 & 0 \\ x_4 & x_3 & x_2 & y_4 & y_3 \\ \hline x_6 & x_5 & 0 & y_6 & 0 \\ x_7 & x_6 & x_5 & y_7 & y_6 \end{array}\right], \qquad M_{(2,1)}(B) = \left[\begin{array}{ccc|cc} x_5 & 0 & 0 & 0 & 0 \\ x_6 & x_5 & 0 & y_6 & 0 \\ x_7 & x_6 & x_5 & y_7 & y_6 \\ \hline x_3 & x_2 & 0 & y_3 & 0 \\ x_4 & x_3 & x_2 & y_4 & y_3 \end{array}\right].$$

By Proposition 1, at least one of these submatrices is nonsingular.

For each $\gamma \in C(k; p_1, \cdots, p_l)$, let $V_\gamma \equiv \{B \in \mathcal{B}(k; p_1, \cdots, p_l)\colon \det M_\gamma(B) \neq 0\}$. By Proposition 1, $\{V_\gamma\}_{\gamma \in C(k; p_1, \cdots, p_l)}$ is a cover of $\mathcal{B}(k; p_1, \cdots, p_l)$. In fact, each V_γ is an open and dense subset of $\mathcal{B}(k; p_1, \cdots, p_l)$. It is easy to see that if $B_1, B_2 \in \mathcal{B}(k; p_1, \cdots, p_l)$ and $B_1 \sim B_2$ then $B_1 \in V_\gamma$ if and only if $B_2 \in V_\gamma$. Thus, V_γ is a union of whole equivalence classes. The next proposition combines several results in [10].

PROPOSITION 2. *Let $B \in V_\gamma$. Then there is a unique matrix $\hat{B}$ in the equivalence class of B such that $M_\gamma(\hat{B})$ has the following form: Let M_{ij} denote the ijth block of $M_\gamma(\hat{B})$ when both the rows and the columns are partitioned according to the partition $(p_1, \cdots, p_l)$ of k. (So M_{ij} is $p_i \times p_j$.) Then the first p_i diagonals of $M_{i1}, \cdots, M_{il}$ (counting from the lower left-hand corner) are zero except for the p_ith diagonal of M_{ii} which is equal to one.*

Remark 8. From the definition of $M_\gamma(B)$ and the fact that B is BRLT, it follows that the diagonals of M_{ij} above the p_jth are automatically zero. This implies that M_{ii} is a $p_i \times p_i$ identity matrix and $M_{ij} = 0$ for $i < j$.

We will say that $B \in V_\gamma$ is in *canonical form with respect to* γ if $M_\gamma(B)$ has the special form described in Proposition 2. Since each equivalence class which is contained in V_γ has a unique representative which is in canonical form (with respect to γ), these representatives are in one-to-one correspondence with those subspaces in $S_A(k; p_1, \cdots, p_l)$ which are spanned by matrices in V_γ. Thus, the matrices in V_γ which are in canonical form parametrize a subset of $S_A(k; p_1, \cdots, p_l)$. Since the collection $\{V_\gamma\}_{\gamma \in C(k; p_1, \cdots, p_l)}$ is a cover of $\mathcal{B}(k; p_1, \cdots, p_l)$, we have a set of parametrizations (one

for each $\gamma \in C(k; p_1, \cdots, p_l)$) such that every $S \in S_A(k; p_1, \cdots, p_l)$ is contained in at least one parametrization.

Remark 9. It is easy to show that each parametrization parametrizes an open and dense subset of $S_A(k; p_1, \cdots, p_l)$. Thus, for many applications it will suffice to use a single one of these parametrizations—since the subspaces omitted form a negligible subset of $S_A(k; p_1, \cdots, p_l)$.

Example 5. As in Examples 3 and 4, let $(m_1, m_2, m_3) = (4, 3, 1)$, $k = 5$, $(p_1, p_2) = (3, 2)$. Then $S_A(5; 3, 2)$ is covered by two parametrizations which correspond to the two elements of $C(5; 3, 2)$:

$$\left[\begin{array}{ccc|cc} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline x_5 & 0 & 0 & 0 & 0 \\ 0 & x_5 & 0 & 1 & 0 \\ 0 & 0 & x_5 & 0 & 1 \\ \hline x_8 & 0 & 0 & y_8 & 0 \end{array}\right] \qquad \left[\begin{array}{ccc|cc} 0 & 0 & 0 & 0 & 0 \\ x_2 & 0 & 0 & 0 & 0 \\ 0 & x_2 & 0 & 1 & 0 \\ 0 & 0 & x_2 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline x_8 & 0 & 0 & y_8 & 0 \end{array}\right].$$

$\gamma = (1, 2)$ $\qquad$ $\gamma = (2, 1)$

Remark 10. In [10], we defined a restricted class of elementary column operations called *elementary cyclic column operations* (ECCO's). Roughly speaking, these are elementary column operations which preserve the BRLT structure of the matrices in $\mathcal{B}(k; p_1, \cdots, p_l)$. Using ECCO's, one can design a simple algorithm to put a matrix $B \in V_\gamma$ in its canonical form with respect to γ. In fact, such an algorithm is implicitly used in the proofs in [10].

Acknowledgment. The results described in this paper represent a portion of the author's doctoral thesis [11] completed under Professor R. W. Brockett. The author also thanks Professor C. I. Byrnes for many helpful discussions.

REFERENCES

[1] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[2] C. I. BYRNES AND C. F. MARTIN, eds., *Geometrical Methods for the Theory of Linear Systems*, Reidel, Dordrecht, 1980.

[3] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377–401.

[4] L. FINESSO AND G. PICCI, *A characterization of minimal square spectral factors*, IEEE Trans. Automat. Control 27 (1982), pp. 122–127.

[5] K. MÅRTENSSON, *On the matrix Riccati equation*, Inform. Sci., 3 (1971), pp. 17–49.

[6] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496–501.

[7] J. RODRIGUEZ-CANABAL, *The geometry of the Riccati equation*, Stochastics, 1 (1973), pp. 129–149.

[8] ———, *The geometry of the Riccati equation*, Ph.D. thesis, Univ. Southern California, Los Angeles, 1972.

[9] M. A. SHAYMAN, *Geometry of the algebraic Riccati equation, Part* I, this Journal, this issue, pp. 375–394.

[10] ———, *On the variety of invariant subspaces of a finite-dimensional linear operator*, Trans. Amer. Math. Soc., to appear.

[11] ———, *Varieties of invariant subspaces and the algebraic Riccati equation*, Ph.D. thesis, Harvard University, Cambridge, MA, 1980.

[12] A. C. M. VAN SWIETEN, *Qualitative behavior of dynamical games with feedback strategies*, Ph.D. thesis, University of Groningen, the Netherlands, 1977.

[13] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control., 16 (1971), pp. 621–634.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0005 $01.25/0

COMMENTS ON "IDENTIFIABILITY OF SPATIALLY VARYING AND CONSTANT PARAMETERS IN DISTRIBUTED SYSTEMS OF PARABOLIC TYPE"*

M. COURDESSES†

Abstract. In the paper [SIAM J. Control and Optim., 15 (1977), pp. 785–802], S. Kitamura and S. Nakagiri gave results on identifiability of constant parameters of a system described by a linear, one-dimensional parabolic partial differential equation with pointwise measurement. Unfortunately, their expression of the solution seems to be incorrect and consequently some of their results are, too. A new expression for the solution is given here, as well as a necessary and sufficient condition for identifiability.

1. Introduction. Following the notation from the paper of S. Kitamura and S. Nakagiri [5], let us consider the system described by

$$\frac{\partial u}{\partial t} = a\frac{\partial^2 u}{\partial x^2} + bu + f(x,t), \qquad x \in (0,1), \quad t > 0, \tag{1}$$

with boundary and initial conditions given as

$$\begin{aligned} &\alpha_0 u(t,0) - (1-\alpha_0)\frac{\partial u}{\partial x}(t,0) = g_0(t), \qquad 0 \leqq \alpha_0 \leqq 1, \\ &\alpha_1 u(t,1) + (1-\alpha_1)\frac{\partial u}{\partial x}(t,1) = g_1(t), \qquad 0 \leqq \alpha_1 \leqq 1, \end{aligned} \tag{2}$$

$$u(x,0) = u_0(x). \tag{3}$$

The measured output y is represented by

$$y(t) = u(x_p, t), \qquad t \geqq 0, \tag{4}$$

where x_p denotes the position of a sensor.

Under some assumptions a unique solution of (1), (2), (3) exists. Kitamura and Nakagiri [5] consider the eigenvalue problem associated with (1) and (2) and give the representation of the solution $u(x,t)$ in terms of the eigenvalues k_n and the eigenfunctions $\psi_n(x)$ [3, p. 795]. We remark that $\psi_n(x)$ do not depend on the parameters a and b. In fact the calculation of the solution gives [2], [4]

$$\begin{aligned} u(x,t) = &\sum_{n=1}^{\infty} \langle u_0, \psi_n \rangle e^{-k_n t}\psi_n(x) + \int_0^t \sum_{n=1}^{\infty} e^{-k_n(t-\tau)}\psi_n(x)\langle \psi_n, f\rangle \, d\tau \\ &+ a\int_0^t \sum_{n=1}^{\infty} (\psi_n(1) - \psi_n'(1))\, e^{-k_n(t-\tau)}\psi_n(x) g_1(\tau)\, d\tau \\ &+ a\int_0^t \sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\, e^{-k_n(t-\tau)}\psi_n(x) g_0(\tau)\, d\tau, \end{aligned} \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denote the inner product of $L_2[0,1]$ (see [2, p. 29]).

* Received by the editors April 21, 1981.

† Laboratoire d'Automatique et d'Analyse des Systèmes du CNRS, 7, Avenue du Colonel Roche, 31400 Toulouse, France, and Université Paul Sabatier, Toulouse, France.

2. Results on identifiability. The results given in [5, Result 12] must be modified, for the parameter a appears in the expression and [5, Lemma 3] does not apply. First we give:

LEMMA. *Let* $\{k_n\}_{n=1,2,\cdots}$ *and* $\{k_n^m\}_{n=1,2,\cdots}$ *be strictly monotone increasing sequences tending to infinity and let*

$$\sum_{n=1}^{\infty} C_n e^{-k_n t} = \sum_{n=1}^{\infty} C_n^m e^{-k_n^m t} \quad \text{for all } t \in [0, \infty).$$

Assume $\sum_{n=1}^{\infty} |C_n|$ *and* $\sum_{n=1}^{\infty} |C_n^m|$ *to be convergent.*

i) *If* C_n *and* $C_n^m \neq 0$, *for all* $n = 1, \cdots, N$, *then* $k_n = k_n^m$ *and* $C_n = C_n^m$ *for* $n = 1, \cdots, N$.

ii) *If* C_n *and* $C_n^m \neq 0$ *for all* $n = 1, 2, \cdots$, *then* $k_n = k_n^m$ *and* $C_n = C_n^m$ *for* $n = 1, 2, \cdots$.

The proof is omitted [1].

RESULT. *Let* $u_0(x)$ *and* $f(x, t) = 0$ *in* (1)–(3). *Let* $g_0(t)$ *and* $g_1(t)$ *belong to* $C[0, \infty)$ *and let*

(i) $g_0(t) \not\equiv 0$ *and* $g_1(t) \equiv 0$,
(ii) $g_0(t) \equiv 0$ *and* $g_1(t) \not\equiv 0$ *or*
(iii) $g_0(t) \not\equiv 0$ *and* $g_1(t) = \beta g_0(t)$,

where β *satisfies the inequality* $\beta(\psi_n(1) - \psi_n'(1)) \neq -(\psi_n(0) + \psi_n'(0))$ *for all* n, $n \in N$ (*the set of natural numbers*).

Then parameters a *and* b *are always identifiable if and only if* x_p *is such that* $\psi_i(x_p) \neq 0$ *for at least one* $i \in N$.

Proof. We shall prove only the case (i). Under the assumptions we obtain from (5):

$$y(t) = a \int_0^t \left(\sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\, e^{-k_n(t-\tau)} \psi_n(x_p) \right) g_0(\tau)\, d\tau$$

or

$$y(t) = a \int_0^t \left(\sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\, e^{-k_n \tau} \psi_n(x_p) \right) g_0(t-\tau)\, d\tau.$$

Put

$$\alpha(t) = a \sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\, e^{-k_n t} \psi_n(x_p),$$

$$\alpha^m(t) = a_m \sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\, e^{-k_n^m t} \psi_n(x_p).$$

Since the difference of outputs $e(t)$ is zero we have also

$$\int_0^t (\alpha(\tau) - \alpha^m(\tau)) g_0(t-\tau)\, d\tau = 0, \qquad t \geqq 0.$$

Then Titchmarsh's theorem [3, p. 34] gives

$$a \sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\, e^{-k_n t} \psi_n(x_p) = a_m \sum_{n=1}^{\infty} (\psi_n(0) + \psi_n'(0))\, e^{-k_n^m t} \psi_n(x_p), \qquad t \geqq 0.$$

Since $\psi_n(0) + \psi_n'(0) \neq 0$ for all $n \in N$ [5, Lemma 4] and $\psi_i(x_p) \neq 0$, our lemma implies that $a = a_m$ and $k_i = k_i^m$, thus $b = b_m$.

The proof of the necessity is similar to that for [5, Result 10].

Acknowledgments. The author would like to thank Professor G. Ryan of the University of Bath and Professor S. Nakagiri of the Faculty of Engineering, Kobe University, for a number of helpful comments on this paper.

REFERENCES

[1] M. COURDESSES, M. P. POLIS AND M. AMOUROUX, *On identifiability of parameters in a class of parabolic distributed systems*, in Proc. IEEE Conference on Decision and Control, Albuquerque, NM, 1980.

[2] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer, Berlin, 1978.

[3] V. DITKIN AND A. P. PRUDNIKOV, *Integral Transforms and Operational Calculus*, 1st ed., Pergamon Press, New York, 1965.

[4] A. PIERCE, *Unique identification of eigenvalues and coefficients in a parabolic problem*, this Journal, 17 (1979), pp. 494–499.

[5] S. KITAMURA AND S. NAKAGIRI, *Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type*, this Journal, 15 (1977), pp. 785–802.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0006 $01.25/0

WIDE SENSE STATIONARY SOLUTIONS OF LINEAR SYSTEMS WITH ADDITIVE NOISE*

L. ARNOLD† AND V. WIHSTUTZ†

Abstract. Let z be a (wide sense) stationary process with spectral measure F^z, acting as noise on

$$(*) \qquad \dot{x} = Ax + Bz, \qquad y = Gx.$$

The aim of this paper is to describe the set of stationary outputs y. A stationary solution x is called hard if it satisfies $(*)$ as a stochastic process, its spectral measure F^x then satisfies

$$(**) \qquad (i\lambda - A)\, dF^x(\lambda)\, (i\lambda - A)^* = dF^z(\lambda).$$

Any solution of $(**)$ is called a soft stationary solution of $(*)$. If $i\lambda \in \sigma(A)$, λ is called critical.

For $B = G = I$, there exists a soft stationary solution if and only if there exists a hard stationary solution on the original probability space if and only if $(i\lambda - A)^{-1} \in L_2(F^z)$ and image $\Delta F^2(\lambda) \subset$ image $(i\lambda - A)$ for critical λ. The intuitive meaning of this condition is that the critical frequencies of the undisturbed system have to be missing in the noise spectrum in a certain sense. Otherwise we encounter resonance. The soft stationary solution is unique if and only if the hard stationary solution is unique on any probability space if and only if all $\operatorname{Re} \lambda_j(A) \neq 0$. The set of all stationary solutions is described. The results carry over to the observable case for general C.

Key words. wide sense stationary processes, linear stochastic systems

1. Introduction. Notions of solution. Let z be a mean square continuous zero mean wide sense stationary stochastic process ("wide sense" is dropped from now on) on the real line $\mathbb{R}$ with values in $\mathbb{R}^m$ defined on a probability (pr.) space $(\Omega, \mathcal{S}, P)$, with spectral representation

$$z(t) = \int_{-\infty}^{\infty} e^{it\lambda}\, d\zeta(\lambda), \qquad E\, d\zeta(\lambda)\, d\zeta(\lambda)^* = dF^z(\lambda).$$

Here, $\zeta(\lambda)$ is the spectral process of z, and the nonnegative definite matrix $dF^z(\lambda)$ its spectral measure. (The star * denotes complex conjugation and transposition, $M^* = \bar{M}^T$.) Let H^z be the closed linear span in $L_2(\Omega, \mathcal{S}, P)$ of the components of $z(t)$, $t \in \mathbb{R}$, with scalars from $\mathbb{C}$. We define a space $L_2(F^z)$ of $\mathbb{C}^{m\times m}$-valued measurable functions on $\mathbb{R}$ as follows: Let μ be a nonnegative measure with respect to which all F^z_{jk} are absolutely continuous (take, e.g., $d\mu = \sum_1^m dF^z_{jj}$) so that $dF^z/d\mu = f(\lambda) \geqq 0$. Then $A(\lambda) \in L_2(F^z)$ if and only if

$$\int_{-\infty}^{\infty} \operatorname{trace} A(\lambda) f(\lambda) A(\lambda)^*\, d\mu < \infty,$$

i.e., for each $k = 1, \cdots, m$

$$I(A, A)_{kk} = \int_{-\infty}^{\infty} \left(\sum_p \sum_q a_{kp}(\lambda) \bar{a}_{kq}(\lambda) f_{pq}(\lambda) \right) d\mu < \infty.$$

Note that it is not required that $\int a_{kp} \bar{a}_{kq} f_{pq}\, d\mu$ exists. The integral $I(A, B) = \int A(\lambda)\, dF^z(\lambda) B(\lambda)^*$ for A, $B \in L_2(F^z)$ is then defined as the matrix with (j, k)th element given by

$$I(A, B)_{jk} = \int_{-\infty}^{\infty} \left(\sum_p \sum_q a_{jp} f_{pq} \bar{b}_{kq} \right) d\mu.$$

* Received by the editors September 21, 1981 and in revised form May 4, 1982.

† Universität Bremen, D2800 Bremen 33, Federal Republic of Germany, and Courant Institute, New York University, New York, New York 10012.

$L_2(F^z)$ is a Hilbert space with scalar product $(A, B) = \text{trace } I(A, B)$, which is isometric to the Hilbert space $(H^z)^m$ with scalar product Eh^*g. Finally, let U_t be the unitary group acting on H^z such that $U_t z(0) = z(t)$. Then the linear time-invariant system

$$\dot{x} = Ax + Bz, \qquad x(0) = x_0, \quad A\ d \times d\text{-matrix}, B\ d \times m\text{-matrix}, \tag{1.1}$$

assigns to any initial r.v. $x_0 \in L_2$ the unique solution

$$x(t) = e^{tA}(x_0 + Z(t)), \qquad Z(t) = \int_0^t e^{-sA} Bz(s)\, ds, \qquad t \in \mathbb{R}, \tag{1.2}$$

(all derivatives and integrals are taken in the mean square sense). We restrict ourselves to zero mean solutions, i.e., to x_0 with $Ex_0 = 0$. Later, in § 3, we will add to (1.1) a linear read-out map $y = Gx$, G a $p \times d$-matrix.

The aim of this paper is to clarify as completely as possible the problem of existence and uniqueness of stationary solutions of (1.1).

A stationary solution x that is given as a stochastic process on some probability space which also carries z is called a *hard* solution. In view of $\dot{x} - Ax = Bz$ the search for hard solutions amounts to looking for $x(t) = \int e^{it}\, d\xi(\lambda)$ for which $(i\lambda - A)\, d\xi(\lambda) = B\, d\zeta(\lambda)$. (Such a solution is thus always stationarily connected with Bz, i.e., the pair (x, Bz) is stationary.) For a nonsingular B this means that $H^z \subset H^x$. We are interested in the converse, $H^x \subset H^z$, i.e., in those solutions which are actually driven by the noise, for which there is thus an $x_0 \in H^z$ such that $x(t) = U_t x_0$. Those x are called *subordinated* (cf. Gikhman and Skorokhod [4, p. 242]).

The spectral measure $dF^x(\lambda) = E\, d\xi(\lambda)\, d\xi(\lambda)^*$ of a (real) stationary solution x satisfies

$$(i\lambda - A)\, dF^x(\lambda)\, (i\lambda - A)^* = B\, dF^z(\lambda) B^*, \qquad dF^x(-\lambda) = \overline{dF^x(\lambda)}. \tag{1.3}$$

Any spectral measure F^x satisfying (1.3) for given F^z is said to be a *soft* solution of (1.1). It is called *subordinated* if $dF^x(\lambda) = R(\lambda)\, dF^z(\lambda) R(\lambda)^*$, $R \in L_2(F^z)$. A soft solution is said to be *realized* on a probability space if there is a hard solution on that space with F^x satisfying (1.3). A subordinated soft solution can always be realized via $d\xi(\lambda) = R(\lambda)\, d\zeta(\lambda)$ whenever z can be defined on that space.

The situation is simple if all $\operatorname{Re} \lambda_j(A) \neq 0$. Then there are always unique soft and hard solutions (which are subordinated) given by $dF^x(\lambda) = (i\lambda - A)^{-1} B\, dF^z(\lambda) B^* (i\lambda - A)^{*-1}$ and $d\xi(\lambda) = (i\lambda - A)^{-1} B\, d\zeta(\lambda)$, respectively. In case all $\operatorname{Re} \lambda_j(A) < 0$ we use $(i\lambda - A)^{-1} = \int_{-\infty}^0 \exp(s(i\lambda - A))\, ds$ to convert x into the more familiar form

$$x(t) = \int_{-\infty}^{\infty} e^{it\lambda} (i\lambda - A)^{-1} B\, d\zeta(\lambda) = \int_{-\infty}^{t} e^{(t-s)A} Bz(s)\, ds,$$

similarly if all $\operatorname{Re} \lambda_j(A) > 0$ (see Bunke [3, p. 46ff.]).

It remains to consider the case where some $\operatorname{Re} \lambda_j(A) = 0$. We call a $\lambda \in \mathbb{R}$ *critical* if $i\lambda$ is an eigenvalue of A (i.e., $i\lambda - A$ is singular), otherwise it is called *noncritical*. Of course, if λ is critical, so is $-\lambda$.

All results presented in this paper remain valid for the discrete time system $x_{n+1} = Ax_n + z_n$ if the obvious changes are made. To our knowledge, only some particular cases with critical λ's have been treated (see Wentzell [6, p. 56], for the nth order scalar equation, or Arnold, Horsthemke and Stucki [1] for $d = 2$, $\lambda_{1,2} = \pm i$).

Analogous problems can be posed in the framework of strict sense stationary processes (see Arnold and Wihstutz [2]). The white noise case was treated by Snyders [5].

2. The case $\dot{x} = Ax + z$.

2.1. Soft solutions. We have to solve (1.3) for $B = I$. For each fixed critical λ, (1.3) becomes the algebraic equation

$$TPT^* = C, \qquad C \geqq 0, \tag{2.1}$$

where $T = i\lambda - A$, $P = \Delta F^x(\lambda) \geqq 0$, $C = \Delta F^z(\lambda) \geqq 0$, with its homogeneous version

$$TPT^* = 0. \tag{2.2}$$

Note that F^x may jump at λ even if F^z is continuous, because $(i\lambda - A)$ is singular. Let us recall a few facts from linear algebra (all subspaces and linear operators are in $\mathbb{C}^d$):

$\mathbb{C}^d = \operatorname{im} T \oplus \ker T^*$, $\operatorname{im} T \perp \ker T^*$ for all T, for Hermitian C $\operatorname{im} C \perp \ker C$. For $C = DD^* \geqq 0$ $\operatorname{im} C = \operatorname{im} D$ and $\ker C = \ker D^*$. $\sqrt{C} \geqq 0$ denotes the square root of $C \geqq 0$. For any linear operator

$$T: \bigoplus_{j=1}^{p} X_j = \mathbb{C}^d \to \mathbb{C}^d = \bigoplus_{k=1}^{q} Y_k$$

there exists a unique decomposition $T = \sum_{j=1}^{p} \sum_{k=1}^{q} T_{kj}$ into linear operators $T_{kj}: \mathbb{C}^d \to \mathbb{C}^d$ with $T_{kj}(X_j) \subset Y_k$, $T_{kj}(X_l) = 0$ for all $l \neq j$. Also, the restriction T_0 of T to $\operatorname{im} T^*$ is a bijection $T_0: \operatorname{im} T^* \to \operatorname{im} T$. Consequently, if for $C \geqq 0$ $\operatorname{im} C \subset \operatorname{im} T$ (or, equivalently, $\ker C \supset \ker T^*$) T^* preserves linear independence of $\operatorname{im} C$ and $\ker C \cap \operatorname{im} T$ as subspaces of $\operatorname{im} T$. Therefore, if we define in $\mathbb{C}^d$ the subspaces

$$X_1 = T^*(\operatorname{im} C), \quad X_2 = T^*(\ker C \cap \operatorname{im} T), \quad X_3 = \ker T,$$

then under the condition $\operatorname{im} C \subset \operatorname{im} T$ we have $\mathbb{C}^d = X_1 \oplus X_2 \oplus X_3$ with $X_1 \oplus X_2 = \operatorname{im} T^* \perp \ker T = X_3$. We call a solution P of (2.1) or (2.2) *C-subordinated* if there is an R such that $P = RCR^*$. A subordinated solution is thus always $\geqq 0$, and for $C = 0$ only $P = 0$ is subordinated.

LEMMA 2.1. *Given the homogeneous equation $TPT^* = 0$, where T is prescribed:*

(i) *There exists a nontrivial (general, Hermitian, nonnegative definite, C-subordinated for $C \neq 0$) solution P if and only if* $\ker T \neq 0$.

(ii) *A (general, Hermitian, nonnegative definite, C-subordinated) matrix P is a solution of $TPT^* = 0$ if and only if*

$$P(\operatorname{im} T^*) \subset \ker T. \tag{2.3}$$

Proof. (i) The condition $\ker T \neq 0$ is obviously necessary. If it holds, then condition (2.3), which is an obvious reformulation of (2.2), can be satisfied for a nontrivial P, e.g., in case $C \neq 0$ by the subordinated

$$P = \begin{cases} 0 & \text{on } \operatorname{im} T^*, \\ R_3 C R_3^* & \text{on } \ker T, \end{cases}$$

where $R_3^*: \ker T \to \operatorname{im} C \neq 0$ is arbitrary while 0 on $\operatorname{im} T^*$. □

LEMMA 2.2. *Given the inhomogeneous equation $TPT^* = C$ with prescribed T and $C \geqq 0$:*

(i) Existence. *There exists a solution $P \geqq 0$ if and only if there exists a subordinated solution if and only if*

$$\ker C \supset \ker T^* \qquad \text{(or, equivalently, } \operatorname{im} C \subset \operatorname{im} T\text{)}. \tag{2.4}$$

(ii) Set of solutions. *An arbitrary solution* $P = QQ^* \geqq 0$ *is given by*

$$Q^* = \begin{cases} U^*\sqrt{C}T_0^{*-1} & \text{on } X_1, \\ 0 & \text{on } X_2, \\ \text{arbitrary} & \text{on } X_3, \end{cases} \tag{2.5}$$

where U *is an arbitrary unitary operator on* $\mathbb{C}^d$. *An arbitrary solution* P *of* (2.1) *also has the form* $P = P_0 + P_h$, *where*

$$P_0 = \begin{cases} T_0^{-1}CT_0^{*-1} & \text{on } X_1, \\ 0 & \text{on } X_2 \oplus X_3 \end{cases} \tag{2.6}$$

is the canonical solution (*derived from* (2.5) *by putting* $Q^* = 0$ *on* X_3) *and* P_h *is a* (*Hermitian*) *solution of the homogeneous equation* (2.2) *such that* $P_0 + P_h \geqq 0$.

(iii) Set of subordinated solutions. *An arbitrary subordinated solution* $P = RCR^*$ *is given by*

$$R^* = \begin{cases} \sqrt{C}_0^{-1}U^*\sqrt{C}T_0^{*-1} \\ \quad + \text{arbitrary into } \ker C & \text{on } X_1, \\ \text{arbitrary into } \ker C & \text{on } X_2, \\ \text{arbitrary} & \text{on } X_3, \end{cases} \tag{2.7}$$

where U *is unitary with invariant subspace* $\operatorname{im} C$.

(iv) Uniqueness. *The solution is unique* (*and equal to the canonical solution* P_0) *if and only if* $\ker T = X_3 = 0$.

Proof. (i) If P satisfies $TPT^* = C$, then for every $Cx \in \operatorname{im} C$ there is a $y = PT^*x$ such that $Ty = Cx$, i.e., $\operatorname{im} C \subset \operatorname{im} T$. Conversely, if (2.4) holds, then $\mathbb{C}^d = \operatorname{im} C \oplus \ker C \cap \operatorname{im} T \oplus \ker T^*$ and $\mathbb{C}^d = T^*(\operatorname{im} C) \oplus T^*(\ker C \cap \operatorname{im} T) \oplus \ker T = X_1 \oplus X_2 \oplus X_3$ because $T_0^* : \operatorname{im} T \to \operatorname{im} T^*$ is bijective. Thus P_0 given by (2.6) is well defined and obviously subordinated. It remains to show that it is a solution. In fact, for each $x = x_1 \oplus x_2 \oplus x_3 \in \operatorname{im} C \oplus \ker C \cap \operatorname{im} T \oplus \ker T^*$, $TP_0T^*x = TT_0^{-1}CT_0^{*-1}T^*x_1 = Cx_1 = Cx$ since $TT_0^{-1} = I = T_0^{*-1}T^*$ on $\operatorname{im} C \subset \operatorname{im} T$ and $x_2 \oplus x_3 \in \ker C \supset \ker T^*$, by (2.4).

(ii) Suppose $P = QQ^*$ solves (2.1). Then $|Q^*T^*x| = |\sqrt{C}x|$ for all $x = x_1 \oplus x_2 \oplus x_3 \in \operatorname{im} C \oplus \ker C \cap \operatorname{im} T \oplus \ker T^* = \mathbb{C}^d$, thus $|Q^*T^*x_2| = 0$ entailing $Q^* = 0$ on X_2 and $|Q^*T^*x_1| = |\sqrt{C}x_1|$ or, equivalently, $|Q^*T^*\sqrt{C}_0^{-1}y| = |y|$ on $\operatorname{im} C$, entailing that $Q^*T^*\sqrt{C}_0^{-1} = U^*$ is isometric on $\operatorname{im} C$ and finally $Q^* = U^*\sqrt{C}T_0^{*-1}$ on X_1, where U can be taken unitary on $\mathbb{C}^d$ without loss of generality. Finally, Q^* is arbitary on X_3. Therefore, a solution has to look like (2.5).

Conversely, if $P = QQ^*$ has form (2.5), then we will prove that $y^*TPT^*x = y^*Cx$ for all x, $y \in \mathbb{C}^d$ yielding (2.1). Start with $x = x_1 \oplus x_2$, $y = y_1 \oplus y_2 \in \operatorname{im} T \oplus \ker T^* = \mathbb{C}^d$, $T^*x_1 = T^*u_1 \oplus T^*u_2$, $y_1^*T = T^*y_1 = v_1^*T \oplus v_2^*T \in X_1 \oplus X_2 = \operatorname{im} T^*$. Then $y^*TPT^*x = y_1^*TPT^*x_1 = (v_1^*T + v_2^*T)QQ^*(T^*u_1 + T^*u_2) = v_1^*TT_0^{-1} \times \sqrt{C}UU^*\sqrt{C}T_0^{*-1}Tu_1 = v_1^*Cu_1 = y_1^*Cx_1 = y^*Cx$. If $Q^* = 0$ on X_3 we obtain P_0 as a particular (subordinated) solution. Clearly $P - P_0 = P_h$ satisfies (2.2) since (2.1) is linear.

(iii) All we have to show is that the R^*'s defined by (2.7) are exactly the solutions of the equation $\sqrt{C}R^* = Q^*$ with Q^* given by (2.5). Of course, every R^* from (2.7) satisfies $\sqrt{C}R^* = Q^*$ with some Q^* from (2.5). To show that every solution of $\sqrt{C}R^* = Q^*$ looks like (2.7) we introduce a second decomposition of $\mathbb{C}^d$ given by $Y_1 = \operatorname{im} C$, $Y_2 = \ker C$ and decompose our operators as follows:

$$\sqrt{C} = \Gamma_{11} + \Gamma_{12} + \Gamma_{21} + \Gamma_{22}, \quad \Gamma_{jk}(Y_k) \subset Y_j, \quad \Gamma_{jk}(Y_{\neq k}) = 0,$$

$$R^* = R_{11}^* + R_{12}^* + R_{13}^* + R_{21}^* + R_{22}^* + R_{23}^*, \qquad R_{jk}^*(X_k) \subset Y_j,$$

$$R_{jk}^*(X_{\neq k}) = 0, \quad Q^* = Q_1^* + Q_2^* + Q_3^*, \quad Q_k^*(X_{\neq k}) = 0.$$

Writing down now the equation $\sqrt{C}R^* = Q^*$ in detail and taking into account that $\Gamma_{12} = \Gamma_{21} = \Gamma_{22} = 0$ and Γ_{11} is invertible on Y_1 with $\Gamma_{11}^{-1} = \sqrt{C_0}^{-1}$ we find that R_{13}, R_{21}, R_{22} and R_{32} are arbitrary, while $R_{12} = 0$ and $\Gamma_{11} R_{11}^* = Q_1^* = U^* \sqrt{C} T_0^{*-1}$ on X_1 into Y_1, from which we conclude that necessarily $U^*(\operatorname{im} C) \subset \operatorname{im} C$ and $R_{11}^* = \sqrt{C_0}^{-1} U^* \sqrt{C} T_0^{*-1}$.

(iv) If $\ker T = 0$ then certainly P_0 is the only solution. If there were two different solutions, then $P_1 - P_2$ would solve the homogeneous equation, whence, by Lemma 2.1(i) $\ker T \neq 0$. □

Remark 2.1. If C, T and P are on $\mathbb{R}^d$ rather than on $\mathbb{C}^d$, the preceding lemmas remain valid with the obvious changes.

Remark 2.2. *Matrix representation* of the solutions of (2.1) and (2.2). As shown by the lemmas, an appropriate basis for representing the solutions as matrices is furnished by $\mathbb{C}^d = X_1 \oplus X_2 \oplus X_3$.

(i) *Homogeneous equation.* A *general* solution P of $TPT^* = 0$ has the matrix representation

$$P = \begin{pmatrix} 0 & 0 & P_{13} \\ 0 & 0 & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix}, \qquad P_{kj} \text{ arbitrary.}$$

P is a *Hermitian* solution if and only if $P_{kj} = P_{jk}^*$, P is a *nonnegative definite* solution if and only if $P_{31} = P_{13} = P_{32} = P_{23} = 0$, $P_{33} = Q_3 Q_3^* \geqq 0$. Finally, P is a *subordinated* solution if and only if it is nonnegative definite and $P_{33} = R_3 C R_3^*$, R_3 arbitrary on $\ker T$ and 0 on $\operatorname{im} T^*$.

(ii) *Inhomogeneous equation.* The *canonical* solution P_0 looks like

$$P_0 = \begin{pmatrix} T_0^{-1} C T_0^{*-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Any *nonnegative definite* solution has the form $P = P_0 + P_h \geqq 0$, where P_h is a Hermitian solution of the homogeneous equation with the following structure: $P_{31} = Q_{31} U^* \sqrt{C} T_0^{*-1} = P_{13}^*$, $P_{32} = 0 = P_{23}$, $P_{33} = Q_{31} Q_{31}^* + Q_{32} Q_{32}^*$ or

$$P = \begin{pmatrix} T_0^{-1} C T_0^{*-1} & 0 & T_0^{-1} \sqrt{C} U Q_{31}^* \\ 0 & 0 & 0 \\ Q_{31} U^* \sqrt{C} T_0^{*-1} & 0 & Q_{31} Q_{31}^* + Q_{32} Q_{32}^* \end{pmatrix} \geqq 0, \tag{2.8}$$

with U unitary and Q_{31}, Q_{32} arbitrary on $\ker T$ and 0 on $\operatorname{im} T^*$. For a *subordinated* $P = RCR^* = P_0 + P_h$ we have for the possible R's

$$R = \begin{matrix} \\ \\ \\ \\ \text{im } C \end{matrix} \begin{matrix} \begin{pmatrix} T_0^{-1} \sqrt{C} U \sqrt{C}^{-1} & R_{12} \\ 0 & R_{22} \\ R_{31} & R_{32} \end{pmatrix} \end{matrix} \begin{matrix} X_1, \\ X_2, \\ X_3, \\ \\ \ker C \end{matrix}$$

with arbitrary R_{12}, R_{22}, R_{31} and R_{32} and U unitary leaving $\operatorname{im} C$ invariant, while the P_h has to satisfy $P_{31} = R_{31} \sqrt{C} U^* \sqrt{C} T_0^{*-1} = P_{13}^*$, $P_{32} = 0 = P_{23}^*$, $P_{33} = R_{31} C R_{31}^*$ or

$$P = \begin{pmatrix} T_0^{-1} C T_0^{*-1} & 0 & T_0^{-1} \sqrt{C} U \sqrt{C} R_{31}^* \\ 0 & 0 & 0 \\ R_{31} \sqrt{C} U^* \sqrt{C} T_0^{*-1} & 0 & R_{31} C R_{31}^* \end{pmatrix}, \tag{2.9}$$

with R_{31} arbitrary on $\ker T$, 0 on $\operatorname{im} T^*$ and U unitary leaving $\operatorname{im} C$ invariant.

Remark 2.3. Equation (2.8) tells us that the $P_h \neq 0$ in the solution $P = P_0 + P_h \geqq 0$ of the inhomogeneous equation is only a nonnegative definite solution of the homogeneous equation if and only if $Q_{31} = 0$. Similarly, by (2.9), the $P_h \neq 0$ in a subordinated $P = P_0 + P_h$ is never subordinated itself.

The preceding lemmas solve the problems concerning soft solutions for a particular critical λ. It remains to piece together the jumps at the various critical λ's and determine what happens in between, so that the final result will be a spectral measure of a stationary process x with values in $\mathbb{R}^d$.

THEOREM 2.1. (Homogeneous equation). a) Existence: *There exists a nontrivial (even subordinate) soft solution, i.e., a spectral measure F_h^x satisfying the homogeneous equation*

$$(i\lambda - A)\, dF_h^x(\lambda)(i\lambda - A)^* = 0 \tag{2.10}$$

if and only if there exists a critical λ.

b) Set of soft solutions: *All solutions F_h^x of* (2.10) *are pure jump measures with possible jumps at critical λ's, where the jump $\Delta F_h^x(\lambda)$ at $\lambda \geqq 0$ is chosen according to Lemma* 2.1(ii) *with $T = i\lambda - A$, $P = \Delta F_h^x(\lambda)$ (for $\lambda = 0$ choose $\Delta F_h^x(0)$ to be real), and $\Delta F_h^x(-\lambda) = \overline{\Delta F_h^x(\lambda)}$. The* subordinated *solutions are those which have subordinated jumps as given by Lemma* 2.1(ii) *with $C = \Delta F^z(\lambda)$.*

Proof. The equation

$$\int_{B \setminus \text{critical } \lambda\text{'s}} (i\lambda - A)\, dF_h^x(\lambda)(i\lambda - A)^* = 0 \quad \text{for each Borel set } B \subset \mathbb{R}$$

is only compatible with $dF_h^x = 0$ off the critical λ's. So if there is no critical λ, there will be no nontrivial solution. Now assume there exists a critical $\lambda \geqq 0$. For this (2.10) reads

$$(i\lambda - A)\Delta F_h^x(\lambda)(i\lambda - A)^* = 0.$$

We can apply Lemma 2.1(i) with $T = i\lambda - A$ (singular) yielding a nontrivial $P = \Delta F_h^x(\lambda)$ (even one that is subordinated to $C = \Delta F^z(\lambda)$ and real if $\lambda = 0$). For $\lambda \leqq 0$ put $\Delta F_h^x(-\lambda) = \overline{\Delta F_h^x(\lambda)}$, which is always possible because P solves $TPT^* = 0$ if and only if $\bar{P}$ solves $\overline{TPT}^* = 0$ and $(i(-\lambda) - A) = (\overline{i\lambda - A})$. □

THEOREM 2.2. (Inhomogeneous equation). a) Existence. *The following statements are equivalent*:

(i) *There exists a soft solution, i.e., a spectral measure F^x satisfying the inhomogeneous equation*

$$(i\lambda - A)\, dF^x(\lambda)(i\lambda - A)^* = dF^z(\lambda). \tag{2.11}$$

(ii) *There exists a soft subordinated solution, i.e.., a spectral measure of the form $dF^x(\lambda) = R(\lambda)\, dF^z(\lambda) R(\lambda)^*$ that satisfies* (2.11).

(iii) $\chi(\lambda)(i\lambda - A)^{-1} \in L_2(F^z)$, *$\chi$ = indicator function of noncritical λ's and*

$$\text{image } \Delta F^z(\lambda) \subset \text{image } (i\lambda - A) \text{ for each critical } \lambda.$$

b) Set of soft solutions. *Any solution of* (2.11) *is given by*

$$F^x = F_0^x + G_h,$$

where the canonical (*subordinated*!) *solution F_0^x has the form*

$$dF_0^x(\lambda) = R_0(\lambda)\, dF^z(\lambda) R_0(\lambda)^*, \qquad R_0(\lambda) = \begin{cases} (i\lambda - A)^{-1}, & \lambda \text{ noncritical}, \\ (i\lambda - A)_0^{-1}, & \lambda \text{ critical}, \end{cases} \tag{2.12}$$

$(i\lambda - A)_0$: im $(i\lambda - A)^* \to$ im $(i\lambda - A)$ *being the (bijective) restriction of* $(i\lambda - A)$ *to* im $(i\lambda - A)^*$. *Each* G_h *is a pure jump function with possible (Hermitian, but not necessarily nonnegative definite) jumps at critical* λ*'s, where* $\Delta G_h(\lambda)$ *solves the homogeneous equation* (2.10) *with* $\Delta F_0^x(\lambda) + \Delta G_h(\lambda) \geqq 0$ *and* $\Delta G_k(-\lambda) = \overline{\Delta G_h(\lambda)}$. *For each fixed critical* λ, *the possible jumps* $\Delta F^x(\lambda)$ *are described by Lemma* 2.2(ii) *for* $T = i\lambda - A$, $P = \Delta F^x(\lambda)$ *and* $C = \Delta F^z(\lambda)$.

c) Set of subordinated solutions. *Any subordinated solution of* (2.11) *is described by*

$$R(\lambda) = \begin{cases} (i\lambda - A)^{-1}, & \lambda \text{ noncritical}, \\ R_\lambda, & \lambda \text{ critical}, \end{cases}$$

where R_λ *for a critical* λ *is given by formula* (2.7) *in Lemma* 2.2(iii) *with* $T = i\lambda - A$, $P = \Delta F^x(\lambda)$ *and* $C = \Delta F^z(\lambda)$.

d) Uniqueness. *The soft solution is unique (and equal to* F_0^x, *thus even subordinated) if and only if there is no critical* λ, *i.e. if and only if all* $\operatorname{Re} \lambda_j(A) \neq 0$. *The subordinated solution is unique (and equal to* F_0^x*) if and only if* $\Delta F^z(\lambda) = 0$ *for each critical* λ.

Proof. a) It suffices to prove (i)$\Rightarrow$(iii)$\Rightarrow$(ii). So assume (i), i.e., the existence of a solution F^x. For critical λ's, F^x satisfies

$$(i\lambda - A)\Delta F^x(\lambda)(i\lambda - A)^* = \Delta F^z(\lambda),$$

which clearly entails im $\Delta F^z(\lambda) \subset$ im $(i\lambda - A)$, while for noncritical λ's

$$\int_{B \setminus \text{critical } \lambda\text{'s}} (i\lambda - A)\, dF^x(\lambda)(i\lambda - A)^* = \int_{B \setminus \text{critical } \lambda\text{'s}} dF^z(\lambda), \qquad B \subset \mathbb{R} \text{ Borel set.}$$

As in the case of a scalar measure we conclude from the last equation that the only possible candidate for F^x for noncritical λ's is

$$dF^x(\lambda) = (i\lambda - A)^{-1}\, dF^z(\lambda)(i\lambda - A)^{*-1}.$$

Since the existence of F^x as a spectral measure was assumed, the right-hand side satisfies

$$\operatorname{tr} \int_{-\infty}^{\infty} \chi(\lambda)(i\lambda - A)^{-1}\, dF^z(\lambda)(i\lambda - A)^{*-1} < \infty,$$

in other words, $\chi(\lambda)(i\lambda - A)^{-1} \in L_2(F^z)$. Thus (iii) is satisfied.

Now assume (iii). Then we claim that F_0^x defined by (2.12) is a solution (which is obviously subordinated). The first condition of (iii) says that

$$\int_{B \setminus \text{critical } \lambda\text{'s}} dF_0^x(\lambda) = \int_{B \setminus \text{critical } \lambda\text{'s}} R_0(\lambda)\, dF^z(\lambda) R_0(\lambda)^*$$

is well defined, while the second condition of (iii) assumes that at critical λ's

$$\Delta F_0^x(\lambda) = R_0(\lambda)\Delta F^z(\lambda) R_0(\lambda)^*$$

is well defined, so F_0^x is a bona fide spectral measure. It obviously satisfies (2.11). Furthermore, because $R_0(-\lambda) = \overline{R_0(\lambda)}$ (and, of course, $dF^z(-\lambda) = dF^z(\lambda)$), $dF_0^x(-\lambda) = \overline{dF_0^x(\lambda)}$. Thus we have found a subordinated solution.

b) and c) The existence proof tells us that in between critical λ's we have no other choice but $dF^x = dF_0^x = (i\lambda - A)^{-1} dF^z (i\lambda - A)^{*-1}$. For critical λ's, the possible jumps are described by Lemma 2.2.

d) follows immediately from b) and c). □

Remark 2.4. *Sufficient condition for noncritical* λ*'s.* The intuitive meaning of the existence condition $\chi(\lambda)(i\lambda - A)^{-1} \in L_2(F^z)$ or

$$\text{tr} \int_{-\infty}^{\infty} \chi(\lambda)(i\lambda - A)^{-1}\, dF^z(\lambda)(i\lambda - A)^{*-1} < \infty \tag{2.13}$$

is that the critical frequencies of the undisturbed system $\dot{x} = Ax$ have to be removed from the noise spectrum. To give the criterion a form that allows more insight assume that A is in complex Jordan canonical form, $A = \text{diag}\,(A_1, \ldots, A_k)$ with $F^z(\lambda)$ partitioned into blocks accordingly. Then (2.13) holds if and only if

$$\text{tr} \int_{-\infty}^{\infty} (i\lambda - A_j)^{-1}\, dF^z_{jj}(\lambda)(i\lambda - A_j)^{*-1} > \infty \tag{2.14}$$

for each critical Jordan block separately (the condition is trivially satisfied for the noncritical blocks). Assume now that $A = i\lambda_0 + H$, where

$$H = \begin{bmatrix} 0 & & & 0 \\ 1 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}$$

is of size d. Then because

$$\text{tr}\,(i\lambda - A)^{-1}\, dF^z(\lambda)(i\lambda - A)^{*-1} = \frac{1}{(\lambda - \lambda_0)^{2d}}\, dF_0(\lambda), \tag{2.15}$$

where

$$dF_0(\lambda) = \sum_{k=1}^{d-1} \sum_{j=1}^{d-1} (i(\lambda - \lambda_0))^{d-k-1}(-i(\lambda - \lambda_0))^{d-j-1}\, \text{tr}\,(H^k dF^z_{kj} H^{j*})$$

is a nonnegative measure, (2.13) is equivalent to

$$\frac{1}{(\lambda - \lambda_0)^d} \in L_2(dF_0). \tag{2.16}$$

If one writes down (2.15) in more detail, one sees that (2.16) is equivalent to

(2.17)

$$\int_{-\infty}^{\infty} \frac{1}{(\lambda - \lambda_0)^{2j}} \left(\sum_{p=1}^{j} \sum_{q=1}^{j} (i(\lambda - \lambda_0))^{p-1}(-i(\lambda - \lambda_0))^{q-1} \right) dF^z_{pq} < \infty \quad \text{for } j = 1, \cdots, d.$$

Note that in general one is not allowed to interchange $\int$ and $\sum$ in (2.17). In the form (2.17) the criterion exhibits most clearly the hierarchy of requirements on the elements of F^z with any new equation added to the system. There is a convenient sufficient condition for (2.17) containing only the diagonal elements of dF^z, namely,

$$\begin{aligned} &\frac{1}{(\lambda - \lambda_0)^d} \in L_2(dF^z_{11}), \cdots, \frac{1}{(\lambda - \lambda_0)^{d-j+1}} \\ &\qquad \in L_2(dF^z_{jj}), \cdots, \qquad \frac{1}{\lambda - \lambda_0} \in L_2(dF^z_{dd}). \end{aligned} \tag{2.18}$$

If (2.18) holds, then all terms appearing in (2.17) exist separately (use the Cauchy–Schwarz inequality), so (2.17) is true. However, (2.18) neglects the possiblity of extinction of the frequency λ_0 due to extreme correlation between components of z (cf. the examples).

If A is arbitrary, condition (2.17) has to be checked for each critical Jordan block separately.

2.2 Hard solutions. We now return to

$$\dot{x} = Ax + z, \qquad x(t) = e^{tA}(x_0 + Z(t)),$$
$$Z(t) = \int_0^t e^{-sA} z(s)\, ds, \qquad t \in \mathbb{R}. \tag{2.19}$$

THEOREM 2.3. a) Existence. *There exists an (even subordinated) hard solution of* (2.19) *on each probability space that can carry z if and only if there exists a soft solution.*

b) Solution set. *Any soft solution can be realized on some probability space. All subordinated soft solutions can be realized on every probability space that can carry z. They are given by*

$$\mathrm{x}(t) = \int_{\mathbb{R}\setminus\text{critical } \lambda\text{'s}} e^{it\lambda}(i\lambda - A)^{-1}\, d\zeta(\lambda) + \sum_{\lambda \text{ critical}} e^{it\lambda} R(\lambda)\Delta\zeta(\lambda),$$

where $R(-\lambda) = \overline{R(\lambda)}$, $R(\lambda)$ *some admissible subordinator at* λ.

c) Uniqueness. *The hard (hard subordinated) solution is unique on* any *probability space that can carry z if and only if the soft (soft subordinated) solution is unique.*

Proof. a) If there is a soft solution then there exists the canonical soft solution F_0^x. This can be realized on any probability space that can carry z via

$$x^0(t) = \int_{-\infty}^{\infty} e^{it\lambda} R_0(\lambda)\, d\zeta(\lambda) \qquad \text{(see (2.12))}.$$

b) Given a soft solution F^x, if it is subordinated, it can be realized wherever z can be realized via $d\xi(\lambda) = R(\lambda)\, d\zeta(\lambda)$. If F^x is not subordinated, we have in general to enlarge the probability space due to the new randomness brought in by jumps at critical λ's. If we are completely free in our choice of the probability space we can realize F^x somewhere as ξ (e.g., as Gaussian process with orthogonal increments and $E\, d\xi\, d\xi^* = dF^x$) and then put $d\zeta = (i\lambda - A)\, d\xi$, thus defining a z subordinated to x.

c) Suppose there is a unique soft solution, i.e., the canonical F_0^x. This happens (by Theorem 2.2d), if and only if there is no critical λ. But then $d\xi^0 = (i\lambda - A)^{-1}\, d\zeta$ is the only solution to

$$\int_B (i\lambda - A)\, d\xi = \int_B d\zeta, \qquad B \text{ Borel set in } \mathbb{R},$$

i.e., the hard canonical solution is unique on any probability space. If the soft subordinated solution is unique, then $\Delta\zeta(\lambda) \equiv 0$ on critical λ's, so again the only solution to

$$\int_B (i\lambda - A) R(\lambda)\, d\zeta = \int_B d\zeta, \qquad B \text{ Borel set in } \mathbb{R},$$

is the canonical one. □

The sum of hard stationary solutions of the homogeneous and the inhomogeneous equation need in general not be stationary. If z is Gaussian then all subordinated solutions are Gaussian.

Example 2.1. $d = 1$, $A = 0$, $\dot{x} = z$, $x(t) = x_0 + \int_0^t z(s)\, ds$, $\lambda^2\, dF^x = dF^z$, $\lambda = 0$ critical. There exists a soft solution if and only if $\Delta F^z(0) = 0$ and $\int_{\mathbb{R}\setminus\{0\}} dF^z(\lambda)/\lambda^2 < \infty$.

In this case the canonical soft solution $dF_0^x(\lambda) = (1/\lambda^2)\, dF^z(\lambda)$ is the unique subordinated solution, while the general solution consists of F_0^x plus a nonnegative jump at $\lambda = 0$. On the hard level, the canonical solution

$$x^0(t) = \int_{-\infty}^{\infty} e^{it\lambda} \frac{1}{i\lambda}\, d\xi(\lambda)$$

is the unique subordinated solution, while all solutions are $x(t) = x^0(t) + c$, c being a r.v. orthogonal to z.

Example 2.2. Existence of a soft solution of

$$\dot{x} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} x + z$$

with $\lambda = 0$ critical is assured if and only if $\Delta F_{11}^z(0) = 0$ and

$$\int_{\mathbb{R}\setminus\{0\}} \lambda^{-2}\, dF_{11}^z < \infty, \qquad \int_{\mathbb{R}\setminus\{0\}} \lambda^{-4}(dF_{11}^z + i\lambda\,(dF_{21}^z - dF_{12}^z) + \lambda^2\, dF_{22}^z) < \infty.$$

Sufficient for the latter is

$$\int_{\mathbb{R}\setminus\{0\}} \lambda^{-4}\, dF_{11}^z < \infty \quad \text{and} \quad \int_{\mathbb{R}\setminus\{0\}} \lambda^{-2}\, dF_{22}^z < \infty.$$

If these conditions hold, $\Delta F^z(0) = \left(\begin{smallmatrix} 0 & 0 \\ 0 & c \end{smallmatrix}\right)$ with $c \geqq 0$. For $\lambda \neq 0$, $dF^x = dF_0^x = (i\lambda - A)^{-1}\, dF^z (i\lambda - A)^{*-1}$, while for $\lambda = 0$ we obtain the following possible jumps:

$$\Delta F^x(0) = \begin{pmatrix} c & \sqrt{c}a \\ \sqrt{c}a & d^2 \end{pmatrix},$$

with two parameters a, $d \in \mathbb{R}$ where $ca^2 \leqq cd^2$. All subordinated soft solutions are described by $R(\lambda) = (i\lambda - A)^{-1}$ for $\lambda \neq 0$, and by

$$R(0) = \begin{pmatrix} r_1 & u \\ r_2 & \rho \end{pmatrix}, \qquad r_1, r_2, \rho \in \mathbb{R}, \quad u = \pm 1$$

yielding subordinated jumps at 0 given by

$$\Delta F^x(0) = R(0)\Delta F^z(0) R(0)^* = c \begin{pmatrix} 1 & u\rho \\ u\rho & \rho^2 \end{pmatrix}, \qquad \rho \in \mathbb{R}, \quad u = \pm 1.$$

All hard subordinated solutions are given by

$$x(t) = \int_{\mathbb{R}\setminus\{0\}} e^{it\lambda} \frac{1}{\lambda^2} \begin{pmatrix} -i\lambda & 0 \\ -1 & -i\lambda \end{pmatrix} d\zeta(\lambda) + \Delta\zeta_2(0) \begin{pmatrix} u \\ \rho \end{pmatrix}.$$

In particular, the (soft or hard) canonical solution, obtained for $\rho = 0$ and $u = +1$, i.e., with

$$F_0^x(0) = \begin{pmatrix} c & 0 \\ 0 & 0 \end{pmatrix}, \qquad \Delta\xi^0(0) = \begin{pmatrix} \Delta\zeta_2(0) \\ 0 \end{pmatrix},$$

is the unique subordinated solution if and only if $c = 0$, i.e., $\Delta F^z(0) = 0$. Any particular soft solution can be realized on the probability space, where z is defined whenever there is a r.v. $\Delta\xi(0)$ with $E\Delta\xi(0)\Delta\xi(0)^* = \Delta F^x(0)$ which is orthogonal to $d\zeta(\lambda)$, $\lambda \neq 0$.

Example 2.3. Existence of a soft solution of

$$\dot{x} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} x + z$$

with $\lambda = \pm 1$ critical is assured if and only if

$$\Delta F^z(1) = c \begin{pmatrix} 1 & -i \\ i & 1 \end{pmatrix} = \overline{\Delta F^z(-1)},\ c \geqq 0,$$

i.e.,

$$\Delta\zeta(1) = \overline{\Delta\zeta(-1)} = \Delta\zeta_1(1) \begin{pmatrix} 1 \\ i \end{pmatrix}$$

and

$$\int_{\mathbb{R}\setminus\{-1,1\}} \frac{1}{(\lambda-1)^2(\lambda+1)^2}((1+\lambda^2)(dF^z_{11}+dF^z_{22})+2i\lambda(dF^z_{21}-dF^z_{12}))<\infty,$$

i.e.

$$\frac{1}{(\lambda-1)(\lambda+1)} \in L_2(F_0), \qquad dF_0 = (1+\lambda^2)(dF^z_{11}+dF^z_{22})+2i\lambda(dF^z_{21}-dF^z_{12}).$$

Sufficient for the latter are the conditions

$$\int_{\mathbb{R}\setminus\{-1,1\}} \frac{dF^z_{11}}{(\lambda-1)^2(\lambda+1)^2}<\infty, \qquad \int_{\mathbb{R}\setminus\{-1,1\}} \frac{dF^z_{22}}{(\lambda-1)^2(\lambda+1)^2}<\infty.$$

Suppose now that the existence conditions are satisfied. Then all possible soft solutions of the inhomogeneous equation are $dF^x(\lambda) = dF^x_0(\lambda) = (i\lambda - A)^{-1}\, dF^z(\lambda)(i\lambda - A)^{*-1}$ for $\lambda \neq \pm 1$, while for $\lambda = \pm 1$ we obtain the following possible jumps:

$$\Delta F^x(1) = \overline{\Delta F^x(-1)} = \begin{pmatrix} d^2 + \sqrt{c}a + \frac{c}{4} & \sqrt{c}b + i\left(d^2 + \frac{c}{4}\right) \\ \sqrt{c}b - i\left(d^2 - \frac{c}{4}\right) & d^2 - \sqrt{c}a + \frac{c}{4} \end{pmatrix},$$

with three parameters a, b, $d \in \mathbb{R}$ such that $c(a^2+b^2) \leqq cd^2$. All subordinated soft solutions are described by $R(\lambda) = (i\lambda - A)^{-1}$ for $\lambda \neq \pm 1$ and by

$$R(1) = \overline{R(-1)} = \frac{1}{4}\begin{pmatrix} s+(r+1)e^{i\psi} & i(s-(r+1)\,e^{i\psi}) \\ -i(t-(r-1)e^{i\psi}) & t-(r-1)\,e^{i\psi} \end{pmatrix},$$

$r, s, t \in \mathbb{C}$, $0 \leqq \psi < 2\pi$, yielding subordinated jumps

$$\Delta F^x(1) = \overline{\Delta F^x(-1)} = \frac{c}{4}\begin{pmatrix} |r+1|^2 & i(r+1)(\bar{r}-1) \\ -i(\bar{r}+1)(r-1) & |\bar{r}-1|^2 \end{pmatrix}, \qquad r \in \mathbb{C}.$$

In particular, for $r = 0$

$$\Delta F^x_0(1) = \frac{c}{4}\begin{pmatrix} 1 & -i \\ i & 1 \end{pmatrix}.$$

All hard subordinated solutions are given by

$$x(t)=\int_{\mathbb{R}\setminus\{\pm 1\}} e^{it\lambda}\frac{1}{1-\lambda^2}(i\lambda+A)\,d\zeta(\lambda)+e^{it}\Delta\xi(1)+e^{-it}\Delta\xi(-1),$$

$$\Delta\xi(1)=\overline{\Delta\xi(-1)}=\frac{1}{2}e^{i\psi}\Delta\zeta_1(1)\begin{pmatrix}1+r\\ i(1-r)\end{pmatrix},\qquad 0\leqq\psi<2\pi,\quad r\in\mathbb{C}.$$

The canonical solution $x^0(t)$ is obtained for $\psi=r=0$. The subordinated solution is unique if and only if $c=0$, i.e., $\Delta F^z(1)=0$, or $\Delta\zeta(1)=0$ almost surely. Any particular nonsubordinated soft solution with jumps $\Delta F^x(\pm 1)$ can be realized on the probability space where z is defined whenever there are two r.v.'s $\Delta\xi(1)$, $\Delta\xi(-1)$ with $\Delta\xi(-1)=\overline{\Delta\xi(1)}$, $E\Delta\xi(1)\,\Delta\xi(1)^*=\Delta F^x(1)$ and $\Delta\xi(1)$, $\Delta\xi(-1)$, $d\zeta(\lambda)$, $\lambda\neq\pm 1$, orthogonal.

We now turn to an existence condition on the hard level. Observe that

$$U_tZ(s)=e^{tA}(Z(t+s)-Z(t)). \tag{2.20}$$

If for $T\to\infty$ or $-\infty$ the mean square limit

$$\text{l.i.m.}\,\frac{1}{T}\int_0^T Z(t)\,dt=Z_0$$

exists, we can take Cesàro limit in (2.20) with respect to s and obtain (putting $x_0=-Z_0\in H^z$) $U_tx_0=e^{tA}(x_0+Z(t))$, i.e., $x^0(t)=U_tx_0$ is a subordinated solution of (2.19).

Conversely, suppose there exists a hard stationary solution $x(t)=\int\exp(i\lambda t)\,d\xi(\lambda)=e^{tA}(x_0+Z(t))$. We have

$$\begin{aligned}\frac{1}{T}\int_0^T Z(t)dt&=-x_0+\frac{1}{T}\int_0^T\int_{-\infty}^{\infty}e^{t(i\lambda-A)}\,d\xi(\lambda)\,dt\\&=-x_0+\int_{-\infty}^{\infty}\left(\frac{1}{T}\int_0^T e^{t(i\lambda-A)}\,dt\right)d\xi(\lambda).\end{aligned}$$

The second term on the right-hand side will surely disappear if $\int_0^T\exp t(i\lambda-A)\,dt/T$ is bounded and tends to zero for $T\to-\infty$ or ∞. By looking at each Jordan block of A separately and confining ourselves to the critical blocks, one sees that this is only the case if $A=0$ or $A=\left(\begin{smallmatrix}0&-b\\ b&0\end{smallmatrix}\right)$, $b\in\mathbb{R}$. We have thus proved

THEOREM 2.4. *Let A be diagonalizable and let all* $\operatorname{Re}\lambda_j(A)=0$. *There exists a hard stationary solution of $\dot{x}=Ax+z$ if and only if*

$$\text{l.i.m.}_{T\to\infty}\,\frac{1}{T}\int_0^T Z(t)\,dt=Z_0 \tag{2.21}$$

exists. In this case $x_0=-Z_0$ is the initial r.v. of the canonical subordinated solution.

Remark 2.5. If z is second order strictly stationary and A is diagonalizable with all $\operatorname{Re}\lambda_j(A)=0$, then (2.21) is also necessary and sufficient for the existence of a second order strictly stationary solution. In fact, (2.21) is certainly necessary. On the other hand, we can do the same thing with (2.20) now reinterpreted in the strict sense (i.e., U_t being the group of shifts on the set of z-measurable functions) ending up with the admissible $x_0=-Z_0\in L_2$, so (2.21) is sufficient. In particular, if z is Gaussian there exists a (Gaussian) stationary solution if and only if (2.21) holds.

Remark 2.6. *Resonance.* If there is no stationary solution, we typically encounter a phenomenon called resonance, meaning that $E|x(t)|^2\to\infty(t\to\infty)$ for any solution.

To investigate this it is enough to restrict ourselves to a critical Jordan block. Take, e.g., the scalar equation $\dot{x} = z$, $x(t) = x_0 + Z(t)$, $Z(t) = \int_0^t z(s)\,ds$; other cases can be treated similarly (see Arnold, Horsthemke and Stucki [2] for $d = 2$).

If $\Delta F^z(0) > 0$ then $E|Z(t)|^2 \sim ct^2$. If $\Delta F^z(0) = 0$ but $\int_{-\infty}^{\infty} \lambda^{-2}\,dF(\lambda) = \infty$, then

$$\sigma^2(t) = E|Z(t)|^2 = \int_0^t \int_0^t C(u-v)\,du\,dv = t\int_{-t}^{t} (1-|s|/t)C(s)\,ds$$

$$= t\int_{-\infty}^{\infty} g_t(\lambda)\,dF^z(\lambda), \qquad g_t(\lambda) = t\left(\sin\frac{\lambda t}{2}\Big/\frac{\lambda t}{2}\right)^2,$$

$C(t)$ being the covariance function of z and $g_t(\lambda)$ a regularization of Dirac's δ-function with $\int_{-\infty}^{\infty} g_t(\lambda)\,d\lambda = 2\pi$. Now assume that F^z has a density f around $\lambda = 0$ with $f(\lambda) = |\lambda|^\alpha + o(|\lambda|^\alpha)(\lambda \to 0)$, $-1 < \alpha \leqq 1$. It is easily seen that for $t \to \infty$

$$\sigma^2(t) \sim \begin{cases} ct^{1-\alpha} & \text{for } -1 < \alpha < 1, \\ c \log t & \text{for } \alpha = 1. \end{cases}$$

3. The case $\dot{x} = Ax + Bz$, $y = Gx$. First observe that all statements of § 2 remain true if z is replaced by Bz. As $Z(t) = \int_0^t \exp(-sA)Bz(s)\,ds$ moves almost surely in the controllable subspace $\mathbb{R}_0 = \sum_{k=1}^{d} A^{k-1}(\operatorname{im} B) \subset \mathbb{R}^d$, the canonical stationary solution subordinated to Bz (if existing), with initial r.v.

$$x_0 = \int_{-\infty}^{\infty} R_0(\lambda)B\,d\zeta \in \mathbb{R}_0,$$

will also stay in $\mathbb{R}_0$ since $x(t) = e^{tA}(x_0 + Z(t))$ and $A\mathbb{R}_0 \subset \mathbb{R}_0$.

Now consider $y = Gx$ and assume first that (G, A) is observable, i.e., $\mathcal{N} = 0$, where

$$\mathcal{N} = \bigcap_{k=1}^{d} \ker(GA^{k-1}) \subset \mathbb{R}^d.$$

In this case we have for any $t \in \mathbb{R}$

$$x(t) = W_\alpha^{-1} \int_{s=-\alpha}^{0} e^{sA'} G' \left(y(s+t) + G\int_{u=s}^{0} e^{uA} Bz(t-u+s)du \right) ds,$$

where $W_\alpha = \int_{-\alpha}^{0} e^{sA'} G' G e^{sA}\,ds > 0$ for all $\alpha > 0$ if and only if $\mathcal{N} = 0$ (see Wonham [7, p. 58]). Keeping in mind that a stationary x is automatically stationarily connected with Bz via $Bz = \dot{x} - Ax$, the last formula immediately yields.

THEOREM 3.1. *On the level of hard solutions there is a one-to-one correspondence between stationary x and stationary (y, Bz), between stationary (x, z) and stationary (y, z) and, in particular, between subordinated x and subordinated y.*

For $\mathcal{N} \neq 0$ we switch to the factor system on $X = \mathbb{R}^d/\mathcal{N}$ (see Wonham [7, p. 59]) which is now observable, so that Theorem 3.1 applies.

Remark 3.1. We do not know whether a stationary y can be produced by a nonstationary x. By Theorem 3.1, such a y is certainly not stationarily connected with Bz. However, this cannot happen if all $\operatorname{Re}\lambda_j(A) \neq 0$ and $(i\lambda - A)^{-1} \in L_2(F^z)$. Indeed, assume that y is stationary and x^0 is the (unique) canonical stationary solution. Then $y(t) - y^0(t) = G\exp(tA)(x_0 - x_0^0)$, whence

$$E|y(t) - y^0(t)|^2 \leqq 2E|y(t)|^2 + 2E|y^0(t)|^2 = c.$$

Since the right-hand side is bounded for all $t \in \mathbb{R}$ we necessarily need $x_0 - x_0^0 \in \mathcal{N}$ to prevent blow-up of the left-hand side. For $\mathcal{N} = 0$ this entails $x_0 = x_0^0$, i.e., the stationary y was produced by the stationary x^0.

REFERENCES

[1] L. ARNOLD, W. HORSTHEMKE AND J. STUCKI, *The influence of external real and white noise on the Lotka–Volterra model*, Biometrical J., 21 (1979), pp. 451–471.

[2] L. ARNOLD AND V. WIHSTUTZ, *Stationary solutions of linear systems with additive and multiplicative noise*, Stochastics, 7 (1982), pp. 133–155.

[3] H. BUNKE, *Gewöhnliche Differentialgleichungen mit zufälligen Parametern*, Akademie-Verlag, Berlin 1972.

[4] I. I. GIKHMAN AND A. V. SKOROKHOD, *The Theory of Stochastic Processes* I, Springer, Berlin-Heidelberg-New York 1974.

[5] J. SNYDERS, *Stationary probability distributions for linear time-invariant systems*, this Journal, 15 (1977), pp. 428–437.

[6] A. WENTZELL, *Theorie zufälliger Prozesse*, Birkhäuser, Basel-Boston-Stuttgart, 1979.

[7] W. M. WONHAM, *Linear Multivariable Control: a Geometric Approach* (2nd ed.), Springer, New York-Heidelberg-Berlin, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0007 $01.25/0

OPTIMAL INPUT/OUTPUT FEEDBACK WITH STRUCTURE CONSTRAINTS*

ROMANO M. DESANTIS†

Abstract. An optimal constrained output feedback problem is studied in the framework of Hilbert resolution space valued random processes. The main result, a necessary condition for optimality, is simultaneously applicable to continuous time, sampled data, finite and infinite dimensional systems. When applied to finite dimensional systems this condition rediscovers in a unified setting most of the classical results already available in the technical literature; when applied to infinite dimensional systems, such as hereditary differential systems, it gives new appropriate extensions of these results.

LIST OF SYMBOLS

H	Hilbert space	T^*	the adjoint of T
P^t, P_t	orthoprojector in H	$\|T\|$	the operator norm of T
x, y, z, w	elements of H	$\operatorname{tr}(T)$	the trace of the operator T
$\|x\|$	the norm of x	$\{e_i\}$	an orthonormal basis of H
$\mathbb{R}$	resolution of the identity in H	$[\Psi(t), H_T(t), \xi(t)]$	state (costate) realization of T
ν	linearly ordered set (usually the real numbers)	(k_T, X_T, g_T)	trajectory state (costate) realization of T
t_0, t_∞	minimum and maximum elements of ν	(F, G, B)	state/costate factorization of T
t	element of ν	$[T]_\alpha$	the α component of T, with $\alpha \in \{M, \mathbf{A}, \mathbf{C}\}$
$[H, P^t]$	Hilbert resolution space with $R = \{P^t, t \in \nu\}$	N	a memoryless feedback compensator
$[T]_M$	the memoryless component of T	$J(N)$	the functional to be minimized in an optimal output feedback problem
$[T]_C$	the strictly causal component of T	A, N, F, B	matrices
F, B, N, L_{1i}, L_{2i}	memoryless operators	A', N', F', B'	transpose matrices
$(H, \Sigma, \mathcal{P})$	a space of Hilbert space valued random processes	L_{2i}, L_{1i}	constraints on the feedback compensator
Σ	family of Borel sets in H	$zI, z^{-1}I$	respectively the forward and backward shift operator in $l_{2\infty}$
$\mathcal{P}$	probability measure on Σ	$L_2(t_0, t_1; H)$	Hilbert space of square integrable functions, f, defined on (t_0, t_1) and such that $f(t) \in H$, $\forall t \in (t_0, t_1)$
ρ, π, ω, η	elements of $(H, \Sigma, \mathcal{P})$	$M_2 \triangleq R^n \times L_2(-h, 0; R^n)$, $\phi \in M_2 \to \phi \triangleq (\phi^0, \phi^1)$, $\phi^0 \in R^n$, $\phi' \in L_2(-h, 0; R^n)$; $\langle \phi, \psi \rangle = \langle \phi^0, \psi^0 \rangle + \langle \phi^1, \psi^1 \rangle$	
$Q(\rho, \mu)$	the cross covariance operator associated with the random processes ρ and μ	W_2^1	the Sobolev space of absolutely continuous functions from $(-h, 0)$ to R^n, with first derivative in $L_2(-h, 0; R^n)$
$[Q(\rho, \mu)]^M$	the cross variance operator associated with $Q(\rho, \mu)$	$W_2 \triangleq R^n \times W_2^1$	
$m(\rho)$	the mean of the random process ρ		
$E[f(\sigma)]$	the expected value of the random variable $f(\rho)$		
T, G, F, B	operators acting on a Hilbert space		

1. Introduction. The present paper is in line with and shares with [Po.1], [Sch.1], [Sa.1], [Tu.1] and [De.1] an interest in the investigation of the role of causality properties (such as strict causality, anticausality, the relation between causality structure and state realization, etc.) and causality related operations (such as the extraction

* Received by the editors May 5, 1981, and in revised form March 23, 1982. This research was supported in part by the Quebec Ministry of Education under grant FCAC-378-6 and in part by the Canadian National Research Council under grant CNRC-A-8244.

† École Polytechnique de Montréal, Département de Génie Electrique, Montréal, Québec, Canada H3C 3A7.

of the additive and/or multiplicative causal part of a noncausal system) in linear optimal control estimation and filtering problems with a quadratic cost functional. Porter [Po.1] opened the way to, and Schumitzki [Sch.1] finalized, the solution of a very general deterministic optimal regulator problem. This solution is solely based on causality and causality related state realization fundamental properties; it is completely independent from special systems characterizations as stationarity, state finite dimensionality, infinite time-horizon, etc.; it encompasses multivariate cases and frequency domain results; it is simultaneously applicable to nonstationary, finite-horizon, and infinite dimensional state space systems. Solutions with similar attributes were subsequently developed in the realm of stochastic filtering [Tu.1], [Sa.1] and yet more recently in the context of a basic problem encompassing many of the classical and nonclassical deterministic and stochastic regulator, estimator and filtering problems [De.1].

In all these developments the system to be designed (estimator, controller and/or filter) is in general simply required to be physically realizable. The novelty of the present study is in that it considers a problem where the system can now be submitted to such constraints as a fixed dynamical structure, a decentralized configuration, a partial exchange of information, etc. The objective is to develop in the context of such a problem a general enough formulation and solution so as to once again achieve a wider range of application of the available results, a better understanding of the role played by causality and related concepts, a better perspective of the relations between a number of various problems, techniques, and results. This is done by adopting the Hilbert resolution space approach which is by now standard in this kind of study; while our notations and concepts are essentially identical to those in [De.1]–[De.3], a brief review will be integrated in the development so as to make the paper self contained and easier to read. Complementary background information can be found in most of the cited references and in particular, in a text book format, in [Po.1], [Sa.1], [Go.1], [Ba.1] and in the forthcoming [Fe.1].

2. Causality and state. Given a linearly ordered set ν with minimum and maximum elements t_0 and t_∞, a family of orthoprojectors, $\mathbb{R}=\{P^t, t\in\nu\}$, is a *resolution of the identity* in the Hilbert space H if:

i) $P^{t_0}H=0$, $P^{t_\infty}H=H$ and $P^kH\supseteq P^lH$ whenever $k>l$;

ii) if $\{P^i\}$ is a sequence of orthoprojectors in $\mathbb{R}$ and there exists an orthoprojector P such that $\{P^ix\}\to Px$ for each $x\in H$, then $P\in\mathbb{R}$.

A Hilbert space H equipped with a resolution of the identity $\mathbb{R}=\{P^t: t\in\nu\}$ is called a *Hilbert resolution space* and is denoted by $[H, P^t]$. Given an orthoprojector $P^t\in\mathbb{R}$, the orthoprojector $I-P^t$ is denoted by P_t. An operator $T:[H,P^t]\to[H,P^t]$ is *causal* if $P^tx=P^ty$ implies $P^tTx=P^tTy$. T is *strictly causal* if it is causal and for any given $\varepsilon>0$ one can find a partition $\{t_0,\xi_1,\cdots,\xi_i,\cdots,\xi_N=t_\infty\}\in\nu$ such that $\sup_i|\Delta T\Delta|<\varepsilon$, where $\Delta_i\triangleq P^{\xi_i}P_{\xi_{i-1}}$. T is *anticausal* (resp. strictly anticausal) if T^*, the adjoint of T, is causal (resp. strictly causal). T is *memoryless* if causal and anticausal at the same time [De.2], [Fe.1].

LEMMA 1. *If T is causal (resp. anticausal), then $TP_t=P_tTP_t$, $P^tT=P^tTP^t$ (resp. $TP^t=P^tTP^t$, $P_tT=P_tTP_t$).*

LEMMA 2. *If T_1 is causal (resp. anticausal) and T_2 is strictly causal (resp. strictly anticausal), then T_1T_2 and T_2T_1 are strictly causal (resp. strictly anticausal).*

A *state realization* (resp. costate realization) of a strictly causal (resp. strictly anticausal) T is a triplet $[\Psi_T(t), H_T(t), \zeta_T(t)]$ such that for each $t\in\nu$ one has [Sch.1], [Fe.1]

$H_T(t)$ is a Hilbert space,

$$\Psi_T(t): P^tH \to H_T(t) \qquad (\text{resp. } P_tH \to H_T(t)),$$

$$\zeta_T(t): H_T(t) \to P_tH \qquad (\text{resp. } H_T(t) \to P^tH),$$

and, for each $u \in H$,

$$\zeta_T(t)\Psi_T(t)u = P_tTP^tu \qquad (\text{resp.} = P^tTP_tu).$$

We will only consider minimal state realizations, that is, realizations with Ψ_T dense in $H_T(t)$ (controllable), and ζ_T $1-1$ (observable). Given a strictly causal (resp. strictly anticausal) T with state realization (resp. costate realization), $[\Psi_T(t), H_T(t), \zeta_T(t)]$, we consider the following space of state (costate) trajectories

$$\tilde{X}_T = \{g: \nu \to U_{t\in\nu}H_T(t) | g(\cdot) = \Psi_T(\cdot)u,\ u \in [H, P^t]\}$$

together with the larger space

$$\tilde{\tilde{X}}_T = \{g: \nu \to U_{t\in\nu}H_T(t) | g(t) \in H_T(t)\} > \tilde{X}_T.$$

This space, $\tilde{\tilde{X}}_T$, comes equipped with the family of projections, $\{P^t\}$, defined by the following property: if $g \in \tilde{\tilde{X}}_T$, then P^tg implies

$$(P^tg)(\tau) = \begin{cases} g(\tau), & \tau \leqq t, \\ 0, & \tau > t. \end{cases}$$

Denoting by X_T the smallest subspace of $\tilde{\tilde{X}}_T$ which contains $\tilde{X}_T$ and is closed under $\{P^t\}$, the pair (G_T, K_T) is called a *state trajectory realization* (costate trajectory realization) if

$$T = K_TG_T,$$

$$G_T: [H, P^t] \to [X_T, P^t] \text{ is strictly causal (anticausal)},$$

$$G_Tu = \Psi_T(\cdot)u,$$

$$K_T: [X_T, P^t] \to [H, P^t] \text{ is memoryless.}$$

In the sequel we will take $[X_T, P^t]$ to be a Hilbert resolution space. A state trajectory realization is *minimal* if it is induced by a minimal state/costate realization. A triplet of operators (F, G, B),

$$B: [H_1, P^t] \to [X_1, P^t], \quad G: [X_1, P^t] \to [X_2, P^t], \quad F: [X_2, P^t] \to [H_2, P^t],$$

with $[X_1, P^t]$ and $[X_2, P^t]$ Hilbert resolution spaces, is said to represent a *state-costate factorization* of T, if $T = FGB$ and the pairs (GB, F) and (G^*F^*, B^*) represent a minimal state and costate trajectory realization of T and T^* respectively.

LEMMA 3. *Let strictly causal T admit a state/costate factorization (F, G, B). If for a memoryless selfadjoint Q one has*

$$T^*QT = T_M + T_{\mathbf{C}} + T_{\mathbf{A}}$$

then there exists a memoryless K such that

$$T_{\mathbf{C}} = KGB, \qquad T_{\mathbf{A}} = B^*G^*K^*.$$

3. Stochastic systems. Let $(H, \Sigma, \mathcal{P})$ denote a probability space of Hilbert space valued random processes, with Hilbert space H, family of Borel sets Σ, and probability measure $\mathcal{P}$ [Ba.1, Chap. 6]. Under appropriate conditions associated with a stochastic

process ρ defined in $(H, \Sigma, \mathcal{P})$ there exists an element $m(\rho) \in H$, (the *mean value* of ρ) and a selfadjoint operator $Q(\rho): H \to H$, (the *covariance* of ρ), such that

$$E(x, \rho) = [x, m(\rho)], \qquad x \in H$$

and

$$E[x, \rho - m(\rho)][\rho - m(\rho), y] = [x, Q(\rho)y] \quad \forall x, y \in H.$$

The symbol "$Ef(\sigma)$" denotes the expected value of the scalar valued random variable $f(\sigma)$ with respect to the probability space underlying ρ. Similarly, with each pair of stochastic processes ρ, π one can associate an operator $Q(\rho, \pi)$, (*cross covariance*), with the property that

$$E(x, \rho - m_\rho)(\pi - m_\pi, y) = (x, Q(\rho\pi)y), \qquad x, y \in H.$$

If $Q(\rho, \pi) = 0$ then ρ and π are *uncorrelated.*

By conveniently defining a measure on the index set ν, a Hilbert resolution space $[H, P^t]$ can be given a causality structure preserving representation $[L_2[\nu; \tilde{H}], \tilde{P}^t]$ where: $L_2[\nu; \tilde{H}]$ denotes the Hilbert space of square integrable functions $\nu \to \tilde{H}$, and $\tilde{P}^t$ is the usual family of truncation operators on $L_2[\nu; \tilde{H}]$, [Ha.1]. With respect to such a representation we will suppose that there exists a memoryless operator $[Q(\rho, \mu)]^M$, the *cross variance* of ρ and μ, such that

$$\langle u_1, [Q(\rho, \mu)]^M u_2 \rangle = E \int \langle u_1(\tau), \rho(\tau) - m_\rho(\tau) \rangle \langle u_2(\tau), \mu(\tau) - m_\mu(\tau) \rangle \, d\tau.$$

If $\rho = \mu$, then $[Q(\rho, \mu)]^M$ will be denoted by $[Q(\rho)]^M$ and referred to as the *variance* of ρ. The trace of an operator T is given by $\operatorname{tr}(T) = \sum_i \langle Te_i, e_i \rangle$, where $\{e_i\}$ is any orthonormal basis in H; T is Hilbert–Schmidt if $\operatorname{tr}(T^*T) < \infty$.

LEMMA 4.

i) $Q(L\rho) = LQ(\rho)L^*$.

ii) $E(\langle \rho, R\rho \rangle)$, *whenever finite, is equal to* $\operatorname{tr} RQ(\rho)$.

iii) $\operatorname{tr} T_1 Q(\rho) T_2 = \operatorname{tr} T_2 T_1 Q(\rho) = \operatorname{tr} Q(\rho) T_2 T_1$.

iv) $\operatorname{tr}[T_1 Q(\rho) + T_1^* Q(\rho)] = 2R_e \operatorname{tr} T_1 Q(\rho)$.

v) *If* $Q(\rho, u)$ *is memoryless then* $[Q(\rho, \mu)]^M = Q(\rho, \mu)$.

vi) *If* T_1 *and* T_2 *are memoryless then* $[T_1 Q(\rho, \mu) T_2]^M = T_1 [Q(\rho, \mu)]^M T_2$.

LEMMA 5. *Let* T *be a Hilbert–Schmidt operator.*

i) *If* T *is strictly causal then* $I + T$ *has a bounded causal inverse.*

ii) *If* T_1 *and* T_2 *are strictly causal then, for any real* $\varepsilon > 0$,

$$\lim_{\varepsilon \to 0} |(I + T_1 + \varepsilon T_2)^{-1} - (I + T_1)^{-1} - \varepsilon (I + T_1)^{-1} T_2 (I + T_1)^{-1}| = O(\varepsilon).$$

iii) *There exist Hilbert–Schmidt* $T_{\mathbf{C}}$, $T_{\mathbf{A}}$ *and* T_M, *respectively strictly causal, strictly anticausal and memoryless such that*

$$T = T_{\mathbf{C}} + T_{\mathbf{A}} + T_M.$$

iv) *For any Hilbert–Schmidt triplet* $T_{\mathbf{C}}$, $T_{\mathbf{A}}$ *and* T_M *one has*

$$[T_M T_{\mathbf{A}}]^M = 0, \qquad [T_{\mathbf{C}}^* T_{\mathbf{A}}]^M = 0.$$

v) *If* T *is strictly causal and has a state/costate factorization* (F, G, B), *then in correspondence to any memoryless and invertible* R, *one can find memoryless* $\bar{N}$ *and* K *such that*

$$R + T^*T = (I + \bar{N}GB)^*(R + B^*KB)(I + \bar{N}GB).$$

4. Optimal output feedback. The following abstract problem statement is inspired from, and can be viewed as a generalization of, a number of more specific formulations which have appeared in recent years in the technical literature (see for example [Le.1], [Ko.1], [Er.1], [We.1]).

Statement P1. Optimal structure constrained output feedback (Fig. 1). Let a strictly causal $T:[H_1, P^t]\to[H_2, P^t]$ and three zero mean valued mutually uncorrelated

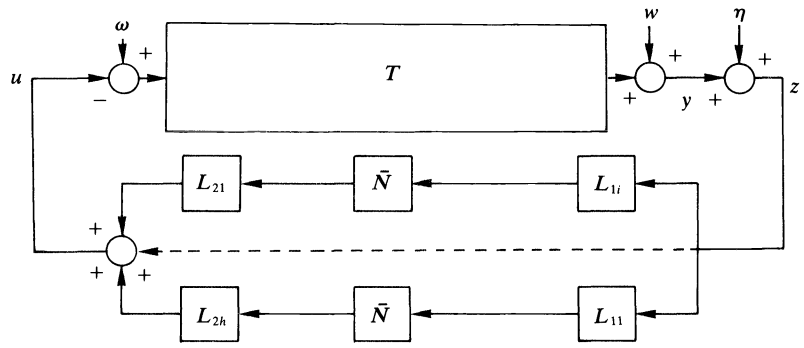


FIG. 1. *Optimal structure constrained output feedback in Hilbert space: determine* $\bar{N}$ *so as to minimize* $J(\bar{N})=\frac{1}{2}E[\langle y, Qy\rangle+\langle u, Ru\rangle]$. $w=P_0TP^0\pi$, *influence of the past input,* $(P^0\pi)$, *over the future output*; y = *output of the plant*; η = *additive noise*; z = *measured noise corrupted output*; u = *control action*; ω = *perturbation*; L_{2i}, L_{1i} = *structure constraints imposed on the feedback controller*; $\bar{N}$ = *the memoryless output feedback compensator to be determined.*

random processes ω, $\pi\in(H_1, \Sigma_1, \mathcal{P}_1)$, $\eta\in(H_2, \Sigma_2, \mathcal{P}_2)$ with $Q(\omega)$ and $Q(\eta)$ memoryless be given. Determine a memoryless $\bar{N}:[H_2, P^t]\to[H_1, P^t]$ such that

i) $I+(\sum_{i=1}^{n} L_{2i}\bar{N}L_{1i})\,T$ has a bounded causal inverse;

ii) for any other memoryless $\tilde{N}$ satisfying i) one has

$$J(N)\leqq J(\tilde{N}),$$

with

$$J(\tilde{N})\triangleq\tfrac{1}{2}E\{\langle y, Qy\rangle+\langle u, Ru\rangle\},\qquad y=P_0T(P_0\omega-P_0u)+w,$$

$$u=\left[\sum_{i=1}^{n} L_{2i}\tilde{N}L_{1i}\right]z,\qquad w=P_0TP^0\pi,$$

$$z=y+\eta,$$

where P^0 is an assigned orthoprojector in $\mathbb{R}$; Q, R, L_{2i}, L_{1i} are assigned memoryless operators with Q and R selfadjoint and positive definite.

Remark 1. The resolution of the identity will be intended to be either discrete or continuous; when the resolution of the identity is continuous, we will take $Q(\eta)=0$ (if $Q(\eta)\neq 0$ one would in general have $J(N)=\infty$ and the problem would no longer be well posed).

Remark 2. The physical interpretation of problem P1 is illustrated in Fig. 1: The random processes ω and η represent respectively the input perturbation and the output measurement noise acting on the plant T; w represents the influence of the past input over the future output; N is the output feedback controller to be chosen so as to minimize a weighted sum of the energies of the output, y, and of the control, u; L_{2i}, L_{1i} allow one to take into account the constraints that one might want to

impose on this controller such as a memoryless input/output, a decentralized configuration, a dynamical controller with an assigned state space dimension. A more detailed discussion about motivation, general historical background and constraint representation can be found in the cited references.

The main results of the paper are embodied in the following.

THEOREM 1. *Under the hypothesis that the required crossvariance operators are well defined one has*:

i) *If T is Hilbert–Schmidt, then a necessary condition for $\bar{N}$ to be a solution of the optimal structure constrained feedback problem is*

$$\sum_{i=1}^{n} L_{2i}^{*}\{[[\bar{T}^{*}\bar{Q}\bar{T}]_{\mathrm{C}}\bar{Q}\bar{T}^{*}]^{M}+[\bar{T}^{*}\tilde{Q}\bar{T}]^{M}NQ(\eta)+RN([\bar{T}\bar{Q}\bar{T}^{*}]^{M}+Q(\eta))\}L_{1i}^{*}=0,$$

where

$$\bar{T}\triangleq T(I+TN)^{-1}, \qquad N\triangleq\sum_{i=1}^{n}L_{2i}\bar{N}L_{1i},$$

$$\tilde{Q}\triangleq Q+N^{*}RN, \qquad \bar{Q}\triangleq P^{0}Q(\pi)P^{0}+P_{0}Q(\omega)P_{0}+NP_{0}Q(\eta)P_{0}N^{*}.$$

ii) *If $\bar{T}$ has a state/costate factorization $(\bar{F},\bar{G},\bar{B})$, then this necessary condition becomes*

$$\sum_{i=1}^{n} L_{2i}^{*}[\bar{B}^{*}K_{0}K_{2}\bar{F}^{*}+(\bar{B}^{*}K_{1}\bar{B}+R)NQ(\eta)+RN\bar{F}K_{2}\bar{F}^{*}]L_{1i}^{*}=0,$$

where K_0, K_1 and K_2 are memoryless and such that

$$K_{0}\bar{G}=[\Pi_{1}]_{\mathrm{C}}, \quad K_{1}=[\Pi_{1}]^{M}, \quad K_{2}=[\Pi_{2}]^{M},$$

$$\Pi_{1}\triangleq\bar{G}^{*}\bar{F}^{*}\tilde{Q}\bar{F}\bar{G}, \qquad \Pi_{2}\triangleq\bar{G}\bar{B}\bar{Q}\bar{B}^{*}\bar{G}^{*}.$$

Proof. i) From the hypothesis that T is Hilbert–Schmidt and strictly causal one has that $\bar{T}=(I+TN)^{-1}T$ is well defined, bounded and strictly causal (Lemma 5). Moreover, observing that

$$y=T(P^{0}\pi+P_{0}\omega+u), \qquad u=-N(y+\eta),$$

it follows that

$$Q(y)=\bar{T}\bar{Q}\bar{T}^{*},$$

where

$$\bar{Q}\triangleq P^{0}Q(\pi)P^{0}+P_{0}Q(\omega)P_{0}+NP_{0}Q(\eta)P_{0}N^{*}$$

and

$$Q(u)=N[Q(y)+Q(\eta)+\bar{T}Q(\eta)+Q(\eta)\bar{T}^{*}]N^{*}.$$

Assuming the required traces to be well defined, one has

$$J(\bar{N})=\tfrac{1}{2}\operatorname{tr}[\tilde{Q}\bar{T}\bar{Q}\bar{T}^{*}+N^{*}RNQ(\eta)+\tfrac{1}{2}\operatorname{tr}N^{*}RN(\bar{T}Q(\eta)+Q(\eta)\bar{T}^{*})],$$

where $\tilde{Q}=Q+N^{*}RN$. Clearly,

$$\tfrac{1}{2}\operatorname{tr}[N^{*}RN\{\bar{T}Q(\eta)+Q(\eta)\bar{T}^{*}\}]$$

is equal to zero: in the discrete case this follows from $\bar{T}$ being strictly causal, in the continuous case, from $Q(\eta)$ being equal to zero (see Remark 1).

Using the continuity of the inverse property (Lemma 5), for any given memoryless $\Delta\bar{N}$ and an arbitrary sufficiently small $\varepsilon > 0$ one has

$$[I + T(N + \varepsilon \Delta N)]^{-1} \cong (I + TN)^{-1} + \varepsilon (I + TN)^{-1} T \Delta N (I + TN)^{-1}$$

where

$$\Delta N = \sum_{i=1}^{n} L_{2i} \Delta \bar{N} L_{1i}.$$

It follows that

$$\Delta J_{\varepsilon}(\bar{N}) \triangleq J(\bar{N} + \varepsilon \Delta \bar{N}) - J(\bar{N}) \cong \varepsilon \text{ tr } \{W \Delta \bar{N}^*\}$$

where

$$W \triangleq \sum_{i=1}^{n} L_{2i}^* \{\bar{T}^* \tilde{Q} \bar{T} \bar{Q} \bar{T}^* + \bar{T}^* \tilde{Q} \bar{T} N Q(\eta) + RN[\bar{T} \bar{Q} \bar{T}^* + Q(\eta)]\} L_{1i}^*.$$

From the arbitrariness of ε, to satisfy the optimality condition one must have $\text{tr}\{W \Delta \bar{N}^*\} = 0$. This, plus the arbitrariness of $\Delta \bar{N}$ implies $[W]^M = 0$, that is,

$$\sum_{i=1}^{n} L_{2i}^* \{[\bar{T}^* \tilde{Q} \bar{T} \bar{Q} \bar{T}^*]^M + [\bar{T}^* \tilde{Q} \bar{T}]^M N Q(\eta) + RN([\bar{T} \bar{Q} \bar{T}^*]^M + Q(\eta))\} L_{1i}^* = 0.$$

Since $\bar{T}^* \tilde{Q} \bar{T}$ is a Hilbert–Schmidt operator, it admits a canonical causality additive decomposition (Lemma 5):

$$\bar{T}^* \tilde{Q} \bar{T} = [\bar{T}^* \tilde{Q} \bar{T}]_M + [\bar{T}^* \tilde{Q} \bar{T}]_{\mathbf{C}} + [\bar{T}^* \tilde{Q} \bar{T}]_{\mathbf{A}}.$$

Using this decomposition one has

$$[\bar{T}^* \tilde{Q} \bar{T} \bar{Q} \bar{T}^*]^M = [[\bar{T}^* \tilde{Q} \bar{T}]_M \bar{Q} \bar{T}^*]^M + [[\bar{T}^* \tilde{Q} \bar{T}]_{\mathbf{A}} \bar{Q} \bar{T}^*]^M + [[\bar{T}^* \tilde{Q} \bar{T}]_{\mathbf{C}} \bar{Q} \bar{T}^*]^M;$$

hence, since the first two elements on the right-hand side of this equation are null,

$$[\bar{T}^* \tilde{Q} \bar{T} \bar{Q} \bar{T}^*]^M = [[\bar{T}^* \tilde{Q} \bar{T}]_{\mathbf{C}} \bar{Q} \bar{T}^*]^M.$$

The optimality condition becomes

$$\sum_{i=1}^{n} L_{2i}^* \{[[\bar{T}^* \tilde{Q} \bar{T}]_{\mathbf{C}} \bar{Q} \bar{T}^*]^M + [\bar{T}^* \tilde{Q} \bar{T}]^M N Q(\eta) + RN([\bar{T} \bar{Q} \bar{T}^*]^M + Q(\eta))\} L_{1i}^* = 0.$$

ii) Let now $(\bar{F}, \bar{G}, \bar{B})$ be a state/costate factorization of $\bar{T}$.
Using the notation

$$\Pi_1 \triangleq \bar{G}^* \bar{F}^* \tilde{Q} \bar{F} \bar{G}, \qquad \Pi_2 \triangleq \bar{G} \bar{B} \bar{Q} \bar{B}^* \bar{G}^*,$$

the necessary conditions become

$$\sum_{i=1}^{n} L_{2i}^* \{[\bar{B}^* [\Pi_1]_{\mathbf{C}} B \bar{Q} \bar{B}^* \bar{G}^* \bar{F}^*]^M + \bar{B}^* [\Pi_1]^M B N Q(\eta) + RN(\bar{F} [\Pi_2]^M F^* + Q(\eta))\} L_{1i}^* = 0.$$

By Lemma 3 there exists a memoryless K_0 such that

$$[\Pi_1]_{\mathbf{C}} = K_0 \bar{G}.$$

Using the notation $K_1 \triangleq [\Pi_1]^M$ and $K_2 \triangleq [\Pi_2]^M$, it follows that

$$\sum_{i=1}^{n} L_{2i}^* \{\bar{B}^* K_0 K_2 \bar{F}^* + (\bar{B}^* K_1 \bar{B} + R) N Q(\eta) + RN \bar{F} K_2 \bar{F}^*\} L_{1i}^* = 0.$$

Theorem 1 unveils an interconnection between causality additive decomposition state realization and constrained feedback optimization which nicely complements the well understood interrelation between causality multiplicative decomposition, unconstrained optimization and Riccati equations [De.1], [Sch.1]. If one poses $Q(\eta)=0$, $\bar{F}=I$, $\sum_{i=1}^{n} L_{2i}\bar{N}L_{1i}=\bar{N}$, then the problem considered in statement P1 becomes the *optimal state feedback* problem considered in these references. In this case the necessary condition for optimality becomes $N=R^{-1}B^*K_0$ where memoryless K_0 must satisfy the equation

$$\bar{G}^*(Q+N^*RN)\bar{G}=K_1+K_0\bar{G}+\bar{G}^*K_0^*.$$

It is easy to verify that one can choose $\bar{G}$ and $\bar{B}$ so that $\bar{G}=(I+GBN)^{-1}P_0G$ and $\bar{B}=B$, where (I, G, B) is a state/costate factorization of T. It follows that K_0 must be such that

$$G^*P_0(I+GB\bar{N})^{-1*}(Q+\bar{N}^*R\bar{N})(I+GB\bar{N})^{-1}P_0G$$
$$=K_1+K_0(I+GB\bar{N})^{-1}P_0G+G^*P_0(I+GB\bar{N})^{-1*}K_0^*.$$

Preoperating and postoperating respectively with $(I+B\bar{N}G)^*$ and $(I+B\bar{N}G)$ and rearranging terms, one has

$$B^*G^*P_0GP_0GB=B^*K_1B+B^*G^*\bar{N}^*(R+B^*K_1B)\bar{N}GB+B^*G^*\bar{N}^*(R+B^*K_1B)$$
$$+(R+B^*K_1B)\bar{N}GB,$$

that is,

$$R+B^*G^*P_0QP_0GB=(I+R^{-1}B^*K_0P_0GB)^*(R+B^*K_{-1}B)(I+R^{-1}B^*K_0P_0GB).$$

One can then conclude that, under the Hilbert–Schmidt hypothesis, the existence of a causality multiplicative decomposition is not only a sufficient condition for optimality, as was recognized in [Sch.1] and [De.1], but it also is a necessary one. The above equation also indicates that the causality decomposition to be considered is in general slightly more complex than suggested in these references where K_1 was systematically supposed to be zero. All this is summarized in the following theorem:

THEOREM 2. *If T is Hilbert–Schmidt and admits a state/costate factorization (F, G, B) then a necessary and sufficient condition for N to be a solution of the optimal state feedback problem is*

$$R+B^*G^*P_0QP_0GB=(I+NP_0GB)^*(R+B^*K_1B)(I+NP_0G).$$

5. Applications. By conveniently specializing the Hilbert spaces and the operators involved, Theorem 1 allows one to rediscover the solutions of most of the various versions of the input/output feedback optimization problems available in the technical literature. As these problems may encompass systems which are not Hilbert–Schmidt, it is of interest to observe that the Hilbert–Schmidt requirement of Theorem 1 is essentially needed to guarantee that

i) $I+TN$ enjoys the invertibility and continuity of the inverse properties described in Lemma 7iii;

ii) Π_1 admits a canonical causality additive decomposition;

iii) $[[\Pi_1]_M\bar{Q}\bar{T}^*]^M=0$ and $[\bar{G}^*K^*\bar{Q}\bar{T}^*]^M=0$.

One can then make Theorem 1 applicable to systems which are not Hilbert–Schmidt by assuming that $(I+TN)$ has a bounded causal inverse and by verifying that properties ii) and iii) do indeed hold in the specific case of interest.

The following examples illustrate the theory.

Example 1. Given real matrices $A(i)$, $B(i)$ and $F(i)$ with dimension $n \times n$, $n \times p$ and $m \times n$ respectively, let the following system be given:

$$x(i+1) = A(i)x(i) + B(i)u(i) + B\omega(i), \qquad x(0) = x_0,$$
$$y(i) = F(i)x(i), \qquad i \in (0, N),$$
$$z(i) = y(i) + \eta(i),$$

where $\omega(i)$ and $\eta(i)$ are uncorrelated zero mean white noise processes with positive definite covariance matrices $Q_\omega(i)$ and $Q_\eta(i)$; x_0 is a random vector independent of $\omega(i)$ and $\eta(i)$ with a zero mean value and covariance matrix $Q(x_0)$. Consider the problem of determining a matrix $N(i)$ of time varying gains such that taking $u(i) = -N(i)z(i)$ one minimizes

$$J[N(\cdot)] = \frac{1}{2} E \sum_{i=0}^{N-1} [\langle x(i+1), Q(i+1)x(i+1)\rangle + \langle u(i), R(i)u(i)\rangle],$$

where $Q(i)$ and $R(i)$ are positive definite matrices.

This problem is brought into the framework of statement P1 by comparing the abstract system represented in Fig. 1 with the more specific block diagram in Fig. 2.

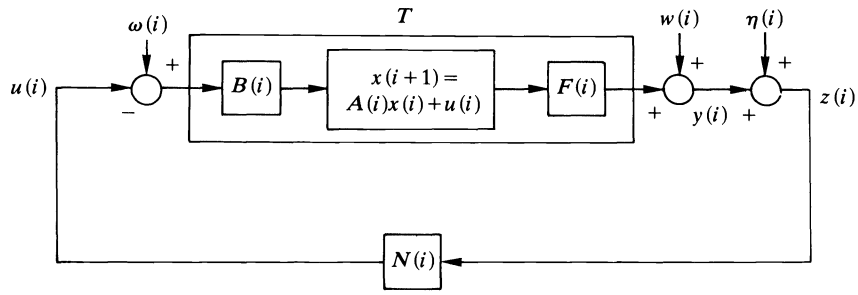


FIG. 2. *Output feedback gains for a linear discrete stochastic control problem*: *determine* $N(i)$, $i \in [0, N-1]$, *so as to minimize the expected value of* $\frac{1}{2}\sum_{i=0}^{N-1} [\langle y(i+1), Q(i+1)y(i+1)\rangle + \langle u(i), R(i)u(i)\rangle]$. $w(i) = \Phi(i, 0)x_0$, *influence of the past input over the future output*; $u(\cdot)$, $x(\cdot)$, $y(\cdot)$ = *control, state and output functions*; $\eta(i)$ = *additive noise*; $\omega(i)$ = *perturbation*; $N(\cdot)$ = *output feedback gain matrix function.*

The strictly causal Hilbert–Schmidt operator T maps in this case $[l_2^p(0, N), P^i] \to [l_2^n(0, N), P^i]$ and is computed in terms of the symbol $F(\cdot)[zI - A(\cdot)]^{-1}B(\cdot)$ where zI represents the forward shift operator; the memoryless operators $Q(\omega)$, $Q(\eta)$, Q and R are computed in terms of the matrices $Q_\omega(i)$, $Q_\eta(i)$, $Q(i)$ and $R(i)$; $w(i) = F(i)\Phi(i, 0)x(0)$, where $\Phi(i, 0)$ is the state transition matrix; $\Sigma L_{2i}\bar{N}L_{1i} = N$. It is easy to verify that all the hypotheses leading to and used in Theorem 1 are satisfied. The necessary conditions for $N(i)$ to be an optimal solution can then be obtained by applying this theorem. Accordingly, one appropriately identifies Π_1 and Π_2 and then computes K_1, K_2 and K_0. In this case one can use the fact that $K_1 = [\Pi_1]^M$ and $K_2 = [\Pi_2]^M$ are given by the memoryless components of Π_1 and Π_2.

Noting that the operators $\bar{G}$ and $\bar{G}^*$ are formally identified with

$$\bar{G} \triangleq [zI - \bar{A}(\cdot)]^{-1}, \qquad \bar{G}^* \triangleq [z^{-1}I - \bar{A}'(\cdot)]^{-1},$$

where $\bar{A}(i) = A(i) - B(i)N(i)F(i)$, one obtains

$$\Pi_1 \triangleq [z^{-1}I - \bar{A}'(\cdot)]^{-1}F'(\cdot)[Q(\cdot) + N'(\cdot)R(\cdot)N(\cdot)]F[zI - \bar{A}(\cdot)]^{-1},$$
$$\Pi_2 \triangleq [zI - \bar{A}(\cdot)]^{-1}[B(Q_\omega(\cdot) + Q_\eta(\cdot))B' + Q(x\,\cdot)\delta(0)][z^{-1}I - \bar{A}'(\cdot)]^{-1}.$$

From here, via simple algebraic manipulations of the type illustrated in [De.1], [Po.1], one finds that the memoryless component of Π_1 is computed by the equation

$$K_1(i) = \bar{A}'(i+1)K_1(i+1)\bar{A}(i+1) + F'(i+1) \cdot [Q(i+1) + N'(i+1)R(i+1)N(i+1)]F(i+1),$$
$$K_1(N-1) = 0$$

while the memoryless part of Π_2 is computed by the equation

$$K_2(i+1) = \bar{A}(i)K_2(i)\bar{A}'(i) + Q_\omega(i) + B(i)N(i)Q_\eta(i)N'(i)B'(i),$$
$$K_2(0) = Q(x_0).$$

Similarly, the operator K_0 such that $K_0 G = (\Pi_1)_C$ turns out to be determined by the relation $K_0(i) = K_1(i)\bar{A}(i)$.

Using the formula given in theorem 1, one has then

$$N(i) = [B'(i)K_1(i)B(i) + R(i)]^{-1}B'(i)K_1(i)\bar{A}(i)K_2(i)F'(i) \cdot [Q_\eta(i) + F(i)K_2(i)F'(i)]^{-1}.$$

In the particular case where $A(\cdot), B(\cdot), F(\cdot), Q(\cdot), R(\cdot), Q_\omega(\cdot)$, and $Q_\eta(\cdot)$ are constant matrices, this result is identical to that given by Ermer and Vandelinde in [Er.1].

Example 2. Let the discrete time invariant system

$$x(i+1) = Ax(i) + Bu(i), \qquad x(0) = x_0,$$
$$y(i) = Fx(i)$$

be given with A, B and F constant matrices with respective dimensions $n \times n$, $n \times p$, and $m \times n$; x_0 is a random vector with a zero mean value and covariance matrix $Q(x_0(0))$. Consider the problem of determining a matrix N of constant output feedback gains such that by taking $u(i) = Ny(i)$ one minimizes the cost functional

$$J(N) = \frac{1}{2} E \sum_{i=0}^{\infty} [\langle y(i+1), Qy(i+1)\rangle + \langle u(i), Ru(i)\rangle]$$

where Q and R are positive definite matrices. This problem is brought into the framework of statement P1 by comparing Fig. 1 with Fig. 3. The operator T to be considered now maps $[l_2^p(0, \infty), P^i] \to [l_2^n(0, \infty), P^i]$ and can be conveniently represented in terms of the transfer function $F(zI - A)^{-1}B$. Note that, though strictly causal, T is not Hilbert–Schmidt. In line with the suggestion in the first paragraph of the

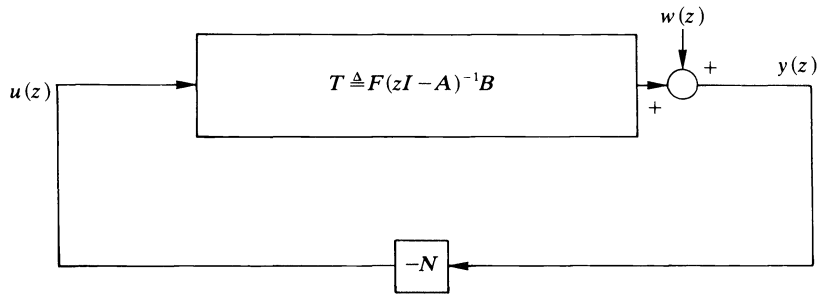


FIG. 3. *Optimal constant output feedback gains problem: determine N so as to minimize the expected value of* $\frac{1}{2}\sum_{i=0}^{\infty}[\langle y(i+1), Qy(i+1)\rangle + \langle u(i), Ru(i)\rangle]$. $w(z) = F(zI - A)^{-1}x_0$, *influence of the past input over the future output*; $u(z), y(z)$ = *control and output functions*; N = *constant output feedback gain matrix.*

present section, we add the hypothesis that $(I+TN)$ has a causal bounded inverse (i.e., the feedback system is stable) and verify by a direct inspection that properties ii) and iii) do indeed hold in the present case. We then apply Theorem 1 just as in the previous example. With a slight abuse of notations one finds

$$\bar{G}=(zI-\bar{A})^{-1}, \quad \bar{G}^*=(z^{-1}I-\bar{A}')^{-1}, \quad \bar{A}\triangleq A-BNF;$$

hence

$$\Pi_1=(z^{-1}I-\bar{A}')^{-1}F'(Q+N'RN)F(zI-\bar{A})^{-1},$$
$$\Pi_2=(zI-\bar{A})^{-1}Q(x_0)(z^{-1}I-\bar{A}')^{-1}.$$

Observing that K_1 and K_2 once again coincide with the memoryless parts of Π_1 and Π_2, one finds that they must satisfy the following Lyapunov equations:

$$\bar{A}'K_1A-K_1=-F'QF-F'N'RNF,$$
$$\bar{A}K_2\bar{A}'-K_2=-Q(x_0).$$

K_0 is easily determined by observing that

$$[\Pi_1]_C=K_1A(zI-\bar{A})^{-1}, \quad \text{hence } K_0=K_1A.$$

One can then conclude: under the hypothesis that N stabilizes the system, a necessary condition for it to be the optimal feedback matrix is

$$(R+B'K_1B)NFK_2F'=B'K_1AK_2F'.$$

This result represents the discrete time version of the by now classical Levine–Athans theorem ([Le.1, Thm. 1]).

Example 3. With $A(t)$, $B(t)$ and $F(t)$ of dimension $n\times n$, $n\times p$ and $m\times n$ respectively, let

$$\dot{x}(t)=A(t)x(t)+B(t)u(t)+B(t)\omega(t), \qquad x(0)=x_0,$$
$$y(t)=F(t)x(t), \qquad t\in[0,t_1],$$

where $w(t)$ is a zero mean white noise random process with a positive definite covariance matrix $Q_\omega(t)$; x_0 is a random vector, independent of $\omega(t)$, with a zero mean value and covariance matrix $Q(x_0)$. The problem of interest is to determine a time varying gains matrix $N(t)$ so as to minimize

$$J[N(\cdot)]=\frac{1}{2}E\left\{\int_0^{t_1}\langle y(t),Q(t)y(t)\rangle+\langle u(t),R(t)u(t)\rangle\,dt+\langle y(t_1),Sy(t_1)\rangle\right\}$$

where $Q(t)$, $R(t)$ and S are positive definite matrices. The picture to compare with Fig. 1 is now represented in Fig. 4. The strictly causal Hilbert–Schmidt $T:[L_2^p[0,t_1],P^t]\to[L_2^m[0,t_2],P^t]$ is formally computed in terms of the symbol $F(\cdot)[sI-A(\cdot)]^{-1}B(\cdot)$, where sI represents the derivative operator; the memoryless operators $Q(\omega)$, Q and R are computed in terms of the matrices $Q_\omega(t)$, $Q(t)$ and $R(t)$; $w(t)=F(t)\Phi(t,0)x_0$, where $\Phi(t,0)$ is the state transition matrix; $\sum_{i=1}^n L_{2i}\bar{N}L_{1i}=\bar{N}$. To obtain the necessary conditions for $N(t)$ to be an optimal controller one establishes the following formal operator representations:

$$G=[sI-\bar{A}(\cdot)]^{-1}, \qquad G^*=[-sI-\bar{A}(\cdot)]^{-1}$$

with

$$\bar{A}(t)=A(t)-B(t)N(t)F(t).$$

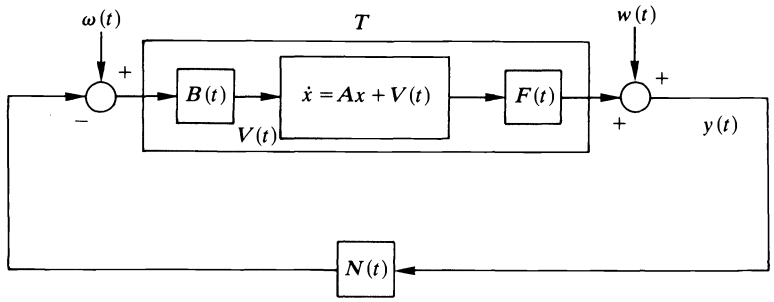


FIG. 4. *Output feedback gains for a linear continuous time stochastic control problem: determine* $N(t)$, $t \in [0, t_1]$, *so as to minimize the expected value of* $\frac{1}{2}\{\int_0^{t_1} [\langle y(t), Q(t)y(t)\rangle + \langle u(t), R(t)u(t)\rangle]\, dt + \langle y(t_1), Sy(t_1)\rangle\}$. $w(t) = \Phi(t, 0)x_0$, *influence of the past input over the future output;* $u(\cdot)$, $x(\cdot)$, $y(\cdot)$ = *control, state and output functions;* $\omega(t)$ = *perturbation;* $N(t)$ = *output feedback gain matrix.*

It follows that

$$\Pi_1 = [-sI - \bar{A}'(\cdot)]^{-1}F'(\cdot)[Q(\cdot) + N'(\cdot)RN(\cdot) + S\delta(t - t_1)]F(\cdot)[sI - \bar{A}(\cdot)]^{-1},$$

$$\Pi_2 = [sI - \bar{A}(\cdot)]^{-1}[B(\cdot)Q_\omega B'(\cdot) + Q(x_0)\delta(t)][-sI - \bar{A}'(\cdot)]^{-1}.$$

The memoryless operator K_0 such that $K_0 G = [\Pi_1]_C$ is computed just as in the previous examples and turns out to be determined by the matrix $K_0(t)$ given by

$$-\dot{K}_0(t) = \bar{A}'(t)K_0(t) + K_0(t)\bar{A}(t) + F'(t)[Q(t) + N'(t)R(t)N(t)]F(t),$$

$$K_0(t_1) = F'(t_1)SF(t_1).$$

K_2 turns out to be determined by

$$\dot{K}_2(t) = \bar{A}(t)K_2(t) + K_2(t)\bar{A}'(t) + BQ_\omega(t)B',$$

$$K_2(0) = Q(x_0).$$

According to Theorem 1, the desired necessary condition for optimality is then

$$R(t)N(t)F(t)K_2(t)F'(t) = -B'(t)K_0(t)K_2(t)F'(t)$$

and, for $F(t)K_2(t)F'(t)$ invertible,

$$N(t) = -R^{-1}(t)B'(t)K_0(t)K_2(t)F'(t)[F(t)K_2(t)F'(t)]^{-1}.$$

This result represents the continuous time version of the Ermer–Vandelinde theorem discussed in Example 1.

Example 4. Consider the hereditary differential system [Ma.1]

$$\dot{x}(t) = A_0(t)x(t) + A_1(t)x(t - h) + B(t)u(t) + B(t)\omega(t),$$

$$y(t) = F(t)x(t), \qquad t \in [t_0, t_1],$$

$$x(t) = \begin{cases} \phi^0, & t = t_0 \\ \phi^1(t - t_0), & t \in [t_0 - h, t_0], \quad \phi^1(\theta) \in L_2[-h, 0; R^n], \end{cases}$$

where: $h > 0$; $A_0(\cdot), A_1(\cdot), B(\cdot)$ and $F(\cdot)$ represent $n \times n$, $n \times n$, $n \times p$, $n \times m$ matrix functions bounded and measurable on $[t_0, t_1]$; $\omega(t)$ is a zero mean white noise random process with a positive definite covariance $Q_\omega(t)$; $[\phi^0, \phi^1(\theta)]$ is an $R^n \times L_2(-h, 0; R^n)$ valued zero mean random process characterized by the covariance operator $Q[\phi^0, \phi^1(\theta)]$. It is once again easy to recognize that the problem of determining a

feedback time varying gain matrix $N(t)$ so as to minimize

$$J[N(\cdot)]=\frac{1}{2}E\left\{\int_{t_0}^{t_1}[\langle y(t), Q(t), y(t)\rangle+\langle u(t), R(t)u(t)\rangle]\,dt\right\},$$

where $u(t)=-N(t)x(t)$ and $Q(t), R(t)$ are positive definite matrices, falls into the framework of statement P1. Observing in particular that the input/output behavior of the system can be described in terms of a strictly causal Hilbert–Schmidt operator $T: L_2(t_0, t_1; R^n)\to L_2(t_0, t_1; R^m)$, from Theorem 1 one has that, as a preliminary step to obtain a meaningful necessary condition for $N(\cdot)$ to be optimal, one has to construct a state/costate factorization of $\bar{T}$.

A state/costate factorization of $\bar{T}$ can be constructed by applying the state space theory developed by Delfour [De.5]. Defining $M_2=R^n\times L_2(-h, 0; R^n)$, with inner product $\langle\phi, \varphi\rangle_{M_2}\triangleq\langle\phi^0, \Psi^0\rangle_{R^n}+\langle\phi^1, \Psi^1\rangle_{L_2}$, and introducing the linear operators $\tilde{B}(t): R^n\to M_2$, $\tilde{B}(t)u\triangleq[B(t)u, 0]$, $u\in R^n$ and $\tilde{F}(t): M_2\to R^n$, $\tilde{F}(t)(\phi^0, \phi^1)\triangleq[F(t)\phi^0, 0]$, $(\phi^0, \phi^1)\in M_2$, one has that the system $\bar{T}$ has the state realization described by

$$\frac{d\tilde{x}(t)}{dt}=\bar{A}(t)\tilde{x}(t)+\tilde{B}(t)u(t), \qquad \tilde{x}(t_0)=(\phi^0, \phi^1),$$

$$y(t)=\tilde{F}(t)\tilde{x}(t)$$

where

$$\tilde{x}(t)\triangleq[x(t), x_t(\cdot)]\in M_2, \quad x_t(\theta)=x(t+\theta)\in L_2(-h, 0; R^n), \quad \theta\in(-h, 0);$$

$\bar{A}(t)$ is an operator $R^n\times W_2^1\triangleq W_2\to M_2$ represented by the matrix of operators

$$\begin{bmatrix}\bar{A}_{00}(t) & \bar{A}_{01}(t)\\ \bar{A}_{10}(t) & \bar{A}_{11}(t)\end{bmatrix}$$

where $\bar{A}_{00}(t): R^n\to R^n$, $\bar{A}_{01}(t): L_2(-h, 0; R^n)\to R^n$, $\bar{A}_{10}(t): R^n\to L_2(-h, 0; R^n)$, $\bar{A}_{11}(t): L_2(-h, 0; R^n)\to L_2(-h, 0; R^n)$.

Note that $\bar{A}(t)$ is uniquely defined by

$$[\bar{A}(t)\tilde{x}(t)]^0=[A_0(t)-B(t)N(t)F(t)]x(t)+A_1(t)x(t-h),$$

$$[\tilde{A}(t)\tilde{x}(t)]^1(\theta)=\frac{d}{d\theta}x_t(\theta), \qquad \theta\in(-h, 0).$$

From this state realization one has the following $(\bar{F}, \bar{G}, \bar{B})$ state/costate factorization of $\bar{T}$. Memoryless $\bar{B}: L_2(t_0, t_1; R^n)\to L_2(t_0, t_1; M_2)$ and $\bar{F}: L_2(t_0, t_1; M_2)\to L_2(t_0, t_1; R^m)$ are defined in the natural way in terms of $\tilde{B}(t)$ and $\tilde{F}(t)$; strictly causal $\bar{G}: L_2(t_0, t_1; M_2)\to L_2(t_0, t_1; M_2)$ is given by the operator $(I(d/dt)-\bar{A}(\cdot))^{-1}$.

To compute K_0 and K_2 one notes that

$$\Pi_1\triangleq\left(-I\frac{d}{dt}-\bar{A}^*(\cdot)\right)^{-1}\bar{F}^*\tilde{\mathbf{Q}}\bar{F}\left(I\frac{d}{dt}-\bar{A}(\cdot)\right)^{-1},$$

$$\Pi_2\triangleq\left(I\frac{d}{dt}-\bar{A}(\cdot)\right)^{-1}\bar{B}\mathbf{Q}\bar{B}^*\left(-I\frac{d}{dt}-\bar{A}(\cdot)\right)^{-1}$$

where

$$\tilde{\mathbf{Q}} \triangleq \begin{bmatrix} Q(\cdot)+N(\cdot)R(\cdot)N(\cdot) & 0 \\ 0 & 0 \end{bmatrix},$$

$$\bar{\mathbf{Q}} \triangleq \begin{bmatrix} Q_\omega(\cdot) & 0 \\ 0 & 0 \end{bmatrix}.$$

The memoryless $K_0, K_0: L_2(t_0, t_1; M_2) \to L_2(t_0, t_1; M_2)$, can be described in terms of a matrix function of operators:

$$K_0(t) \triangleq \begin{bmatrix} K_{00}(t) & K_{01}(t) \\ K_{10}(t) & K_{11}(t) \end{bmatrix}$$

where $K_{00}(t): R^n \to R^n$; $K_{01}(t): L_2(-h, 0; R^n) \to R^n$; $K_{10}(t): R^n \to L_2(-h, 0; R^n)$; $K_{11}(t): L_2(-h, 0; R^n) \to L_2(-h, 0; R^n)$.

The equation specifying K_0 is given by

$$\left(-I\frac{d}{dt}-\bar{A}^*(\cdot)\right)^{-1}\bar{F}^*(\cdot)\tilde{\mathbf{Q}}(\cdot)\bar{F}(\cdot)\left(I\frac{d}{dt}-\bar{A}(\cdot)\right)^{-1}$$

$$= K_0(\cdot)\left(I\frac{d}{dt}-\bar{A}(\cdot)\right)^{-1}+\left(-I\frac{d}{dt}-\bar{A}^*(\cdot)\right)^{-1}K_0^*(\cdot),$$

which implies $K_0^*(t) = K_0(t)$ and

$$\dot{K}_0(t) = -\bar{A}(t)^*K_0(t) - K_0(t)\bar{A}(t) - \tilde{F}^*(t)\tilde{\mathbf{Q}}(t)\tilde{F}(t),$$
$$K_0(t_1) = 0.$$

Similarly $K_2: L_2(t_0, t_1; M_2) \to L_2(t_0, t_1; M_2)$ is given by the following operator differential equation [Be.1, § 5.9]:

$$\dot{K}_2(t) = -\bar{A}(t)K_2(t) - K_2(t)\bar{A}^*(t) + \bar{Q}(t),$$
$$K_2(t_0) = Q[\phi^0, \phi^1(\theta)].$$

As has been shown in [De.6] and [Be.1], operator differential equations of the above type are equivalent to a well understood set of coupled ordinary and partial differential equations. Thus the computation of K_0 and K_2 presents no theoretical difficulty and, once implemented, it allows one to finally obtain the necessary conditions of optimality. The complexity associated with these computations is on the other hand quite formidable. As a final observation note that if $\tilde{F} = I$, then Theorem 2 gives the solution $N(t) = R^{-1}(t)\bar{B}^*(t)K_0(t)$, where $K_0(t)$ is now computed via the following Riccati operator differential equation

$$-\dot{K}_0(t) = \tilde{A}^*(t)K_0(t) + K_0(t)\tilde{A}(t) - K_0(t)\tilde{B}(t)R^{-1}(t)\tilde{B}^*(t)K_0(t) + \tilde{\mathbf{Q}}(t),$$
$$K_0(t_1) = 0$$

where

$$\tilde{A}(t) = [\bar{A}(t)] - [\tilde{B}(t)N(t)\tilde{F}(t)].$$

We rediscover then the well known state feedback solution initially proposed by Mitter and Delfour [De.6].

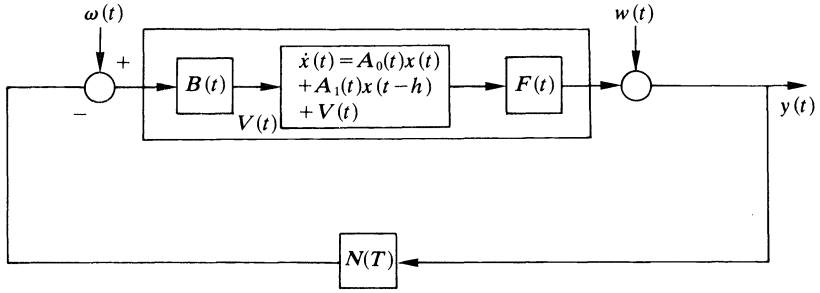


FIG. 5. *Output feedback gains for a linear functional differential stochastic control problem: determine* $N(t)$, $t \in [t_0, t_1]$ *so as to minimize the expected value of* $\frac{1}{2}\{\int_{t_0}^{t_1} [\langle y(t), Q(t)y(t)\rangle + \langle u(t), R(t)u(t)\rangle]\, dt + \langle u(t_1), Sy(t_1)\rangle\}$. $w(t)$ = *influence of the initial state over the future output*; $u(\cdot)$, $y(\cdot)$ = *control and output functions*; $x(t+\theta)$, $\theta \in (-h, 0)$, *state of the system*; $\omega(t)$ = *perturbation*; $N(t)$ = *output feedback gain matrix.*

6. Conclusion. From Theorem 1 one can see that the role played by state and causality theory in constrained linear/quadratic optimization is just as important as that played in the unconstrained optimization case. In particular, computational details aside, the design of an optimal linear/quadratic structure constrained input/output compensator has been shown to require a sequence of well defined and physically meaningful state and causality operations which are completely independent of the specific description of the system. These operations are: the construction of a state/costate factorization, the extraction of strictly causal and memoryless components of a noncausal system, the determination of a state variance operator. In the case of a minimal and discrete resolution of the identity, this last operation corresponds once again to the extraction of a memoryless component; in the case of a continuous resolution of the identity, it coincides with the extraction of a strictly causal component.

From a more practical point of view, the interest of Theorem 1 is in providing a necessary condition which is simultaneously applicable to continuous time, sampled data, finite and infinite dimensional systems. When applied to finite dimensional systems, this condition allows one to rediscover in a unified setting most of the results already available in the technical literature (Examples 1, 2, 3); when applied to infinite dimensional systems, such as hereditary systems, new results are obtained (Example 4); the formidable computational complexity associated with these results opens in turn the way to a new research avenue in the development of efficient computational algorithms. The interest in Theorem 2 is in showing that the existence of a causality multiplicative decomposition, in addition to being a sufficient condition for a solution of the optimal state feedback problem (as established in [Sch.1] and [De.1]), is also a necessary one. By a dual formularization of statement P1 and Theorems 1 and 2, all this could be quickly transferred into the context of constrained filtering and estimation problems. The pattern to follow is clearly indicated in [De.1].

REFERENCES

[Ba.1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Series on Applications of Mathematics, vol. 3, Springer-Verlag, New York, 1976.

[Be.1] A. BENSOUSSAN, *Filtrage optimal des systèmes linéaires*, Dunod, Paris, 1971.

[De.1] R. M. DESANTIS, R. SAEKS AND L. J. TUNG, *Basic optimal estimation and control problems in Hilbert space*, Math. Systems Theory, 12 (1978), pp. 175–203.

[De.2] ———, *Causality in systems analysis*, Proc. IEEE, 64 (1976), pp. 36–44.

[De.3] ———, *Causality for non-linear systems in Hilbert space*, Math. Systems Theory, 7 (1974), pp. 323–337.

[De.4] ———, *Optimal structure constrained output feedback for hereditary systems*, (in preparation).

[De.5] M. C. DELFOUR, *State theory of linear hereditary differential systems*, J. Math. Anal. Appl., 60 (1977), pp. 8–35.

[De.6] ———, *The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution*, Appl. Math. Optim., 3 (1977), pp. 101–162.

[Er.1] C. M. ERMER AND V. D. VANDELINDE, *Output feedback gains for a linear discrete stochastic control problem*, IEEE Trans. Automat. Control, AC-20 (1973), pp. 154–157.

[Fe.1] A. FEINTUCH AND R. SAEKS, *System Theory: A Hilbert Space Approach*, Series on Applications of Mathematics, Springer-Verlag, New York, 1982.

[Go.1] I. Z. GOHBERG AND M. G. KREIN, *Theory and Applications of Volterra Operators in Hilbert Space*, AMS Transl. Math. Monographs 24, American Mathematical Society, Providence, RI, 1970.

[Ha.1] P. R. HALMOS, *Introduction to Hilbert Space and the Theory of the Spectral Multiplicity*, Chelsea, New York, 1957.

[Ko.1] R. L. KOSUT, *Suboptimal Control of Linear Time Invariant Systems Subject to Control Structure Constraints*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 557–563.

[Kw.1] H. KWAKERNAK AND R. SIVAN, *Linear Optimal Control Systems*, Wiley-Interscience, New York, 1972.

[Le.1] W. S. LEVINE, T. L. JOHNSON AND M. ATHANS, *Optimal limited state variable feedback controllers for linear systems*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 44–48.

[Ma.1] A. MANITIUS, *Optimal control of hereditary systems*, in Control Theory and Topics in Functional Analysis, vol. 3, International Atomic Energy Agency, Vienna, 1976, pp. 43–178.

[Po.1] W. A. PORTER, *Modern Foundations of Systems Theory*, Macmillan, New York, 1966.

[Po.2] ———, *A basic optimization problem in linear systems*, Math. Systems Theory, 5 (1971), pp. 20–44.

[Sa.1] R. SAEKS, *Resolution Space Operators and Systems*, Springer-Verlag, New York, 1973.

[Sa.2] ———, *Reproducing kernel resolution space and its application*, J. Franklin Inst., 302 (1976), pp. 331–335.

[Sch.1] A. SCHUMITZKY, *State feedback control for general linear systems*, in Proc. of the International Symposium on the Mathematics of Networks and Systems, T. H. Delft, July 1979, pp. 194–200.

[Tu.1] L. J. TUNG, R. SAEKS AND R. M. DESANTIS, *Wiener–Hopf filtering in Hilbert resolution space*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 702–705.

[We.1] C. L. WENK AND C. H. KNAPP, *Parameter optimization in linear systems with arbitrarily constrained controller structure*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 496–501.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0008 $01.25/0

GLOBAL BEHAVIOR OF GENERALIZED EQUATIONS: A SARD THEOREM*

A. REINOZA†

Abstract. In this paper we are concerned with the global behavior of solutions of generalized equations. Regularity properties of solutions, which are relevant from an algorithmic and theoretical viewpoint, are shown to be generic. This is the fundamental ingredient to establish and prove a Sard theorem for generalized equations.

Key words. nonlinear programming, transversality, subdifferential

1. Introduction. In this paper we shall address questions concerning the global behavior of generalized equations, that is, inclusions of the form

$$0 \in f(x) + \partial \Psi_C(x) \tag{GE}$$

where $f: \mathbb{R}^n \to \mathbb{R}^n$ is a function, C is a polyhedral convex set $\Psi_C(\cdot)$ is the indicator function of C, i.e.,

$$\Psi_C(x) := \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise,} \end{cases}$$

and $\partial \Psi_C(\cdot)$ is the subdifferential operator of the indicator function, i.e.,

$$\partial \Psi_C(x) := \begin{cases} \{v \mid \langle v, z - x \rangle \leqq 0 \text{ for all } z \in C\} & \text{if } x \in C, \\ \varnothing & \text{otherwise.} \end{cases}$$

A generalized equation is a unified way of representing problems from different fields, e.g., mathematical programming, complementarity and mathematical economics, among others. When $C = \mathbb{R}^n_+$, we have the nonlinear complementarity problem first introduced by Cottle [1]. This problem has received a great deal of attention in the last decade. At about the same time that the paper of Cottle appeared, Hartman and Stampacchia [8] proved the following result: If $f: \mathbb{R}^n \to \mathbb{R}^n$ is a continuous function on a convex compact set C in $\mathbb{R}^n$, then there is an x^* in C such that for all $x \in C$

$$\langle x - x^*, f(x^*) \rangle \geqq 0. \tag{VIP}$$

This problem is known in the literature as the variational inequality problem. Karamardian [10], [11] and Moré [15] noticed that the nonlinear complementarity problem (NLCP), is actually a (VIP) if C is a polyhedral convex cone. They used results from variational inequality theory to prove existence of solutions for the (NLCP).

Questions concerning the existence of solutions for problems, in different fields but representable as generalized equations, have been addressed in a somewhat unrelated form; see for example [1], [7], [10], [11], [12], [13], [14], [15], [16]. With the generalized equation setup, these questions are considered in a unified way, thus covering a wide range of applications; see Robinson [19], [20].

The ideas presented here were developed in [17] as a theoretical ground to establish a degree theory for generalized equations. Here we will devote ourselves to

* Received by the editors November 21, 1980, and in revised form March 10, 1982.

† Departamento de Matemáticas y Ciencia de la Computación, Universidad Simón Bolívar, Caracas, Venezuela.

answering some questions about the global behavior of generalized equations. To be more specific, let us introduce the perturbed generalized equation:

$$(\mathrm{GE}_p) \qquad 0 \in f(x,p) + \partial \Psi_C(x)$$

where $f: \mathbb{R}^n \times P \to \mathbb{R}^n$ is a function, P is a topological space, and for some $p_0 \in P$, (GE_{p_0}) is the unperturbed generalized equation.

Globally, we are interested in the "size" of the subset of P for which the generalized equation is well behaved. It will be shown that under suitable conditions three regularity properties, to be defined in § 2, are generically satisfied—that is to say, they are satisfied for all p in a residual set $P_0 \subseteq P$ such that $P \backslash P_0$ is of measure zero in P.

We shall take advantage of the polyhedrality of C. The relative interiors of the faces of C form a partition of C, and the relative interior of each face is a manifold in $\mathbb{R}^n$. We will use some results from differential topology to get more insight into the nature of the solutions of the generalize equation. Thus, questions concerning the behavior of the solutions are reduced to questions about the behavior of the zeros of a vector field on a manifold.

Spingarn [23] has used the parametric transversality theorem, a consequence of Sard's theorem, to show the genericity of the optimality conditions for the general nonlinear programming problem, and we shall use the same approach here for generalized equations. Sard's theorem is the cornerstone of several currently active investigations in optimization; see for example Eaves and Scarf [2], Eaves [3], Saigal and Simon [22].

As a consequence of the global behavior of generalized equations, we shall establish a Sard theorem for regular values of the multivalued function $\Gamma: \mathbb{R}^n \times P \to \mathbb{R}^n$ given by

$$\Gamma(x,p) := f(x,p) + \partial \Psi_C(x).$$

2. Regularity properties. The concept of regularity of the solutions to problems such as nonlinear programming, variational inequalities, complementarity and mathematical economics, has been widely discussed in the literature. Its relevance, from both algorithmic and a theoretic point of view, comes from the necessity of imposing conditions on the problem to ensure good local and global behavior of the solutions. These concepts were extended by Robinson [18] in a natural way to generalized equations, and it is the goal of this section to present them and to derive some of their consequences.

DEFINITION 2.1. Let x^* be a solution of (GE). We say that x^* is a *nondegenerate solution* of (GE) if and only if

$$x^* \in \operatorname{ri} F(x^*),$$

where $F(x^*) = \{x \in C \mid \langle x - x^*, f(x^*) \rangle = 0\}$, and $\operatorname{ri} F(x^*)$ is the relative interior of $F(x^*)$. (It can be easily shown that $F(x^*)$ is a face of C.)

To give an idea of the meaning of the nondegeneracy property, let us consider the complementarity problem, i.e., $C = \mathbb{R}^n_+$.

In this case

$$F(x^*) := \{x \geqq 0 \mid x_i = 0 \text{ if } f_i(x^*) > 0\}$$

and $x^* \in \operatorname{ri} F(x^*)$ is equivalent to strict complementarity slackness.

As was shown in [17], a more suitable way of characterizing nondegeneracy is the following: x^* is a nondegenerate solution of (GE), i.e., $x^* \in \operatorname{ri} F(x^*)$, if and only

if $-f(x^*) \in \operatorname{ri} \partial\Psi_C(x^*)$. From now on we will use interchangeably these two equivalent forms of characterizing nondegeneracy.

It is a well-known result (see [21]) that the relative interiors of the faces of a polyhedral convex set C form a partition of C. Moreover, the relative interior of each face of C is a manifold in $\mathbb{R}^n$. Thus the solutions of (GE) lie on finitely many disjoint manifolds, and when we restrict ourselves to one of these manifolds, questions about the behavior of the solutions of (GE) are reduced to questions about the behavior of the zeros of a vector field on a manifold. Specifically, let x^* be a nondegenerate solution of (GE), and

$$\tau: \operatorname{ri} F(x^*) \to T_{x^*}(\operatorname{ri} F(x^*)),$$

a function defined by $\tau: \Pi_0 \bar{f}$, where $T_{x^*}(\operatorname{ri} F(x^*))$ is the tangent space of $\operatorname{ri} F(x^*)$ at x^*, $\bar{f} := f|\operatorname{ri} F(x^*)$ is the restriction of f to $\operatorname{ri} F(x^*)$ and $\Pi: \mathbb{R}^n \to T_{x^*}(\operatorname{ri} F(x^*))$ is the orthogonal projection. The inclusion $\partial\Psi_C(x^*) \subseteq T_{x^*}(\operatorname{ri} F(x^*))^{\perp}$ implies that x^* is a zero of τ. Thus, the behavior of x^* as a zero of τ gives us information about the behavior of x^* as a solution of (GE). Next, we connect the nondegeneracy property for (GE) and the topological concept of nondegeneracy of the zeros of the vector field τ.

DEFINITION 2.2. Let x^* be a solution of (GE). We say that x^* is a *strongly nondegenerate solution* if:

(i) x^* is a nondegenerate solution of (GE);

(ii) x^* is a nondegenerate zero of τ.

From the definition of nondegenerate zero of a vector field and a simple algebraic argument, we get, in terms of a transversality condition on f, the following characterization of strong nondegeneracy: x^* is a strongly nondegenerate solution of (GE), if and only if x^* is a nondegenerate solution of (GE) and

$$f'(x^*)[T_{x^*}(\operatorname{ri} F(x^*))] + T_{x^*}(\operatorname{ri} F(x^*))^{\perp} = \mathbb{R}^n,$$

that is, when $C = \mathbb{R}^n$, we are dealing with a system of equations, $f(x) = 0$. If x^* is a solution then $F(x^*) = \mathbb{R}^n$, and so (i) is always satisfied and (ii) means that $f'(x^*)$ is nonsingular. Thus, strong nondegeneracy is in this case the usual concept of regular point.

The alternative way of characterizing strong nondegenerate solutions will allow us to establish the genericity of the following important regularity property:

DEFINITION 2.3. Let x^* be a solution of (GE). We say that the *strong positivity conditions* (SPC) hold at x^* for (GE) if, for all $h \neq 0$ such that

(i) $\langle h, f(x^*) \rangle = 0$,

(ii) $x^* + h \in C$,

(iii) $0 \in f(x^*) + f'(x^*)h + \partial\Psi_C(x^*)$,

one has $\langle h, f'(x^*)h \rangle > 0$.

It was shown in [17] that the (SPC) are sufficient for a solution to be isolated. Hence, the genericity of the (SPC) implies that isolation of the solutions of (GE) is a generic property.

The next result is the bridge that allows us to prove genericity of the (SPC).

PROPOSITION 2.1. *Let $f: \mathbb{R}^n \to \mathbb{R}^n$ be a function Fréchet differentiable at a nondegenerate solution x^* of* (GE). *Then the* (SPC) *hold at x^* if and only if x^* is strongly nondegenerate.*

Proof. ($\Rightarrow$) If the implication is false, then from the definition of nondegenerate zero of a vector field, there is some $h \neq 0$ with

$$h \in T_{x^*}(\operatorname{ri} F(x^*)) \tag{1}$$

and such that $\tau'(x^*)h = 0$. Now, $\tau'(x^*) = \Pi_0 f'(x^*)$ so that $\Pi(f'(x^*)h) = 0$, but this holds if and only if

$$-f'(x^*)h \in T_{x^*}(\operatorname{ri} F(x^*))^{\perp}. \tag{2}$$

We know that $T_{x^*}(\operatorname{ri} F(x^*))^{\perp}$ is the affine hull of $\partial\Psi_C(x^*)$, and $-f(x^*) \in \operatorname{ri} \partial\Psi_C(x^*)$, hence for λ small enough

$$0 \in f(x^*) + f'(x^*)\lambda h + \partial\Psi_C(x^*).$$

On the other hand, $x^* + \lambda h \in F(x^*)$ for λ small enough, so that

$$0 = \langle (x^* + h) - x^*, f(x^*)\rangle = \langle \lambda h, f(x^*)\rangle.$$

Thus, λh satisfies conditions (i)–(iii) of the (SPC), but from (1) and (2),

$$\langle h, f'(x^*)h\rangle = 0,$$

which contradicts the assumption that the (SPC) hold at x^*.

($\Leftarrow$) Now, let us assume that x^* is a strongly nondegenerate solution of (GE), so that f satisfies the transversality condition

$$f'(x^*)[T_{x^*}(\operatorname{ri} F(x^*))] + T_{x^*}(\operatorname{ri} F(x^*))^{\perp} = \mathbb{R}^n.$$

For any h such that $x^* + h \in C$ and $\langle f(x^*), h\rangle = 0$, i.e., $h \in T_{x^*}(\operatorname{ri} F(x^*))$, since $-f(x^*) \in \partial\Psi_C(x^*) \subset T_{x^*}(\operatorname{ri} F(x^*))^{\perp}$, we have that

$$0 \in f(x^*) + f'(x^*)h + \partial\Psi_C(x^*)$$

if and only if $f'(x^*)h = 0$, but this holds if and only if $h = 0$.

Hence, the (SPC) hold vacuously. □

We finish this section by illustrating the meaning of strong nondegeneracy for the nonlinear programming problem

$$\min\{\theta(x) | g(x) \leqq 0, h(x) = 0\}, \tag{NLP}$$

where $\theta : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^k$ are C^2 functions. Let us consider the generalized equation

$$0 \in f(w) + \partial\Psi_C(w) \tag{3}$$

where $C = \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^k$, $w = (x, u, v) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k$ and

$$f(w) = \begin{pmatrix} \theta'(x) + ug'(x) + vh'(x) \\ -g(x) \\ -h(x) \end{pmatrix}.$$

Now, $w^* = (x^*, u^*, v^*)$ is strong nondegenerate solution of (3) if and only if:

a) w^* satisfies the Kuhn–Tucker conditions, i.e.,

$$\theta'(x^*) + u^* g'(x^*) + v^* h'(x^*) = 0,$$

$$h(x^*) = 0, \qquad g(x^*) \leqq 0,$$

$$\langle u^*, g(x^*)\rangle = 0, \qquad u^* \geqq 0;$$

b) $w^* \in \operatorname{ri} F(w^*)$. Here, $F(w^*) = \{(x, u, v) | u \geqq 0,\ u_i = 0 \text{ for all } i \in I\}$, where $I = \{i | g_i(x^*) = 0\}$. Thus, $w^* \in \operatorname{ri} F(x^*)$ means that $u_i^* > 0$ if and only if $g_i(x^*) = 0$ for all $i \in I$, i.e., strict complementarity slackness holds at x^*.

c) The (SPC) holds at w^* for (3). In this case, the (SPC) reduces to: for each $h \neq 0$ such that

$$\langle g_i'(x^*), h\rangle = 0, \qquad i \in I,$$
$$\langle h_j'(x^*), h\rangle = 0, \qquad j = 1, \cdots, k,$$

one has $\langle h, L''(w^*)h\rangle > 0$, where

$$L(x, u, v) := \theta(x) + ug(x) + vh(x),$$

i.e., the second order sufficiency conditions hold for (NLP).

d) From the alternative characterization of nondegeneracy we have:

$$f'(w^*)[T_{w^*}(\operatorname{ri} F(w^*))] + T_{w^*}(\operatorname{ri} F(w^*))^{\perp} = \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^k. \tag{4}$$

Here, $T_{w^*}(\operatorname{ri} F(w^*)) = \mathbb{R}^n \times \mathbb{R}^I \times \mathbb{R}^k$ where $\mathbb{R}^I = \{u \in \mathbb{R}^m \mid u_i = 0, \quad i \notin I\}$, $T_{w^*}(\operatorname{ri} F(W^*))^{\perp} = 0 \times (\mathbb{R}^I)^{\perp} \times 0$, and

$$f'(w^*) = \begin{pmatrix} L''(w^*) & g'(x^*) & h'(x^*) \\ -g'(x^*)^T & 0 & 0 \\ -h'(x^*)^T & 0 & 0 \end{pmatrix}.$$

From (4) we get that

$$g'(x^*)^T(\mathbb{R}^I) + (\mathbb{R}^I)^{\perp} = \mathbb{R}^m \quad \text{and} \quad h'(x^*)^T(\mathbb{R}^k) = \mathbb{R}^k,$$

i.e., the gradients of the active constraints are linearly independent.

Conditions (a)–(d) are what Spingarn [23], [24] calls strong second order conditions (SSOC) for the (NLP).

3. Generic properties. With the parametric transversality theorem as a tool, Spingarn [24] proved that the strong second order conditions (SSOC) for the nonlinear programming problem are generically necessary for optimality. With the same approach, we will extend Spingarn's results to prove that isolation and nondegeneracy are generic properties of the solutions of generalized equations. R. Saigal and C. Simon [22], using transversality arguments on the function f, proved for the nonlinear complementarity problem that isolation and nondegeneracy of the solutions are generic properties. Our result is also an extension to generalized equations of Saigal's and Simon's result. Next, we define the family of perturbed problems we are interested in.

DEFINITION 3.1. Let $f: \mathbb{R}^n \times P \to \mathbb{R}^n$ be a function, P a topological space and $p_0 \in P$ some base value at which $f(x) = f(x, p_0)$ for all $x \in \mathbb{R}^n$, i.e., $f(x, p_0)$ is the function in the unperturbed problem. The perturbed generalized equation is defined for each $p \in P$ as:

$$0 \in f(x, p) + \partial \Psi_C(x). \tag{GE_p}$$

The following result will give us the link, between our setup and the parametric transversality theorem, to prove genericity of the properties we are dealing with.

THEOREM 3.1. *Let $X \subseteq \mathbb{R}^n$ and $P \subseteq \mathbb{R}^s$ be C^1 manifolds. Let $f: \mathbb{R}^n \times P \to \mathbb{R}^n$ be a C^1 function such that the function $p \to f(x, p)$ is of rank n at all $(x, p) \in \mathbb{R}^n \times P$. Then there is a residual subset $P_0 \subset P$, such that $P \backslash P_0$ is of measure zero in P and for all $p \in P_0$, the vector field*

$$\tau(\cdot, p): X \to \mathbb{R}^n$$

is zero transversal on X, where

$$\tau(x, p) := \pi_x \circ \bar{f}(x, p),$$

$\pi_x : \mathbb{R}^n \to T_x(X)$ *is the orthogonal projection, and*

$$\bar{f}(\cdot, p) := f(\cdot, p)|X.$$

Proof. Spingarn [23, Prop. 1.23] proved this result for f being the gradient of a function $F : \mathbb{R}^n \times P \to \mathbb{R}$, i.e., $f(x, p) = F'(x, p)$ where F' denotes the derivative with respect to the first argument. However, his proof extends to any function satisfying the requirements of the theorem. □

We can now state the principal result of this section.

THEOREM 3.2. *Let C be a polyhedral convex set. Let $P \subset \mathbb{R}^s$ be a C^1 manifold and $f : \mathbb{R}^n \times P \to \mathbb{R}^n$ a C^1 function such that the function $p \to f(x, p)$ is of rank n at all $(x, p) \in \mathbb{R}^n \times P$. Then there is a residual set $P_0 \subset P$ such that $P \backslash P_0$ is of measure zero in P, and for all $p \in P_0$, if x is a solution of* (GE_p) *then x is strongly nondegenerate (hence also an isolated solution).*

Proof. Spingarn in his dissertation [23, Prop. 4.1], proved that there is a residual set $Q \subset P$, such that $P \backslash Q$ is of measure zero in P and for all $p_0 \in Q$ and $x_0 \in C - f(x_0, p_0) \in \partial \Psi_C(x_0)$ implies that $-f(x_0, p_0) \in \operatorname{ri} \partial \Psi_C(x_0)$, i.e., x_0 is a nondegenerate solution of (GE_{p_0}).

Let C_i, $i = 1, \cdots, M$, be the faces of C. By Theorem 3.1, for each $i = 1, \cdots, M$ there is a residual set P_i such that $P \backslash P_i$ is of measure zero in P, and for all $p \in P_i$, the vector field $\tau_i(\cdot, p) : \operatorname{ri} C_i \to T_x(\operatorname{ri} C_i)$ is zero transversal on $\operatorname{ri} C_i$. This holds for all $x \in \operatorname{ri} C_i$; here $\pi_i(\cdot, p) = \pi_i \circ \bar{f}(\cdot, p)$, $\pi_i : \mathbb{R}^n \to T_x(\operatorname{ri} C_i)$ is the orthogonal projection and $\bar{f}(\cdot, p) := f(\cdot, p)|\operatorname{ri} C_i$.

Let $P_0 = Q \cap (\cap_{i=1}^{M} P_i)$; then P_0 is a residual set such that $P \backslash P_0$ is of measure zero in P, and for each $p \in P_0$ if x is a solution of (GE_p), then x is a nondegenerate solution of (GE_p). Also, since x is a zero of $\tau_i(\cdot, p)$, for some $i \in \{1, \cdots, M\}$ such that $C_i = F(x)$, and $\tau_i(\cdot, p)$ is zero transversal, we have that x is a nondegenerate zero of $\tau_i(\cdot, p)$; that is, for each $p \in P_0$, the solutions of (GE_p) are all strongly nondegenerate. □

As we have mentioned before, now we can clearly see that the strong positivity conditions are a generic property, and hence so is isolation.

4. A Sard theorem. Let C be a polyhedral convex set, P a C^1 manifold in $\mathbb{R}^s$ and $f : \mathbb{R}^n \times P \to \mathbb{R}^n$ a function. In this section, we will define the concept of regular value and establish a Sard theorem for the multivalued function $\Gamma : \mathbb{R}^n \times P \to \mathbb{R}^n$ defined by

$$\Gamma(x, p) = f(x, p) + \partial \Psi_C(x).$$

DEFINITION 4.1. With C, P and f as before, we say that $p \in P$ is a *regular value* of Γ, if for all

$$x \in \Gamma(\cdot, p)^{-1}(0) = \{x | 0 \in f(x, p) + \partial \Psi_C(x)\}$$

x is a strongly nondegenerate solution of (GE_p). Otherwise, we say that p is a *singular value* of Γ.

THEOREM 4.1. *Let C be a polyhedral convex set, $P \subset \mathbb{R}^s$ a C^1 manifold and $f : \mathbb{R}^n \times P \to \mathbb{R}^n$ a C^1 function such that the function $p \to f(x, p)$ is of rank n for all $(x, p) \in \mathbb{R}^n \times P$. Then the set of singular values of Γ is of measure zero in P.*

Proof. This is a direct consequence of Theorem 3.2. □

Next, as an application of our Sard theorem, we establish for three well-known problems that isolation and nondegeneracy are generic properties.

Problem 1. *The nonlinear programming problem.* Let us consider the family of nonlinear programming problems

$$(\mathrm{NLP}_p) \qquad \min \{\theta(x, p) | g(x, p) \leqq 0, h(x, p) = 0\}$$

where $\theta : \mathbb{R}^n \times P \to \mathbb{R}$, $g : \mathbb{R}^n \times P \to \mathbb{R}^m$ and $h : \mathbb{R}^n \times P \to \mathbb{R}^k$ are C^2 functions, and $P \subseteq \mathbb{R}^s$ is a C^2 manifold. Spingarn [23], [24] additionally restricting x to be in a C^2 cyrtohedron (see [23] for a definition), proved that the (SSOC) are generically necessary for optimality. In our case, the cyrtohedron is $\mathbb{R}^n$, the (SSOC) reduce to conditions (a)–(d) of § 2.

COROLLARY 4.1. *Let P, θ, g and h as before, $f : \mathbb{R}^r \times P \to \mathbb{R}^r$ be a function defined by:*

$$f(x, u, v, p) := \begin{pmatrix} L'(x, u, v, p) \\ -g(x, p) \\ -h(x, p) \end{pmatrix}$$

where $r = m + n + k$, and $L(x, u, v, p) := \theta(x, p) + ug(x, p) + vh(x, p)$. If the function $p \to f(x, u, v, p)$ is of rank r at all $(x, u, v, p) \in \mathbb{R}^r \times P$, then there exists a residual set $P_0 \subset P$ such that $P \backslash P_0$ is of measure zero in P, and for each $p \in P_0$, if x^* is a local minimizer of (NLP$_p$), *there exist $(u^*, v^*) \in \mathbb{R}^m_+ \times \mathbb{R}^k$ such that $w^* = (x^*, u^*, v^*)$ satisfies the* (SSOC) *for* (NLP$_p$).

Proof. Let us consider the family of perturbed generalized equations

$$(\mathrm{GE}_p) \qquad 0 \in f(w, p) + \partial \Psi_C(w),$$

where $w = (x, u, v)$ and $C = \mathbb{R}^n \times \mathbb{R}^m_+ \times \mathbb{R}^k$.

From Spingarn [23, Thm. 3.27] and our Sard theorem, there is a residual set $P_0 \subset P$ such that $P \backslash P_0$ is of measure zero in P, and for each $p \in P_0$, if x^* is a local minimizer for (NLP$_p$) there exist $(u^*, v^*) \in \mathbb{R}^m_+ \times \mathbb{R}^k$ such that $w^* = (x^*, u^*, v^*)$ is a strong nondegenerate solution of (GE$_p$). The result follows from the comment at the end of § 2. □

Problem 2. *The nonlinear complementary problem* (NCP). Let $P \subseteq \mathbb{R}^s$ be a C^1 manifold and $f : \mathbb{R}^n \times P \to \mathbb{R}^n$ a C^1 function, and consider the family of perturbed nonlinear complementarity problems

$$(\mathrm{NCP}_p) \qquad x \geqq 0, \quad f(x, p) \geqq 0, \quad \langle x, f(x, p) \rangle = 0.$$

COROLLARY 4.2. *Let P and f as before. If the function $p \to f(x, p)$ is of rank n at each $(x, p) \in \mathbb{R}^n \times P$, then there exists a residual set $P_0 \subset P$, such that $P \backslash P_0$ is of measure zero in P, and for each $p \in P_0$, if x^* is a solution of* (NCP$_p$) *then:*

(i) $x^* + f(x^*, p) > 0$.

(ii) *The principal minor of $f'(x^*, p)$ corresponding to the positive components of x^* is different from zero. Moreover, the solution set of* (NCP$_p$) *is discrete.*

Proof. Let us consider the family of perturbed generalized equations

$$(\mathrm{GE}_p) \qquad 0 \in f(x, p) + \partial \Psi_{\mathbb{R}^n_+}(x).$$

By Theorem 3.1, there is a residual set $P_0 \subset P$ such that $P \backslash P_0$ is of measure zero in P, and for each $p \in P_0$, if x^* is a solution of (GE$_p$), then x^* is strongly nondegenerate, i.e., $x^* \in \operatorname{ri} F(x^*)$ and the (SPC) hold at x^* for (GE$_p$).

Note that x^* is solution of (GE$_p$) if and only if x^* is solution of (NCP$_p$); also, $x^* \in \operatorname{ri} F(x^*)$ if and only if strict complementarity slackness holds, i.e., $x^* + f(x^*) > 0$ holds.

Let A be an $n \times n$ matrix, $x \in \mathbb{R}^n$, $I, J \subseteq N := \{1, \cdots, n\}$. Then A_{IJ} denotes the submatrix of A with elements a_{ij} with $i \in I$, $j \in J$; and x_I denotes the subvector of x with components x_i, $i \in I$. Now, let $I := \{i \mid x_i^* > 0\}$ and $J := N/I$. Any h satisfying

conditions (i)–(iii) of the (SPC), is such that $h_J = 0$ and $f'_{II}(x^*, p)h_I = 0$, which implies that

$$\langle h, f'(x^*, p)h\rangle = \langle h_I, f'_{II}(x^*, p)h_I\rangle = 0.$$

Hence, the determinant of $f'_{II}(x^*, p)$ must be different from zero. Finally, the (SPC) imply isolation, i.e., the solution set of (NCP_p) is discrete. □

Problem 3. Let P be a C^1 manifold, and $f: \mathbb{R}^n \times P \to \mathbb{R}^n$, a C^1 function. If the function $p \to f(x, p)$ is of rank n at each $(x, p) \in \mathbb{R}^n \times P$, then, from Theorem 3.1, for almost all $p \in P$, if $x^* \in f^{-1}(\cdot, p)\,(0)$ then x^* is a strongly nondegenerate solution of the system $f(x, p) = 0$, i.e., $f'(x^*, p)$ is nonsingular. Thus, Theorem 3.1 reduces in this case to Sard's theorem for mappings.

REFERENCES

[1] R. W. COTTLE, *Nonlinear programs with positively bounded Jacobians*, SIAM J. Appl. Math., 14 (1966), pp. 147–158.

[2] B. C. EAVES AND H. SCARF, *The solution of systems of piecewise linear equations*, Math. Oper. Res. 1 (1975), pp. 1–27.

[3] B. C. EAVES, *A short course in solving equations with PL homotopies*, SIAM-AMS Proceedings, 9 (1976), pp. 73–143.

[4] J. W. DANIEL, *Stability of the solution of definite quadratic programs*, Math. Programming, 5 (1973), pp. 41–53.

[5] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley, New York, 1968.

[6] A. V. FIACCO, *Sensitivity analysis for nonlinear programming using penalty methods*, Math. Programming, 10 (1976), pp. 287–311.

[7] G. J. HABETLER AND A. L. PRICE, *Existence theory for generalized nonlinear complementarity problems*, J. Optim. Theory Appl., 7 (1971), pp. 223–239.

[8] P. HARTMAN AND G. STAMPACCHIA, *On some nonlinear elliptic differential functional equations*, Acta Math., 115 (1966), pp. 271–310.

[9] M. W. HIRSCH, *Differential Topology*, Springer-Verlag, New York, 1976.

[10] S. KARAMARDIAN, *The nonlinear complementarity problem with applications, Part* 1, J. Optim. Theory Appl., 4 (1969), pp. 87–98.

[11] ———, *The nonlinear complementarity problem with applications, Part* 2, J. Optim. Theory Appl., 4 (1969), pp. 167–181.

[12] ———, *Generalized complementarity problem*, J. Optim. Theory Appl. 8 (1971), pp. 161–168.

[13] M. KOJIMA, *A unification of the existence theorems of the nonlinear complementarity problem*, Math. Programming, 9 (1975), pp. 257–277.

[14] N. MEGIDDO AND M. KOJIMA, *On the existence and uniqueness of solutions in nonlinear complementarity theory*, Math Programming, 12 (1977), pp. 110–130.

[15] J. J. MORÉ, *Coercivity conditions in nonlinear complementarity problems*, this Journal, 16 (1974), pp. 1–15.

[16] ———, *Classes of functions and feasibility conditions in nonlinear complementarity problems*, Math. Programming, 6 (1974), pp. 327–338.

[17] A. REINOZA, *A degree for generalized equations*, Doctoral dissertation, Univ. of Wisconsin–Madison, Madison, WI, 1979.

[18] S. M. ROBINSON, Unpublished manuscript, 1977.

[19] ———, *An implicit-function theorem for generalized variational inequalities*, Tech. Summ. Rep. 1672, Mathematics Research Center, University of Wisconsin–Madison, 1976.

[20] ———, *Strongly regular generalized equations*, Math. Oper. Res. 5 (1980), pp. 43–62.

[21] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[22] R. SAIGAL AND C. B. SIMON, *Generic properties of the complementarity problem*, Math. Programming, 4 (1973), pp. 324–335.

[23] J. E. SPINGARN, *Generic conditions for optimality in constrained minimization problems*, Doctoral dissertation, Univ. of Washington, Seattle, 1977.

[24] ——— (1980), *On optimality conditions for structured families of nonlinear programming problems*, Math. Programming, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0009 $01.25/0

STABILIZATION OF LINEAR SYSTEMS BY NOISE*

L. ARNOLD†, H. CRAUEL† AND V. WIHSTUTZ†

Abstract. It is proved that the biggest Lyapunov number $\lambda_{\max}$ of the system $\dot{x} = (A + F(t))x$, where A is a fixed $d \times d$ matrix and $F(t)$ is a zero mean strictly stationary matrix-valued stochastic process, satisfies $1/d$ trace $A \leqq \lambda_{\max}$. On the other hand, for each $\varepsilon > 0$ there is a process $F(t)$ for which $\lambda_{\max} \leqq 1/d$ trace $A + \varepsilon$. In particular, the system $\dot{x} = Ax$ can be stabilized by zero mean stationary parameter noise if and only if trace $A < 0$. The stabilization can be accomplished by a one-dimensional noise source. The results carry over to the case where A is a stationary process. They are also true for $F(t)$ = white noise.

Key words. linear stochastic systems, stochastic stability, Lyapunov numbers

1. Introduction. Let $A(t)$, $t \in \mathbb{R}$, be a (strictly) stationary measurable $d \times d$-matrix valued stochastic process on a probability (pr.) space $(\Omega, \mathcal{F}, P)$ with finite mean (thus locally integrable almost surely). This includes the case of a constant matrix A. All notions and facts from the theory of stationary processes used in this paper, such as the associated group of shifts T_t, $t \in \mathbb{R}$, invariant sets and variables, ergodicity etc., can be found in Rozanov [13, Chap. IV]. The linear system

$$\dot{x} = A(t)x, \qquad x(0) = x_0, \tag{1.1}$$

assigns to any initial random variable (r.v.) x_0 the solution $x(t; x_0) = \Phi(t)x_0$, where $\Phi(t)$ is the fundamental matrix with $\Phi(0) =$ identity. We call the system (1.1) stable, if its trivial solution $x \equiv 0$ is exponentially stable, in other words if for any initial random variable x_0 the Lyapunov number

$$\lambda(x_0(\omega), \omega) = \overline{\lim_{t \to \infty}} \frac{1}{t} \log |\Phi(t, \omega)x_0(\omega)|$$

of the corresponding solution is negative almost surely.

The aim of this paper is to find necessary and sufficient conditions for a system (1.1) to be stabilizable by parameter noise, but without interfering in a deterministic way. The idea behind our approach is to model how nature would stabilize a system $\dot{x} = Ax$ with constant A, say, by perturbing the parameters, i.e., the entries of A, but keeping the average fixed and equal to the original value. More precisely, we look for a stationary process $F(t)$ with $EF(t) = 0$ such that

$$\dot{x} = (A(t) + F(t))x \tag{1.2}$$

is stable. The mean zero assumption excludes 'dishonest' things like pole displacement for a control system $\dot{x} = Ax + u$ via feedback control $u = Fx$, a method one typically would apply to a system which is amenable to human interference. The interesting problem of stabilization by stochastic feedback based on observation ($\dot{x} = Ax + Bu$, $y = Cx$, $u = F(t)y$) will be treated elsewhere.

For $d = 1$, the ergodic theorem for stationary (A, F) implies

$$x(t) = x_0 \exp t\frac{1}{t}\int_0^t (A(\tau) + F(\tau))\, d\tau$$

$$\sim x_0 \exp tE(A(0) + F(0)|\mathcal{I}),$$

* Received by the editors January 15, 1982, and in revised form June 2, 1982. This paper was written while the authors were visiting the Courant Institute of Mathematical Sciences, New York University.

† Forschungsschwerpunkt Dynamische Systeme, Universität Bremen, Bibliothekstraße, Postfach 330 440, 2800 Bremen 33, West Germany.

$\mathcal{I}$ denoting the sigma algebra of invariant sets of (A, F). Since $E(A(0)+F(0)|\mathcal{I})<0$ and $EF(0)=0$ entail $EA(0)<0$, a one-dimensional system (1.2) can never be stable unless (1.1) was already stable. Thus we can restrict ourselves to the case $d \geqq 2$.

For $d=2$ Khasminskii [7] stabilized a particular system by applying two white noise sources, while Arnold [1] showed (somewhat heuristically) that any deterministic time invariant system $\dot{x}(t)=Ax(t)$ with trace $A<0$ can be stabilized by one real noise source. The main result of this paper is just a generalization and a rigorous proof of the result in [1]. For general d and $A(t)=A$ constant Willems and Aeyels [19] state algebraic conditions on F leaving the stability properties of (1.1) unchanged.

Our approach is based on the following fundamental theorem of Oseledec [12] (see Wihstutz [18] and Crauel [3] for more recent accounts):

THEOREM 1.1 (Multiplicative ergodic theorem). *Let $A(t)$ satisfy the conditions mentioned above. Then we have for the system* (1.1):

(i) *State space decomposition according to growth*: *For P-almost all $\omega \in \Omega$ there are numbers $r=r(\omega), 1 \leqq r \leqq d$, $\lambda_1(\omega)<\lambda_2(\omega)<\cdots<\lambda_r(\omega)$ and linear subspaces ('Oseledec spaces') $E_1(\omega), \cdots, E_r(\omega)$ with dimension $d_i(\omega)=\dim E_i(\omega)$ such that*

$$R^d = \bigoplus_{i=1}^{r} E_i(\omega)$$

and

$$\lim_{t\to\pm\infty} \frac{1}{t} \log|\Phi(t,\omega)x_0(\omega)| = \lambda_i(\omega) \quad \textit{iff } x_0(\omega) \in E_i(\omega).$$

(ii) *Domain of attraction of $E_i(\omega)$*:

$$\lim_{t\to\infty} \frac{1}{t} \log|\Phi(t,\omega)x_0(\omega)| = \lambda_i(\omega) \quad \textit{iff } x_0(\omega) \in V_i(\omega)\backslash V_{i-1}(\omega),$$

where $V_i(\omega)=\bigoplus_{j=1}^{i} E_j(\omega)$.

(*Note that all Lyapunov numbers are limits rather than* lim sup's).

(iii) *Center of gravity of Lyapunov numbers*:

$$\sum_{i=1}^{r(\omega)} d_i(\omega)\lambda_i(\omega) = E(\operatorname{trace} A(0)|\mathcal{I}) = \operatorname{trace} E(A(0)|\mathcal{I}),$$

$\mathcal{I}$ being the sigma algebra of invariant sets of $A(t)$.

(iv) *Invariance properties*: *The r.v.'s r, λ_i and d_i, $i=1, \cdots, r$ are shift invariant (thus constant if A is ergodic), while $\Phi(t,\omega)E_i(\omega)=T_tE_i(\omega)$, T_t being the group of shifts associated with the stationary process A.*

COROLLARY 1.1. *For the biggest Lyapunov number λ_r of* (1.1) *we always have*

$$\bar{\lambda} = \frac{1}{d} \operatorname{trace} E(A(0)|\mathcal{I}) \leqq \lambda_r \quad a.s.$$

In fact, $\lambda_r<\bar{\lambda}$ would imply $\sum_{i=1}^{r} d_i\lambda_i < \operatorname{trace} E(A(0)|\mathcal{I})$ contradicting Theorem 1.1 (iii).

Note that for a deterministic A the Lyapunov numbers of (1.1) are the real parts of the eigenvalues of A, while the Oseledec spaces are its generalized eigenspaces. Of course, (1.1) is exponentially stable if and only if $\lambda_r<0$ a.s.

The projection $s(t)=x(t)/|x(t)|$ of a nontrivial solution $x(t)$ of (1.1) onto the unit sphere S^{d-1} satisfies the nonlinear equation

$$\dot{s} = h(A,s), \quad h(A,s)=(A-q(A,s)I)s, \quad s(0)=x_0/|x_0|,$$

with

$$q(A, s) = s' \frac{A' + A}{2} s$$

(prime denoting the transpose), while

$$|x(t)| = |x_0| \exp t \frac{1}{t} \int_0^t q(A(\tau), s(\tau))\, d\tau.$$

Thus the Lyapunov number of $x(t; x_0)$ has the form

$$\lambda(x_0, \omega) = \lim_{t \to \infty} \frac{1}{t} \int_0^t q(A(\tau), s(\tau))\, d\tau,$$

which by the ergodic theorem is equal to

$$E_\mu q(A(0), s(0)) = \iint_{\mathbb{R}^{d \times d} \times S^{d-1}} q(A, s)\, d\mu(A, s)$$

provided (A, s) is stationary and ergodic with one-dimensional distribution μ on $\mathbb{R}^{d \times d} \times S^{d-1}$.

If $A(t)$ is a Markov process, we restrict ourselves to so-called Markov solutions of (1.1), i.e., those solutions x such that the pair (A, x) is a Markov process. This amounts to restricting the admissible initial values (see Arnold and Kliemann [2]). For example, (A, x) will certainly be Markov if A is a diffusion and x_0 depends only on the past history of $A(t)$ for $t \leqq 0$. In general, Markov solutions will not be able to attain all Lyapunov numbers existing by Theorem 1.1. On the contrary, typically only the biggest Lyapunov number λ_r can be realized.

2. The main result. Let us first assume that A is deterministic. The general case will be reduced to this particular one later in Remark 2.2.

Our main result is as follows.

THEOREM 2.1. *Given the system $\dot{x} = Ax$, A a fixed $d \times d$ matrix, and the system $\dot{x} = (A + F(t))x$ parametrically perturbed by a stationary ergodic measurable $d \times d$ matrix valued stochastic process with finite mean:*

(i) *For any choice of $F(t)$ with trace $EF(t) = 0$ we always have for the biggest Lyapunov number $\lambda_{\max}(A + F)$ of the perturbed system*

$$\frac{1}{d} \operatorname{trace} A \leqq \lambda_{\max}(A + F).$$

(ii) *For any $\varepsilon > 0$ there is an $F(t)$ with $EF(t) = 0$ such that*

$$\frac{1}{d} \operatorname{trace} A \leqq \lambda_{\max}(A + F) \leqq \frac{1}{d} \operatorname{trace} A + \varepsilon.$$

In particular, a linear system $\dot{x} = Ax$ can be stabilized by zero mean parameter noise if and only if trace $A < 0$.

Proof. (i) By Corollary 1.1 and the ergodicity of $F(t)$, trace $A =$ trace $E(A + F(t)) \leqq d\lambda_{\max}(A + F)$.

(ii) The idea of the proof is as follows: Note first that rotational invariance of the Lebesgue measure λ on S^{d-1} implies

$$\int_{S^{d-1}} q(A, s)\, d\lambda(s) = \frac{1}{d} \operatorname{trace} A, \qquad q(A, s) = \frac{1}{2} s'(A' + A)s.$$

We now pick for each $\varepsilon > 0$ a noise process $F_\varepsilon(t)$ with state space Y in the set of skew-symmetric matrices (whence $q(A+F_\varepsilon, s) = q(A, s)$), such that $(F_\varepsilon, s_\varepsilon)$ has invariant probability μ_ε on $Y \times S^{d-1}$ and

$$\lambda_{\max}(A+F_\varepsilon) = \iint_{Y\times S^{d-1}} q(A, s)\, d\mu_\varepsilon.$$

We are done if we could show $\mu_\varepsilon \to p \times \lambda$ weakly as $\varepsilon \to 0$, p some probability on Y, since then

$$\iint_{Y\times S^{d-1}} q(A, s)\, d\mu_\varepsilon \to \iint_{Y\times S^{d-1}} q(A, s)\, d(p\times\lambda) = \int_{S^{d-1}} q(A, s)\, d\lambda(s) = \frac{1}{d}\operatorname{trace} A.$$

Step 1 (*Choice of noise*). Let $\xi(t)$ be a stationary ergodic diffusion process on a compact connected analytic Riemannian manifold M of nonzero dimension with invariant probability p whose generator G is defined on the C^∞ functions on M. Take, e.g., $\xi(t) =$ Brownian motion on M with generator $G = \Delta/2$, $\Delta =$ Laplace-Beltrami operator on M (see McKean [10] or Dynkin [4, p. 159]). Let $F: M \to S(d)$ be an analytic mapping of M into the linear space of skew-symmetric $d \times d$ matrices $S(d)$ such that $\int_M F(y)\, dp(y) = 0$ and with additional constraints to be specified later. Define a stationary $S(d)$-valued process $F_\varepsilon(t)$ by

$$F_\varepsilon(t) = \frac{1}{\varepsilon} F(\xi_\varepsilon(t)), \qquad \xi_\varepsilon(t) = \xi\left(\frac{t}{\varepsilon}\right).$$

Then of course $EF_\varepsilon(t) = 0$, $q(A+F_\varepsilon, s) = q(A, s)$ and G/ε generates ξ_ε. The pair process $(\xi_\varepsilon, s_\varepsilon)$, where $\dot{s}_\varepsilon = (A + q(A, s_\varepsilon)I)s_\varepsilon + F_\varepsilon(t)s_\varepsilon = h(s_\varepsilon) + F(\xi_\varepsilon)s_\varepsilon/\varepsilon$, is a diffusion process on $M \times S^{d-1}$ with generator

$$L_\varepsilon = H + \frac{1}{\varepsilon}L + \frac{1}{\varepsilon}G,$$

where $H = \langle h(s), \operatorname{grad}_s\rangle$, $L = \langle F(\xi)s, \operatorname{grad}_s\rangle$. We have $C^\infty \subset D(H) \cap D(L/\varepsilon) \cap D(G/\varepsilon) \subset D(L_\varepsilon)$. Due to compactness, L_ε has at least one invariant pr. μ_ε on $M \times S^{d-1}$, and any sequence of μ_ε's has a subsequence (μ_{ε_n}) that converges weakly to a probability μ_0 on $M \times S^{d-1}$ with marginal probability p on M. Our aim is to identify μ_0 with $p \times \lambda$.

Step 2 (μ_0 *is invariant for* $L+G$). The L_ε-invariance of μ_ε implies

$$0 = \varepsilon \int L_\varepsilon f\, d\mu_\varepsilon = \varepsilon \int Hf\, d\mu_\varepsilon + \int L_0 f\, d\mu_\varepsilon, \qquad L_0 = L + G,$$

for all $f \in D(L_\varepsilon)$. Thus for $\varepsilon_n \to 0$ we obtain

$$\int L_0 f\, d\mu_0 = 0$$

at least for all $f \in C^\infty \subset \cap_\varepsilon D(\varepsilon L_\varepsilon)$. We proceed similarly to Echeverria [5] and observe that $R_\alpha C^\infty \subset C^\infty$, $R_\alpha = (\alpha I - L_0)^{-1}$ the resolvent operator of L_0, from which

$$\int R_\alpha L_0 f\, d\mu_0 = \int L_0 R_\alpha f\, d\mu_0 = 0 \quad \text{for } f \in C^\infty$$

follows. This entails

$$\int f\, d\mu_0 = \int (\alpha R_\alpha - R_\alpha L_0) f\, d\mu_0 = \alpha \int R_\alpha f\, d\mu_0, \quad \text{all } \alpha > 0,$$

or

$$\int_0^\infty e^{-\alpha t}\left(\int f\,d\mu_0\right)dt = \int_0^\infty e^{-\alpha t}\left(\int T_t f\,d\mu_0\right)dt,$$

T_t being the semigroup generated by L_0, whence $C^\infty \subset H = \{f: f \text{ bounded and } \int f\,d\mu_0 = \int T_t f\,d\mu_0\}$. Since C^∞ is closed under multiplication and H is closed with respect to both pointwise monotone and uniform convergence, the monotone class theorem (see Williams [20, p. 49]) yields that H contains all bounded measurable functions on $M \times S^{d-1}$. Therefore, μ_0 is an invariant probability for L_0.

Step 3 (μ_0 *is unique*). The pair (ξ, s) with generator $L_0 = L + G$, where $\dot{s} = F(\xi)s$, ξ generated by G, is certainly a sample-continuous Feller–Markov process with state space $M \times S^{d-1}$. We have a unique invariant probability for (ξ, s) if the associated deterministic control system

$$\dot{s} = F(u(t))s, \quad u \text{ continuous control with values in } M,$$

is completely approximately controllable (see Arnold and Kliemann [2]). This will be satisfied if we have even complete controllability (i.e., for every $s \in S^{d-1}$ the positive orbit is equal to S^{d-1}) using only piecewise constant controls. In other words, we need a collection of values $\xi_i \in M$, $i \in I$, such that the control system given by the family of vector fields $(F(\xi_i)s)_{i\in I}$ on S^{d-1} is controllable. As we have conservative vector fields on a compact Riemannian manifold, a result of Lobry [9] applies: We have controllability if and only if

$$\operatorname{rank}(F(\xi_i)s)_{i\in I} \equiv d-1 \quad \text{on } S^{d-1}.$$

Here $\operatorname{rank}(X_i)_{i\in I}$ at a point s is defined as the dimension of the subspace in the tangent space $T_s S^{d-1}$, generated by the smallest family of vector fields containing $(X_i)_{i\in I}$ and being closed under the Lie bracket operation.

As we are only concerned with existence at that point, we can choose, e.g., $M = I =$ unit sphere in the linear space $S(d)$ of $d \times d$ skew-symmetric matrices, $F =$ identity, which will certainly satisfy the rank condition, so that μ_0 is unique.

Step 4 ($\mu_0 = p \times \lambda$). $p \times \lambda$ *is invariant for* L_0 *if and only if for all* $f \in D(L_0)$

$$0 = \int_{M\times S^{d-1}} L_0 f\,d(p\times\lambda) = \int_M \left(\int_{S^{d-1}} Lf\,d\lambda(s)\right)dp(\xi) + \int_{S^{d-1}}\left(\int_M Gf\,dp(\xi)\right)d\lambda(s).$$

As in Step 2, it suffices to check this for $f \in C^\infty$.

The generator $L = \langle F(\xi)s, \operatorname{grad}_s\rangle$ corresponds to the process (ξ, s) with $\xi(t) = \xi =$ const and $\dot{s} = F(\xi)s$. Thus $T_t f(\xi, s) = f(\xi, U(t)s)$ with orthogonal $U(t) = \exp tF(\xi)$ leaving the Lebesgue measure on S^{d-1} invariant. Therefore

$$\int_{S^{d-1}} T_t f(\xi, s)\,d\lambda(s) = \int_{S^{d-1}} f(\xi, s)\,d\lambda(s) \quad \text{and} \quad \int_{S^{d-1}} Lf(\xi, s)\,d\lambda(s) = 0$$

for each fixed $\xi \in M$. On the other hand,

$$\int_M Gf(\xi, s)\,dp(\xi) = 0 \quad \text{for each fixed } s \in S^{d-1}$$

because p is invariant for G. Thus $p \times \lambda = \mu_0$ is the unique invariant probability for L_0, and any convergent subsequence of (μ_ε) tends to μ_0.

Step 5 (μ_ε *is unique*). Since L_ε and $\varepsilon L_\varepsilon = \varepsilon H + L_0$ have the same invariant probabilities it suffices to show that the pair (ξ, s), $\dot{s} = \varepsilon h(s) + F(\xi)s$, $h(s) = (A + q(A, s)I)s$, has a unique invariant probability μ_ε on $M \times S^{d-1}$. This will again

follow from complete controllability of the associated deterministic control system $\dot{s} = \varepsilon h(s) + F(u)s$, u continuous control with values in M. We can look at this system as a perturbation of the system $\dot{s} = F(u)s$ which we assume to be completely controllable by means of finitely many vector fields $(F(\xi_1)s, \cdots, F(\xi_n)s)$, $\xi_i \in M$ (see the choice made in Step 3). As for $\varepsilon \to 0$

$$X_j^\varepsilon = \varepsilon h(s) + F(\xi_j)s \to X_j = F(\xi_j)s \qquad (j = 1, \cdots, n)$$

in the C^k topology of the set of n-tuples of C^∞ vector fields and since the set of completely controllable n-tuples is open in this topology (cf. Sussmann [17]), we have complete controllability for all ε small enough, provided we have complete controllability for $\varepsilon = 0$.

Step 6. Since $\mu_\varepsilon \to p \times \lambda$ weakly,

$$\int q(A + F_\varepsilon, s)\, d\mu_\varepsilon = \int q(A, s)\, d\mu_\varepsilon \to \int q(A, s)\, d(p \times \lambda) = \frac{1}{d} \operatorname{trace} A.$$

The proof will be completed if we can show that

$$\lambda_{\max}(A + F_\varepsilon) = \int q(A, s)\, d\mu_\varepsilon.$$

If we choose the initial r.v. $x_0^\varepsilon = s_\varepsilon(0)$ such that $(s_\varepsilon(0), \xi_\varepsilon(0))$ is independent of the Wiener process driving ξ and distributed according to μ_ε, then the corresponding solution of $\dot{x}_\varepsilon = (A + F_\varepsilon(t))x_\varepsilon$ is a Markov solution with exact Lyapunov number

$$\lambda_\varepsilon = \lambda(x_0^\varepsilon) = \int q(A, s)\, d\mu_\varepsilon.$$

We would be done if $\lambda(x_0) \leqq \lambda_\varepsilon$ for any initial r.v. x_0. Let $e_1, \cdots, e_d$ be the canonical basis in $\mathbb{R}^d$ and $x_0 = \sum_i^d \alpha_i e_i$. By an elementary property of Lyapunov numbers

$$\lambda(x_0) \leqq \max_i \lambda(e_i).$$

The law of large numbers (see Arnold and Kliemann [2]) yields $P_z(E_\varepsilon) = 1$ for μ_ε-almost all $z \in M \times S^{d-1}$, where

$$E_\varepsilon = \left\{ \omega : \lim_{t \to \infty} \frac{1}{t} \int_0^t q(A, s_\varepsilon)\, d\tau = \lambda_\varepsilon \right\}.$$

We would have $P_z(E_\varepsilon) = 1$ for *all* $z \in M \times S^{d-1}$ if $(\xi_\varepsilon, s_\varepsilon)$ were strongly Feller, whence for $i = 1, \cdots, d$

$$P(\lambda(e_i) = \lambda_\varepsilon) = \int_M P_{(\xi, e_i)}(E_\varepsilon)\, dp(\xi) = 1.$$

But the strong Feller property of $(\xi_\varepsilon, s_\varepsilon)$ follows from condition (E) of Ichihara and Kunita [8], which is satisfied if we make the choice of M, ξ and F, e.g., as in Step 3. □

Remark 2.1 (*Choice of noise*). In the proof of Theorem 2.1 stabilization was accomplished by a noise source of dimension $N = d(d-1)/2 - 1$. However, we can do the same for any d with a one-dimensional noise source. In fact, choose $M = S^1 \subset \mathbb{R}^2$, $\xi(t) =$ Brownian motion on S^1 with generator $\Delta/2$ and invariant probability λ, described by the Stratonovich equation in $\mathbb{R}^2$

$$d\xi = (Id - \xi\xi') \circ dW,$$

W = Wiener process in $\mathbb{R}^2$, $\xi = (\xi_1, \xi_2)' \in \mathbb{R}^2$. Let $F: S^1 \to S(d)$ be the polynomial

$$F(\xi_1) = \sum_{k=0}^{2(d-1)} F_k \xi_1^{2k+1}, \quad \xi_1 \in [-1, 1], \quad F_k \in S(d),$$

of degree $4d-3$ satisfying

$$F^{2k}(0) = 0, \qquad F^{(2k+1)}(0) = G_k, \qquad k = 0, \cdots, d-1,$$
$$F(\pm j/d) = \pm D_j, \qquad j = 1, \cdots, d-1.$$

The $G_k \in S(d)$ are picked to yield condition (E) in Ichihara and Kunita [8] at $\xi = (0, 1)' \in S^1$ and some $s_0 \in S^{d-1}$. The analyticity of $S^1 \times S^{d-1}$ and of all vector fields involved guarantees condition (E) at each point of $S^1 \times S^{d-1}$, thus $(\xi_\varepsilon, s_\varepsilon)$ is strongly Feller. The $D_j \in S(d)$ are picked to guarantee complete controllability of $\dot{s} = F(u)s$, $u \in C(\mathbb{R}^+, S^1)$, which entails uniqueness of μ_ε. This is ensured for

$$D_j = \begin{Bmatrix} 0 & & & & 0 \\ & \ddots & & \iddots & \\ & & 0 & 1 & \\ & & -1 & 0 & \\ & \iddots & & \ddots & \\ 0 & & & & 0 \end{Bmatrix} \leftarrow j\text{th row}, \qquad j = 1, \cdots, d-1.$$

$\uparrow$
jth column

Finally, $F(-\xi_1) = -F(\xi_1)$ guarantees $EF_\varepsilon(t) = 1/\varepsilon \int_{S^1} F(\xi_1)\, d\lambda(\xi) = 0$. Therefore, all steps of the proof of Theorem 2.1 go through with this choice of M, ξ and F.

Remark 2.2 (*Parameter-excited systems*). The theorem remains true if $A(t)$ is a stationary process with finite mean rather than a constant matrix. Let $F_0(\xi(t))$ be the stationary zero mean process with Markovian $\xi(t)$ reducing the growth of the system $\dot{x} = (EA(0))x$ which exists according to Theorem 2.1. Then the zero mean process $F(t) = F_0(\xi(t)) - (A(t) - EA(0))$ has the same growth reducing effect on $\dot{x} = A(t)x$ which becomes

$$\dot{x} = (A(t) + F(t))x = (EA(0) + F_0(\xi(t)))x$$

now driven by the Markovian $\xi(t)$. This procedure has the drawback of coupling F too much to A. However, if $A(t) = A(\eta(t))$, $\eta(t)$ a nice diffusion on a compact connected Riemannian manifold of nonzero dimension, A a nice function from that manifold into the space $\mathbb{R}^{d \times d}$ of all $d \times d$ matrices, we can choose the noise $F(t) = F(\xi(t))$ as before, but with ξ independent of η. Then the proof of Theorem 2.1 goes through with ξ replaced by (ξ, n).

Keeping in mind Theorem 1.1 (iii), we have thus proved the following corollary.

COROLLARY 2.1. *The system $\dot{x} = A(t)x$, $A(t)$ a measurable stationary process with finite mean, can be stabilized by stationary zero mean parameter noise $F(t)$ if and only if*

$$\operatorname{trace} EA(0) < 0.$$

The white noise case. The result is also true for white noise. The reasoning is pretty much the same as in the real noise case with some simplifications, so that we can skip many details.

Parametric perturbation of $\dot{x} = Ax$ by white noise means that we have to study the (Stratonovich) stochastic differential equation

$$dx = Ax\,dt + \sum_{r=1}^{m} C_r x \circ dW_r, \tag{2.1}$$

$W = (W_1, \cdots, W_m)'$ m-dimensional Brownian motion, $C_1, \cdots, C_m$ fixed $d \times d$ matrices, or, equivalently, the Ito equation

$$dx = A^0 x\,dt + \sum_{r=1}^{m} C_r x\,dW_r, \qquad A^0 = A + \tfrac{1}{2}\sum_{r=1}^{m} C_r^2.$$

The projection $s = x/|x|$ onto S^{d-1} is a diffusion process governed by

$$ds = (A - s'As)s\,dt + \sum_{r=1}^{m} (C_r - s'C_r s)s \circ dW_r, \tag{2.2}$$

while

$$|x(t)| = |x_0| \exp t\lambda(t), \qquad \lambda(t) = \frac{1}{t}\int_0^t q(s(\tau))\,d\tau + o(t),$$

$$q(s) = s'As + \tfrac{1}{2}\sum_{r=1}^{m} (s'C_r^2 s + s'C_r' C_r s - 2(s'C_r s)^2)$$

(see Khasminskii [7] and Arnold and Kliemann [2, § IV. B]). Thus, as in the real noise case, the Lyapunov numbers (of the Markov solutions) of (2.1) are determined by the long term behavior of $q(s)$ along the (Markov) solutions of (2.2). If s were ergodic on S^{d-1} with invariant probability μ, supp μ spanning $\mathbb{R}^d$, the law of large numbers tells us that the Lyapunov number for any solution starting in a fixed $s_0 \in S^{d-1}$ is equal to

$$\lambda_{\max} = \int_{S^{d-1}} q(s)\,d\mu(s)$$

for μ-almost all $s_0 \in S^{d-1}$ (for the exceptional set of s_0's we have $\lambda(s_0) \leqq \lambda_{\max}$).

THEOREM 2.2 (i) *For given A and for any choice of $C_1, \cdots, C_m$ the biggest Lyapunov number $\lambda_{\max}$ of* (2.1) *satisfies*

$$\frac{1}{d}\operatorname{trace} A \leqq \lambda_{\max}.$$

(ii) *For any $\varepsilon > 0$ there is a choice of $m = d - 1$ matrices C_r for which*

$$\frac{1}{d}\operatorname{trace} A \leqq \lambda_{\max} \leqq \frac{1}{d}\operatorname{trace} A + \varepsilon.$$

In particular, a linear system $\dot{x} = Ax$ can be stabilized by white parameter noise if and only if trace $A < 0$.

Proof. (i) Let $\Phi(t)$ be the fundamental matrix of (2.1) (i.e., the collection of solutions $\varphi_1, \cdots, \varphi_d$ starting in $e_1, \cdots, e_d$, resp.). Then $\Theta(t) = \det \Phi(t)$ satisfies

$$d\Theta = (\operatorname{trace} A)\Theta\,dt + \sum_{r=1}^{m} (\operatorname{trace} C_r)\Theta \circ dW_r, \qquad \Theta(0) = 1,$$

whose solution is

$$\Theta(t) = \exp\left((\operatorname{trace} A)t + \sum_{r=1}^{m} (\operatorname{trace} C_r)W_r(t)\right),$$

so that $(\log \Theta(t))/t \to \operatorname{trace} A$ almost surely. On the other hand

$$|\det \Phi(t)| \leqq \|\Phi(t)\|^d \leqq c\Big(\sum_{j=1}^{d} \|\varphi_j(t)\|^2\Big)^{d/2}$$

for some $c > 0$ entailing

$$\operatorname{trace} A = \lim_{t\to\infty} \frac{1}{t} \log|\det \Phi(t)| \leqq \overline{\lim_{t\to\infty}} \frac{d}{2t} \log \sum_{j=1}^{d} \|\varphi_j(t)\|^2$$

$$= d \max_j \lambda(e_j) = d\lambda_{\max},$$

which proves (i).

(ii) Choose $C_r = D_r/\varepsilon$, $\varepsilon > 0$, $r = 1, \cdots, d-1 = m$, where D_r is as in Remark 2.1. Note that because $s'C_r s = 0$ and $s'C_r^2 s = -s'C_r'C_r s$ we have $q(s) = s'As$ independent of the C_r's, and (2.2) simplifies as follows:

$$ds = (A - s'As)s\,dt + \frac{1}{\varepsilon}\sum_{r=1}^{m} D_r s \circ dW_r. \tag{2.3}$$

The deterministic nonlinear control system on S^{d-1} given by

$$\dot{s} = (A - s'As)s + \frac{1}{\varepsilon}\sum_{r=1}^{m} u_r D_r s, \qquad u_r \text{ continuous with values in } \mathbb{R},$$

is completely approximately controllable for each $\varepsilon > 0$, because the rotations represented by $\sum_1^m u_r D_r$ can move the system to any point so fast that the nonlinear part $(A - s'As)s$ is dominated. Thus s is ergodic with unique invariant probability μ_ε having $\operatorname{supp} \mu_\varepsilon = S^{d-1}$ and

$$\lambda_{\max}^{\varepsilon} = \int_{S^{d-1}} s'As\, d\mu_\varepsilon$$

is the (maximal) Lyapunov number for μ_ε-almost all starts $s_0 \in S^{d-1}$. Now we conclude with arguments similar to the ones in the real noise case that for $\varepsilon \to 0$ $\mu_\varepsilon \to \mu_0$ weakly, where μ_0 is the unique invariant probability of the process on S^{d-1} governed by

$$ds = \sum_{r=1}^{m} D_r s \circ dW_r$$

with generator

$$L_0 = \langle A^0 s, \operatorname{grad}_s \rangle + \tfrac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d} b_{jk}\frac{\partial^2}{\partial s_j\, \partial s_k},$$

$$A^0 = \tfrac{1}{2}\sum_{r=1}^{m} D_r^2, \qquad (b_{jk}) = \sum_{r=1}^{m} D_r ss' D_r'.$$

One easily checks that $L_0^* 1 = 0$ whence $\mu_o = \lambda =$ Lebesgue measure on S^{d-1}. The proof is completed by observing that

$$\lambda_{\max}^{\varepsilon} \to \int_{S^{d-1}} s'As\, d\lambda(s) = \frac{1}{d}\operatorname{trace} A \qquad (\varepsilon \to 0). \qquad \square$$

Remark 2.3. Note that we need $d-1$ independent noise sources, in contrast to the real noise case, where stabilization can be realized by a one-dimensional noise source.

3. Examples.

Linear systems. The harmonic linear oscillator with random damping and random restoring force

$$\ddot{y} + (2\beta + \xi_2(t))\dot{y} + (\omega^2 + \xi_1(t))y = 0, \tag{3.1}$$

$\beta, \omega^2 \in \mathbb{R}^+$, $(\xi_1(t), \xi_2(t))$ stationary with values in $Y \subset \mathbb{R}^2$ and $E\xi_1(t) = E\xi_2(t) = 0$, or, equivalently, $\dot{x} = A(t)x$ with $x = (y, \dot{y})'$ and

$$A(t) = \begin{pmatrix} 0 & 1 \\ -\omega^2 - \xi_1(t) & -2\beta - \xi_2(t) \end{pmatrix}$$

has been dealt with in numerous papers. For a survey see Arnold and Kliemann [2, § IV. C]. There one can find a plot of the function $\lambda_{\max} = \lambda_{\max}(\beta, \sigma^2)$ for the case $\omega^2 = 1$, $\xi_2 \equiv 0$, ξ_1 an Ornstein–Uhlenbeck process with variance σ^2. In particular, $\lambda_{\max}(0, \sigma^2) > 0$ (cf. Molčanov [11]).

Typically, the system (3.1) becomes unstable if the noise is "turned on". However, because trace $EA(t) = -2\beta$ the system can be perturbed by a zero mean noise $F(t)$ to have $-\beta \leqq \lambda_{\max}(A + F) \leqq -\beta + \varepsilon$ for any $\varepsilon > 0$. Take, e.g.,

$$F(t) = \begin{pmatrix} 0 & \eta(t) \\ \xi_1(t) - \eta(t) & \xi_2(t) \end{pmatrix},$$

η a nice fast diffusion with $E\eta = 0$ and big variance. This gives

$$A(t) + F(t) = \begin{pmatrix} 0 & 1 + \eta(t) \\ -\omega^2 - \eta(t) & -2\beta \end{pmatrix}.$$

The result is a stability improvement even for the deterministic case $\xi_1 \equiv \xi_2 \equiv 0$ and $\beta > 1$ because $\lambda_{\max}(A) = -\beta + \sqrt{\beta^2 - 1} \sim -1/2\beta \to 0$ $(\beta \to \infty)$, while $\lambda_{\max}(A + F) \approx -\beta$.

Nonlinear systems. If we linearize a nonlinear system

$$\dot{x} = f(x, \xi), \qquad \xi \text{ stationary noise}, \tag{3.2}$$

around a stationary solution x^0 we obtain the multiplicative noise linear system

$$\dot{y} = A(x^0(t), \xi(t))y, \qquad A(x^0, \xi) = \frac{\partial f}{\partial x}\bigg|_{x = x^0}. \tag{3.3}$$

If x^0 is stationarily connected with ξ, then $A(t) = A(x^0(t), \xi(t))$ is a stationary process. Therefore, we have Oseledec's multiplicative ergodic theorem (Theorem 1.1) for (3.3). The stable manifold theorems of Ruelle ([14], [15]) tell us that we have locally the same situation for the original nonlinear system (3.2) around the solution x^0. In particular, if all Lyapunov numbers of (3.3) are negative almost surely, then the stationary solution x^0 of (3.2) is exponentially stable, and x^0 can be stabilized by noise if and only if trace $EA(t) < 0$. The stabilizing noise can have the form $F(t)(x - x^0(t))$ added to the right-hand side of (3.2), with $F(t)$ appropriately chosen.

As an example, we treat the:

Mathematical model of the hypercycle. The hypercycle is a mechanism introduced by Eigen and Schuster [6] to describe prebiotic evolution. Schuster and Sigmund [16] proved that the hypercycle, described by the nonlinear system

$$\dot{x}_j = x_j(k_j x_{j-1} - \Phi(x_1, \cdots, x_d)), \qquad j = 1, \cdots, d,$$

$$\Phi(x_1, \cdots, x_d) = \sum_{r=1}^{d} k_r x_r x_{r-1}, \qquad x_0 \equiv x_d, \quad k_r \geqq 0,$$

has a unique equilibrium point in the interior of the (invariant) concentration simplex $x_j \geqq 0$, $\sum_{r=1}^{d} x_r = 1$, given by $k_1 x_d = k_2 x_1 = \cdots = k_d x_{d-1}$. This equilibrium point is asymptotically stable if and only if $d \leqq 4$. However, as the eigenvalues of the linearized $(d-1)$-dimensional system (after scaling) are

$$\lambda_j = K \exp(2\pi i j/d), \qquad j = 1, \cdots, d-1, \quad i = \sqrt{-1}, \quad K > 0,$$

we have trace $A = \sum_1^{d-1} \lambda_j = -K < 0$. Thus, by appropriately perturbing the reaction rates k_r by random noise, the hypercycle can be stabilized for $d > 4$ so that all its Lyapunov numbers around the equilibrium point are close to $-K/d < 0$.

Note added in proof. S. M. Meerkov considers the problem of stabilization by deterministic periodic parameter-excitation, which he calls vibrational control. See S. M. Meerkov, *Principle of vibrational control: theory and applications*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 755–762. Under an observability condition he arrives at the same criterion:

$$\text{trace } A < 0.$$

REFERENCES

[1] L. ARNOLD, *A new example of an unstable system being stabilized by random parameter noise*, Inform. Comm. Math. Chem. (1979), pp. 133–140.

[2] L. ARNOLD AND W. KLIEMANN, *Qualitative theory of stochastic systems*, in Probabilistic Analysis and Related Topics, Vol. 3, A. T. Bharucha-Reid, ed., Academic Press, New York, 1982.

[3] H. CRAUEL, *Ergodentheorie linearer stochastischer Systeme*, Diplomarbeit, Universität Bremen, 1981.

[4] E. B. DYNKIN, *Markov Processes*—I, Springer-Verlag, Berlin–Göttingen–Heidelberg, 1965.

[5] P. E. ECHEVERRIA, *A test for invariant measures of Markov processes*, Ph.D. Thesis, Courant Institute, New York University, New York, 1978.

[6] M. EIGEN AND P. SCHUSTER, *The Hypercycle, a Principle of Natural Selforganisation*, Springer-Verlag, Berlin–Heidelberg–New York, 1979.

[7] R. Z. KHASMINSKII, *Stochastic stability of differential equations*, Sijthoff and Noordhoff, Alphen aan den Rijn, 1980 (translation of the Russian edition, Nauka, Moscow, 1969).

[8] K. ICHIHARA AND H. KUNITA, *A classification of the second order degenerate elliptic operators and its probabilistic characterization*, Z. Wahrsch. Verw. Geb., 30 (1974), pp. 235–254.

[9] C. LOBRY, *Controllability of nonlinear systems on compact manifolds*, this Journal, 12 (1974), pp. 1–4.

[10] H. P. MCKEAN, *Stochastic Integrals*, Academic Press, New York, 1969.

[11] S. A. MOLČANOV, *The structure of eigenfunctions of one-dimensional unordered structures*, Izv. Akad. Nauk SSSR, 42 (1978), pp. 72–103; English transl. Math. USSR Izv., 12 (1978), pp. 69–101.

[12] V. I. OSELEDEC, *A multiplicative ergodic theorem. Lyapunov characteristic numbers for dynamical systems*, Trans. Moscow Math. Soc., 19 (1968), pp. 197–231.

[13] YU. A. ROZANOV, *Stationary Random Processes*, Holden-Day, San Francisco, 1967.

[14] D. RUELLE, *Ergodic theory of differentiable dynamical systems*, Publ. Math. IHES, 50 (1979), pp. 275–306.

[15] ———, *Characteristic exponents and invariant manifolds in Hilbert space*, Preprint, IHES/P/80/II, March 1980.

[16] P. SCHUSTER AND K. SIGMUND, *A mathematical model of the hypercycle*, in Dynamics of Synergetic Systems, H. Haken, ed., Springer-Verlag, Berlin–Heidelberg–New York, 1980, pp. 170–178.

[17] H. J. SUSSMANN, *Some properties of vector field systems that are not altered by small perturbations*, J. Differential Equations 20 (1976), pp. 292–315.

[18] V. WIHSTUTZ, *Ergodic theory of linear parameter-excited systems*, in Stochastic systems, M. Hazewinkel and J. C. Willems, eds., Proceedings of the NATO–ASI, Les Arcs 1980, Reidel, Dordrecht, 1981, pp. 205–218.

[19] J. C. WILLEMS AND D. AEYELS, *An equivalence result for moment stability criteria for parametric stochastic systems and Ito equations*, Internat. J. Systems Sci., 7 (1976), pp. 577–590.

[20] D. WILLIAMS, *Diffusions, Markov Processes, and Martingales Vol. 1, Foundations*, John Wiley, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0010 $01.25/0

A LYAPUNOV-LIKE CHARACTERIZATION OF ASYMPTOTIC CONTROLLABILITY*

EDUARDO D. SONTAG†

Abstract. It is shown that a control system in $\mathbf{R}^n$ is asymptotically controllable to the origin if and only if there exists a positive definite continuous functional of the states whose derivative can be made negative by appropriate choices of controls.

Key words. stabilization, controllability, Lyapunov functions, nonlinear control

1. Introduction. Lyapunov techniques have long been used in studying control problems for a system $\dot{x}(t)=f(x(t), u(t))$: Controlling so as to diminish the value of a suitable positive definite function is an obvious way of achieving stabilization, and feedback laws can be analyzed through the use of such a function—see for instance the books Barbashin [1970], Lefschetz [1965] and Letov [1961]. Sometimes one considers Lyapunov functions in conjunction with other techniques, like the analysis of sliding modes—see for instance Utkin [1977]; in these and other applications, the natural Lyapunov functions are often nondifferentiable.

In this paper we deal with the relation between the property of asymptotic controllability (every state can be driven, asymptotically, to a desired state "0", plus a local condition) and the existence of a positive definite continuous function V whose derivative can be made negative by appropriate choices of controls. If not only is the system asymptotically controllable but in fact there is a (suitable smooth) feedback law $K(\cdot)$ such that the closed loop system $\dot{x}(t)=f(x(t), K(x(t)))$ is asymptotically stable, then an inverse Lyapunov theorem can be applied to this closed loop system in order to obtain a V as above. Inverse Lyapunov results for classical (no control) systems have a long history themselves, with important contributions by Persidski, Malkin, Massera and others; a good reference is Hahn [1978]. In general, however, a continuous K fails to exist, even for very simple systems—see for instance the discussion in Sontag and Sussmann [1980]—so such an argument cannot be applied to conclude the existence of V.

The main result of this paper is that, for asymptotically controllable systems, a V as above always exists. We allow relaxed ("chattering") controls when testing the derivative of V. (Since relaxed directions belong to the convex hull of ordinary ones, the latter suffice in the C^1 case.) The proof will be based on a combination of some basic optimal control concepts and classical Lyapunov techniques (in particular, those of Zubov [1964]). For results somewhat related to this note, the reader may wish to consult the references Tokumaru et al. [1969] (gives a sufficient Lyapunov-like condition), Jacobson [1977] (gives a local necessary and sufficient criterion for a special class of systems), and Vinter [1980] (gives a time-varying functional characterizing nonreachability from a given point). Both the results and the techniques used here, however, are different from those in these references.

2. Definitions and statement of results. The systems to be studied are given by differential equations

$$\dot{x}(t)=f(x(t), u(t)) \tag{2.1}$$

* Received by the editors May 18, 1981, and in revised form May 7, 1982. This research was supported in part by U.S. Air Force grant AFOSR-80-0196.

† Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903.

with states $x(t)$ in $X := \mathbf{R}^n$, and input values $u(t)$ in a locally compact metric space U for which the balls $\{\nu | d(\mu, \nu) \leqq r\}$ are compact for each ν in U (d = distance in U). A special element "0" is distinguished in U, and the state $x = 0$ of X is an equilibrium point, i.e., $f(0, 0) = 0$. The map f is locally Lipschitz in (x, u). The set of (*ordinary*) *controls* $\mathbf{U}$ is the set of measurable and locally essentially bounded functions $u: \mathbf{R}_+ \to U$. Here $\mathbf{R}_+$ denotes the set of nonnegative reals; sometimes (depending on the context) $\mathbf{R}_+$ denotes positive reals. By abuse of notation, we shall use the same terminologies for controls defined only on a finite interval; these may be extended arbitrarily outside the interval of interest. Solutions of (2.1) are assumed to be unique and to exist locally (in t) for all controls; suitable Carathéodory-type conditions insure this. (Note that we do *not* wish to impose the (somewhat unrealistic) assumption that the solutions are always defined for all $t \geqq 0$, i.e., that there are no finite escape times.)

Given the distinguished "zero" input value 0, let $|\mu| := d(\mu, 0)$ for μ in U. The set U_r consists of all those μ with $|\mu| \leqq r$, and $\mathbf{U}_r$ is the set of all measurable $u: \mathbf{R}_+ \to U_r$, seen as a subset of $\mathbf{U}$. The set of *generalized control values* is the set $W = W(U)$ of probability measures on U; the subset of those measures supported n U_r is W_r. The set W may be topologized using the weak topology, and one introduces then the space of *relaxed controls* $\mathbf{W}$ as the set of measurable functions $w: \mathbf{R}_+ \to W$. For the topology on $\mathbf{W}$ see the references below; we shall only need to know the continuous dependence facts mentioned later. The subspaces $\mathbf{W}_r$ correspond to the relaxed controls $w(t)$ which are in W_r a.e.; each of these subspaces is sequentially compact and—identifying μ in U with the Dirac measure concentrated at $\{\mu\}$—contains (*densely*) the corresponding $\mathbf{U}_r$. A *bounded* relaxed control w is one belonging to some $\mathbf{W}_r$; the infimum of the r for which w is in $\mathbf{W}_r$ is denoted by $\|w\|$. Note that, for ordinary controls, $\|u\|$ becomes the essential supremum of the values $|u(t)|$, t in $\mathbf{R}$. (The notation $\|x\|$ will be used also for the Euclidean norm on X, but this should cause no confusion.) For details on relaxed controls, see Warga [1972], or the (very clear) presentation in Gamkrelidze [1978]; the paper Arstein [1978] summarizes most of the needed facts.

There is a natural definition of solution of (2.1) when relaxed (rather than ordinary) controls are used; see the above references for details. The solution at time t for the initial condition $x(0) = \xi$ and control w will be denoted by $x(t; \xi, w)$ or just by $x(t)$ if both ξ and w are clear from the context. For any given ξ and w there is an open set $Y := I \times N \times M$ containing $(0, \xi, w)$ such that $x(t; \eta, v)$ is well-defined for any (t, η, v) in Y. Further, if this solution is known to be defined for $0 \leqq t \leqq T$, then the map $(t, \eta, v) \to x(t; \eta, v)$ is continuous on $[0, T] \times N \times M$, for some open N, M.

We are now ready to introduce our definitions and state the main result.

DEFINITION 2.2. The system (2.1) is *asymptotically* (null-) *controllable* (a.c., for short) if and only if the following properties hold:

(i) (global part) for each ξ in X there exists an (ordinary) control u such that $x(t) = x(t; \xi, u)$ is defined for all $t \geqq 0$ and $x(t) \to 0$ as $t \to \infty$;

(ii) (local part) for each $\varepsilon > 0$ there exists a $\delta > 0$ such that for any state with $\|\xi\| < \delta$ there is a u as in (i) such that also $\|x(t)\| \leqq \varepsilon$ for all $t \geqq 0$;

(iii) (bounded controls) there exist positive η and k such that, if the ξ in (ii) satisfies also $\|\xi\| < \eta$, then the control u can be chosen with $\|u\| \leqq k$.

The above seems to be the obvious definition of a.c. if one is to model the usual uniform asymptotic stability notion in the controlled case. (Or intuitively, if the given system is to be seen as the open-loop part of some abstract closed-loop stable system.) The requirement (iii) of a bound on magnitudes of inputs required for controlling small states seems physically (and mathematically) reasonable. In fact, in order to

model "regulation with internal stability" one adds the requirement that $u(t) \to 0$ as $t \to \infty$; see Sontag [1982]. Analogous results would hold in that case. Observe that both of the (local) parts (ii) and (iii) would hold, for instance, if $U = \mathbf{R}^m$, f is differentiable at $(0, 0)$ and the linearization of (2.1) at the origin is stabilizable in the usual sense.

Given a function $V: X \to \mathbf{R}$, a state ξ, and a relaxed control w defined on an interval containing $t = 0$, let

$$\dot{V}_w(\xi) := \lim_{t \to 0^+} \frac{1}{t}[V(x(t; \xi, w)) - V(\xi)]. \tag{2.3}$$

Consider the following four properties of such a function V:

(2.4a) V is continuous;

(2.4b) $V(\xi) > 0$ for $\xi > 0$, and $V(0) = 0$ (V is positive definite);

(2.4c) the set $\{\xi | V(\xi) < r\}$ is bounded, for each r (V is proper);

(2.4d) for each ξ in X there is a relaxed w with $\dot{V}_w(\xi) < 0$, and there are positive numbers k and η such that w can be chosen with $\|w\| < k$ whenever $\|\xi\| < \eta$.

The following is an easy consequence of the above:

(2.4e) for each $\varepsilon > 0$ there is a $\theta > 0$ such that $V(\xi) < \theta$ implies $\|\xi\| < \varepsilon$.

In the next section we prove:

THEOREM 2.5. *The system* (2.1) *is asymptotically controllable if and only if there exists a V satisfying properties* (2.4a)–(2.4e).

The definition (2.3) of the (Dini) derivative along a trajectory is one immediate generalization of that used in the standard (no control) case; see for example Rouche et al. [1977]. (We could have used in in this definition a lim sup instead of a lim inf; in that case Theorem 2.5 *still holds*: The sufficiency statement becomes weaker, while the necessary part can be proved in exactly the same way.)

3. Proof of Theorem 2.5. We first establish the easy part:

A. Sufficiency. Let V, k, η be as in (2.4), and let $\varepsilon > 0$. Take a θ as in (2.4e) such that $V(\xi) < \theta$ implies that $\|\xi\| < \min\{\eta, \varepsilon\}$. A state x will be called *nicely reachable* from a given state ξ if and only if there exists an (ordinary) control u and a time $T \geqq 0$ such that:

(3.1a) $x = x(T; \xi, u)$;

(3.1b) $V(x(t; \xi, u)) < 2V(\xi)$ for $0 \leqq t \leqq T$;

(3.1c) if $V(\xi) < \theta$ then also $\|u\| < k$.

Let

$$\alpha(\xi) := \inf\{V(x) | x \text{ nicely reachable from } \xi\}. \tag{3.2}$$

Either $\alpha(\xi) = 0$ for all ξ or $\alpha(\xi) \neq 0$ for some ξ.

Case I. $\alpha(\xi) = 0$ for all ξ. Pick any ξ, and choose a ξ_1 nicely reachable and with $V(\xi_1) < V(\xi)/2$. Iterate the construction starting with ξ_1. One obtains in this way a sequence $\{\xi_i\}$ with $V(\xi_i) \to 0$ as $i \to \infty$ (hence also $\xi_i \to 0$) and such that $\xi_i = x(t_i; \xi, w)$ for an increasing sequence $\{t_i\}$ and a fixed w (obtained by concatenation). Let $T := \sup\{t_i\}$; then $V(x(t)) \to 0$ (for $x(t) := x(t; \xi, w)$) as $t \to T$. If $T < \infty$, extend w by

$w(t) := 0$ for $t \geqq T$; in any case one concludes that $x(t) \to 0$ as $t \to \infty$, with *all* $x(t)$ nicely reachable from ξ. This gives the first part of the a.c. definition. Pick now a $\delta > 0$ such that $\|x\| < \delta$ implies $V(x) < \theta/2$. If $\|\xi\| < \delta$ for the above ξ, then all $x(t)$ in the obtained trajectory satisfy $V(x(t)) < 2V(\xi) < \theta$. It follows that $\|x(t)\| < \varepsilon$, as required in (ii) of the a.c. definition. Finally, part (iii) is satisfied by construction, using the same k and any $\eta' > 0$ for which $\|\xi\| < \eta'$ implies $V(\xi) < \theta$.

Case II. $\alpha(\xi) > 0$ for some ξ. We shall derive a contradiction. Let $\{x_n\}$ be a sequence of nicely reachable states with $V(x_n) \to \alpha := \alpha(\xi)$. All these x_n belong to the compact set

$$\{x \mid V(x) \leqq V(\xi)\}; \tag{3.3}$$

replacing $\{x_n\}$ by an appropriate subsequence we may assume that

$$x_n \to \zeta, \qquad V(\zeta) = \alpha \leqq V(\xi). \tag{3.4}$$

By property (2.4d), there is then a sequence $t_i \to 0^+$ and a relaxed w with $V(x(t_i; \xi, w)) < V(\zeta) = \alpha$ for each t_i. Further, if $V(\xi) < \theta$ then also $V(\zeta) < \theta$ so one may pick such a w with $\|w\| < k$. It follows from the continuity of $V(x(t; \zeta, w))$ on t that there is an i such that (with $t := t_i$) also

$$V(x(s; \zeta, w)) < \tfrac{3}{2}V(\xi) \quad \text{for } 0 \leqq s \leqq t. \tag{3.5}$$

Thus for (w', ζ') sufficiently close to (w, ζ) it holds that

$$V(x(s; \zeta', w')) < 2V(\xi) \quad \text{for } 0 \leqq s \leqq t. \tag{3.6}$$

Pick an ordinary control $w' = u$ such that this holds and such that also

$$V(x(t; \zeta, u)) < \alpha. \tag{3.7}$$

If $V(\xi) < \theta$, require also that $\|u\| < k$. (Recall that ordinary controls are dense in $\mathbf{W}_k$.)

Let now $z_n := x(s; x_n, u)$. For large enough n, (3.4)–(3.6) give a z_n nicely reachable from ξ and with $V(z_n) < \alpha$. This contradicts the minimality of α.

B. Some bounding functions. We now start proving that a.c. implies the existence of a V as above. We shall need a sequence of basic lemmas. In order to simplify notations, $g(\pm\infty)$ will mean $\lim_{p \to \pm\infty} g(p)$, and $g(0) := \lim_{p \to 0^+} g(p)$ for a function defined on positive reals only. A fixed asymptotically controllable system is assumed given; the numbers k and η are as in the a.c. definition.

LEMMA 3.8. *There exist a positive number* $p_0 < 1$ *and maps* $\tau, \phi, \mu : \mathbf{R}_+ \to \mathbf{R}$, $m : \mathbf{R} \to \mathbf{R}_+$ *and* $K : X \to \mathbf{U}$, *where* ϕ, μ *and* m *are continuous,* m *is strictly decreasing,* ϕ *is strictly increasing and* μ *is nondecreasing, such that the following properties hold*:

a) $m(-\infty) = +\infty$, $m(+\infty) = 0$, $m(0) = 1$;
b) $p \leqq \phi(p)$ *for all* p, $\phi(0) = 0$;
c) $\tau(p) = 0$ *for* $0 \leqq p \leqq p_0$;
d) $\mu(p) = k$ *for* $0 \leqq p \leqq p_0$, $\mu(+\infty) = +\infty$;
e) *for each* $\xi \neq 0$, *with* $x(t) := x(t; \xi, K(\xi))$:
   (i) $\|K(\xi)\| \leqq \mu(\|\xi\|)$,
   (ii) $\|x(t)\| \leqq \phi(\|\xi\|)$ *for* $t \geqq 0$,
   (iii) $x(t) \to 0$ *as* $t \to \infty$,
   (iv) *for* $t \geqq \tau(\|\xi\|)$, $\|x(t)\| \leqq m(t - \tau(\|\xi\|))$.

Proof. Part 1. We shall first construct sequences of nonnegative numbers $\{\varepsilon_i\}$, $\{T_i\}$ and $\{b_i\}$, i in $\mathbf{Z}$, such that: $\{\varepsilon_i\}$ is strictly increasing, $\varepsilon_i \to 0$ (resp. $+\infty$) as $i \to -\infty$ (resp. $+\infty$), and so that for each ξ for which $\|\xi\| < \varepsilon_i$ there is an (ordinary) control u satisfying $\|u\| < b_i$ and $\|x(t; \xi, u)\| < \varepsilon_{i+1}$ for all $t \geqq 0$ and also $\|x(t; \xi, u)\| < \varepsilon_{i-1}$ for $t \geqq T_i$.

Let $\varepsilon_0 := \frac{1}{2}$. By induction on Definition 2.2 one concludes the existence of a sequence of numbers ε_i, $i \leqq -1$, such that: for each ξ with $\|\xi\| \leqq \varepsilon_{i-1}$ there is a control u with $\|u\| \leqq k$, $\|x(t; \xi, u)\| < \varepsilon_i$ and $x(t) \to 0$ as $t \to \infty$; one may take the $\{\varepsilon_i\}$ strictly decreasing and approaching 0 as $i \to -\infty$. Further, one may assume that $\varepsilon_{-1} \leqq \eta$. Consider now a fixed $i < 0$ and take a ξ with $\|\xi\| \leqq \varepsilon_i$. There is then an u as above and some $T = T(\xi)$ such that $\|x(T)\| < \varepsilon_{i-2}$ for the corresponding solution. By continuity of $x(\cdot\,; \cdot, u)$ there is an open neighborhood $H(\xi)$ of ξ such that, for each z in $H(\xi)$, $\|z(t)\| < \varepsilon_{i+1}$ for $0 \leqq t \leqq T$ and $\|z(T)\| < \varepsilon_{i-2}$, where $z(t) := x(t; z, u)$. By construction of ε_{i-2}, there is for each such $z(T)$ a control v with $\|v\| < k$ (not necessarily the same u) such that

$$\|x(t; z(T), v)\| < \varepsilon_{i-1}, \qquad t \geqq 0. \tag{3.9}$$

Concatenating the restriction of u to $[0, T]$ with this v, one concludes that for each z in $H(\xi)$ there is some (ordinary) control with the resulting trajectory having $\|z(t)\| < \varepsilon_{i-1}$ for *all* $t \geqq T(\xi)$ while keeping $\|z(t)\| < \varepsilon_{i+1}$ for all t. (Note that the input to be applied in order to achieve this depends on the particular z; for the original u there may be no neighborhood on which this controllability is achieved.) The $H(\xi)$ cover the ball of radius ε_i; pick a finite subcover. Let $T_i :=$ largest of the $T(\xi)$ for this subcover. With all $b_i := k$ the sequences $\{\varepsilon_i\}, \{b_i\}, \{T_i\}$ satisfy the requirements for $i < 0$.

We now define the sequences for $i \geqq 0$, by induction on increasing i. Assume that ε_i, b_{i-1} and T_{i-1} have been already defined (recall for the first step that $\varepsilon_0 = \frac{1}{2}$). By property (i) in the definition of a.c., it follows that for each ξ with $\|\xi\| \leqq \varepsilon_i$ there exists some u and some $T = T(\xi)$ with $\|x(T)\| < \varepsilon_{i-2}$. By induction, it is possible to control $x(T)$ in such a way as to stay in the ball of radius ε_{i-1}. These further controls can be chosen with $\|v\| < b_{i-1}$. An argument like the one in the previous paragraph gives a fixed T_i such that each state ξ as above is controlled to $\|x(T_i)\| < \varepsilon_{i-1}$ by appropriate choice of controls. Further, all these controls are obtained by concatenating one of a finite number of controls u_j (finite subcover argument) with controls with $\|v\| < b_{i-1}$. Let b_i be larger than b_{i-1} and all $\|u_j\|$. To complete the induction step we need to define ε_{i+1}. Consider the set

$$\{x(t; \xi, u) \mid \|\xi\| \leqq \varepsilon_i, \|u\| \leqq b_i, 0 \leqq t \leqq T_i\}. \tag{3.10}$$

Since this set is compact, it is contained in the interior of some ball of radius ε_{i+1}. For simplicity of future arguments, we shall assume that the sum of the T_i, $i < 0$, is infinite; larger T_i's can always be chosen in the above constructions. This completes Part 1.

Part 2. Let $\phi : \mathbf{R}_+ \to \mathbf{R}$ be any continuous strictly increasing function such that $\phi(0) = 0$ and, for all i

$$\phi(p) > \varepsilon_{i+1} \quad \text{for } p \text{ in } [\varepsilon_{i-1}, \varepsilon_i]. \tag{3.11}$$

Let $p_0 := \varepsilon_{-1}$. Take $\mu : \mathbf{R}_+ \to \mathbf{R}$ to be any continuous nondecreasing function having $\mu(p) = b_0$ for $0 \leqq p \leqq p_0$, $\mu(+\infty) = +\infty$, and such that, for all i,

$$\mu(p) \geqq b_i \quad \text{for } p \text{ in } (\varepsilon_{i-1}, \varepsilon_i]. \tag{3.12}$$

Denote $t_0 := 0$ and $t_i := T_{-1} + \cdots + T_{-i}$ for $i > 0$. Let m be as in a) and satisfying, for $i \geqq 0$,

$$m(t) > \varepsilon_{-i} \quad \text{for } t \text{ in } [t_i, t_{i+1}]. \tag{3.13}$$

Let τ be the step function with value 0 for $p \leqq p_0$ and for $i \geqq 0$,

$$\tau(p) = T_0 + \cdots + T_i \quad \text{for } p \text{ in } (\varepsilon_{i-1}, \varepsilon_i]. \tag{3.14}$$

The open-loop "choice" function K is introduced merely for notational convenience; no smoothness of any kind is required. Given a state ξ, say with $\varepsilon_{i-1} < \|\xi\| \leqq \varepsilon_i$, find a control u_1 sending ξ to $\xi_1 = x(T_i; \xi, u)$ with $\|u_1\| < b_i$, $\|\xi_1\| < \varepsilon_{i-1}$, and all intermediate states with $\|x(t)\| < \varepsilon_{i+1}$. Repeat with ξ_1, finding a u_2, ξ_2. Iterating this construction, let $K(\xi)$ be the concatenation of all the u_j thus obtained. The corresponding trajectory satisfies $x(t) \to 0$ as $t \to \infty$. By construction, $\|K(\xi)\| < b_i$, which is less than $\mu(\|\xi\|)$ by (3.12). Also, $\|x(t)\| < \varepsilon_{i+1}$ for all $t \geqq 0$, and by (3.11) also $\|x(t)\| < \phi(\|\xi\|)$.

We now need only establish property (iv). Assume first that $\|\xi\| < \varepsilon_i$, with $i = -1$. Then $\tau(\|\xi\|) = 0$, and the above construction insures that $\|x(t)\| < \varepsilon_{-j}$ when t is in $[t_j, t_{j+1}]$. By (3.13), property (iv) holds. If $i < -1$, the trajectory has $\|x(t)\| < \varepsilon_{-j}$ when t is in $[t_j - L_i, t_{j+1} - L_i]$, where

$$L_i := T_{-1} + \cdots + T_{i+1}, \tag{3.15}$$

for any $j \geqq -i-1$, and $\|x(t)\| < \varepsilon_{i+1}$ for all $t \geqq 0$. If t is in $[t_j, t_{j+1}]$, $j < -i-1$, then $\|x(t)\| < \varepsilon_{-j}$ by the latter fact; if $j \geqq -i-1$ then $t \geqq t_j > t_j - L_i$ gives again that $\|x(t)\| < \varepsilon_{-j}$. Again by (3.13), property (iv) of (3.8c) holds. There remains the case $i \geqq 0$. In that case, after time

$$T = T_i + T_{i-1} + \cdots + T_0 = \tau(\|\xi\|) \tag{3.16}$$

a state $x(t)$ with $\|x(t)\| < \varepsilon_{-1}$ is reached, and after that the trajectory has states bounded by $m(t-T)$ (by the case $i = -1$).

LEMMA 3.17. *(The notations of the previous lemma still hold). There exist continuous strictly increasing functions $N, \psi, \nu: \mathbf{R}_+ \to \mathbf{R}_+$, with $N(0) = 0$, $N(+\infty) = +\infty$, such that the following properties hold. For any state ξ and for any relaxed control w, let*

$$R(\xi, w) := \int_0^\infty N(\|x(t; \xi, w)\|)\, dt + \max\{\|w\| - k, 0\} \tag{3.18}$$

if the solution is defined for all $t \geqq 0$, and $R(\xi, w) := +\infty$ otherwise. Then, for each ξ:

(a) *$R(\xi, K(\xi)) < \infty$;*

(b) *if $\|\xi\| \leqq p_0$ then $R(\xi, K(\xi)) \leqq \phi(\|\xi\|)$;*

(c) *if $R(\xi, w) < R(\xi, K(\xi))$ for some w, then $\|x(t; \xi, w)\| \leqq \psi(\|\xi\|)$ for all $t \geqq 0$ and $\|w\| \leqq \nu(\|\xi\|)$;*

(d) *for each $\alpha > 0$ there is a $\theta > 0$ such that if $R(\xi, w) < \alpha$ for some w, then $\|\xi\| \leqq \theta$;*

(e) *for each $\alpha > 0$ there is a $\beta > 0$ such that $\|\xi\| > \alpha$ implies that $R(\xi, w) > \beta$ for all w.*

Proof. Let $n := m^{-1}: \mathbf{R}_+ \to \mathbf{R}$ be the inverse function of m. Define the function

$$N_1(p) := p \exp[-n(p)]. \tag{3.19}$$

Both N_1 and $\exp[-n(\cdot)]$ are strictly increasing continuous functions, $N_1(0) = 0$, $N_1(+\infty) = +\infty$.

For each triple of positive numbers (a, b, c), choose the quantity $\gamma(a, b, c) > 0$ in such a way that the following property holds: if w is any control with $\|w\| \leqq a$, and if ξ_1, ξ_2 are any states with

$$\xi_2 = x(T; \xi_1, w) \tag{3.20}$$

for some $T > 0$ and the trajectory having

$$b \leqq \|x(t; \xi_1, w)\| \leqq c \quad \text{for } 0 \leqq t \leqq T, \tag{3.21}$$

then

$$\int_0^T N_1(\|x(t)\|)\,dt \geqq \gamma(a,b,c)\|\xi_1-\xi_2\|. \tag{3.22}$$

Such a quantity always exists, because the integrand is bounded below by $N_1(b)$ (so that the left side is at least $T\cdot N_1(b)$), while $\|\xi_1-\xi_2\|$ is bounded by TB, where B is a bound on the values $f(x,\mu)$ for $\|x\|\leqq c$ and $|\mu|\leqq a$.

We shall define inductively a nondecreasing sequence of maps N_j, $j\geqq 1$, starting with (3.19), and will then let (pointwise)

$$N := \lim_{j\to\infty} N_j. \tag{3.23}$$

This limit will be finite because the following property will hold by construction for each j:

$$p\leqq\phi(j) \quad \text{implies } N_j(p)=N_{j+1}(p) \quad (=N(p)). \tag{3.24}$$

(Note that $\phi(j)\to+\infty$ as $j\to\infty$.) Further, all the functions N_j (hence N itself) will be continuous (or even C^∞ if desired). Assume then that N_i has been defined for all $i\leqq j$, in such a way that (3.24) holds for $i=1,\cdots,j-1$. Introduce the quantities

$$r_j := N_j(\phi(j))\tau(j)+\phi(j), \tag{3.25}$$

$$L_j(\xi,w) := \int_0^\infty N_j(\|x(t;\xi,w)\|)\,dt \tag{3.26}$$

with $L_j(\xi,w) := +\infty$ if there is a finite escape time.

Take now any ξ with $\|\xi\|\leqq j$. When $t\geqq\tau(j)$, it follows from Lemma 3.8e(iv) that

$$\|x(t;\xi,K(\xi))\|<m(t-\tau(\|\xi\|))\leqq m(t-\tau(j)) \tag{3.27}$$

and hence also $\|x(t)\|<m(0)=1<\phi(1)$ for these t. Thus,

$$\int_{\tau(j)}^\infty N_j(\|x(t)\|)\,dt = \int_{\tau(j)}^\infty N_1(\|x(t)\|)\,dt \tag{3.28}$$

(by (3.24)), and by (3.27) this is less than

$$\phi(1)\int_{\tau(j)}^\infty \exp\left[-n(m(t-\tau(j)))\right]dt, \tag{3.29}$$

which equals $\phi(1)$. Since $\|x(t;\xi,K(\xi))\|<\phi(j)$ for all $t\geqq 0$ (again from Lemma 3.8e), one has the bound

$$L_j(\xi,K(\xi))\leqq r_j. \tag{3.30}$$

Define also

$$\gamma_j := \gamma(r_j+\mu(j)+k,\ \phi(j)+1,\ \phi(j)+2). \tag{3.31}$$

Let g_j be any continuous nonnegative function from $\mathbf{R}_+$ to $\mathbf{R}$ which vanishes outside the interval $[\phi(j),\phi(j)+3]$ and such that

$$g_j(p) := \frac{r_j+\mu_j}{\gamma_j} \tag{3.32}$$

for p in the interval $[\phi(j)+1,\phi(j)+2]$. The induction step is provided by the definition

$$N_{j+1}(p) := (1+g_j(p))N_j(p). \tag{3.33}$$

Note that (3.24) indeed holds. This completes the construction of N.

Finally, pick any ψ and ν continuous, increasing, and such that

$$\nu(p) > r_j + \mu(j) + k, \tag{3.34}$$

and

$$\psi(p) > \phi(j) + 2, \tag{3.35}$$

whenever p is in $[j-1, j]$.

We now prove that (a) to (e) hold. Define L as in (3.26) but using N in the integrand. Pick a ξ with $\|\xi\| < j$. The trajectory corresponding to the control $K(\xi)$ has $\|x(t)\| < \phi(j)$ for $t \geqq 0$, so, by (3.24), $L(\xi, K(\xi)) = L_j(\xi, K(\xi)) < r_j$. Thus

$$R(\xi, K(\xi)) \leqq r_j + \mu(j), \tag{3.36}$$

and in particular (a) holds. Assume that j was chosen so that $\|\xi\| > j-1$. Suppose now that w is a control for which $R(\xi, w) < R(\xi, K(\xi))$. By (3.36),

$$\|w\| \leqq r_j + \mu(j) + k < \nu(\|\xi\|), \tag{3.37}$$

as wanted in (c). Now assume that for the corresponding trajectory there would be some t with $\|x(t)\| > \psi(\|\xi\|) > \phi(j) + 2$. There are then times $t_1 < t_2$ such that $\|x(t_1)\| = \phi(j) + 1$, $\|x(t_2)\| = \phi(j) + 2$ and

$$\phi(j) + 1 \leqq \|x(t)\| \leqq \phi(j) + 2 \tag{3.38}$$

for $t_1 \leqq t \leqq t_2$. Hence

$$\begin{aligned} R(\xi, w) &\geqq \int_{t_1}^{t_2} N(\|x(t)\|)\, dt \geqq \int_{t_1}^{t_2} N_{j+1}(\|x(t)\|)\, dt \\ &\geqq (1 + g_j(\phi(j)+1)) \int_{t_1}^{t_2} N_1(\|x(t)\|)\, dt \geqq r_j + \mu(j) \geqq R(\xi, K(\xi)). \end{aligned} \tag{3.39}$$

This contradicts the choice of ξ and w. So (c) holds. To prove (b), let $\|\xi\| \leqq p_0$. Then $\tau(\|\xi\|) = 0$ and $\|K(\xi)\| \leqq k$, so, for $w = K(\xi)$,

$$R(\xi, K(\xi)) = \int_0^\infty N_1(\|x(t)\|)\, dt \leqq \phi(\|\xi\|) \int_0^\infty e^{-t}\, dt = \phi(\|\xi\|). \tag{3.40}$$

We now establish (d) and (e). Given $\alpha > 0$ choose any integer $j > 0$ so that $\alpha + k < \mu(j)$. If $R(\xi, w) < \alpha$ then $\|w\| < \mu(j)$. We claim that $\|\xi\| < \theta := \phi(j) + 2$. Otherwise, $\|\xi\| \geqq \phi(j) + 2$ implies that there exist $t_1 < t_2$ with $\|x(t_1)\| = \xi(j) + 2$, $\|x(t_2)\| = \xi(j) + 1$, and all $\|x(t)\|$ bounded by these values for $t_1 \leqq t \leqq t_2$. By an argument similar to the one used above, $R(\xi, w)$ can be proved to be larger than $r_j + \mu(j) > \alpha$, a contradiction. Thus (d) holds. Assume now that $\|\xi\| > \alpha$, and let $\gamma' := \gamma(k+1, \alpha/2, \alpha)$. Let w be given. If $\|w\| > k+1$ then $R(\xi, w) > 1$; otherwise $R(\xi, w) > \alpha\gamma'/2$. With

$$\beta := \min\left\{k+1, \frac{\alpha\gamma'}{2}\right\}, \tag{3.41}$$

(e) is also established.

C. The function V.

$$V(\xi) := \inf\{R(\xi, w) | w \text{ relaxed control}\}. \tag{3.42}$$

Note that $V(0)=0$, so by (3.18b) the function V is continuous at zero; V is always finite by (3.18a). By (3.18d), the sets $\{\xi | V(\xi)<\alpha\}$ are always bounded. By (3.18e), $V(\xi)>0$ for $\xi \neq 0$.

LEMMA 3.43. *Let $(\xi_n, w_n)\to(\xi, w)$ as $n\to\infty$, and assume that all $R(\xi_n, w_n)$ and $R(\xi, w)$ are finite. Then $R(\xi, w)\leqq \underline{\lim}\, R(\xi_n, w_n)$.*

Proof. Let $x_n(t) := x(t; \xi_n, w_n)$ and $x(t) := x(t; \xi, w)$. Then $N(\|x_n(t)\|)$ converges to $N(\|x(t)\|)$ for each t. By Fatou's lemma,

$$\int_0^\infty N(\|x(t)\|)\,dt \leqq \underline{\lim} \int_0^\infty N(\|x_n(t)\|)\,dt. \tag{3.44}$$

If the measures $w_j(t)$ are all supported in some U_r, for some subsequence $\{w_j\}$, then also $\|w\|\leqq r$. It follows that

$$\|w\| \leqq \underline{\lim}\, \|w_n\|. \tag{3.45}$$

Thus also

$$\max\{\|w\|-k, 0\} \leqq \underline{\lim}\, (\max\{\|w_n\|-k, 0\}), \tag{3.46}$$

and the conclusion follows from (3.44) and (3.46) and the elementary fact that always $\underline{\lim}\, a_n + \underline{\lim}\, b_n \leqq \underline{\lim}\, (a_n + b_n)$.

LEMMA 3.47. *For each ξ there is a w^* with $R(\xi, w^*) = V(\xi)$.*

Proof. Let $\{R(\xi, w_n)\}$ be a minimizing sequence. By (3.18d) we may assume that all $\|w_n\| \leqq \nu(\|\xi\|) =: r$ and

$$\|x_n(t)\| \leqq \psi(\|\xi\|) \text{ for } t \geqq 0. \tag{3.48}$$

By sequential compactness of $\mathbf{W}_r$ (a subsequence of) $\{w_n\}$ converges to a control w^* in $\mathbf{W}_r$. The solution $x(t) := x(t; \xi, w^*)$ is defined for all t, because (3.48) implies that $\|x(t)\|$ is bounded independently of t. By (3.43),

$$R(\xi, w^*) \leqq \lim R(\xi, w_n) = V(\xi), \tag{3.49}$$

so $V(\xi) = R(\xi, w^*)$, as wanted.

LEMMA 3.50. *V is lower semicontinuous.*

Proof. Let $\{\xi_n\}$ be a sequence converging to ξ, and write $V(\xi_n) = R(\xi_n, w_n)$. For a suitable subsequence one may assume that $w_n \to w$, for an appropriate w. (All $\|w_n\|$ are bounded by $\nu(\|\xi\|)+\delta$, some δ.) Thus

$$V(\xi) \leqq R(\xi, w) \leqq \underline{\lim}\, R(\xi_n, w_n) = \underline{\lim}\, V(\xi). \tag{3.51}$$

LEMMA 3.52. *V is continuous.*

Proof. We only need to establish upper semicontinuity. Pick $\varepsilon > 0$. Choose a positive $\delta < p_0$ so that

$$p < \delta \text{ implies } \phi(p) < \frac{\varepsilon}{2}. \tag{3.53}$$

Pick any state ξ, and let $R(\xi, w) = V(\xi)$. Since $x(t) := x(t; \xi, w)$ necessarily converges to zero, there is some T with $\|x(T)\| < \delta$. There is also a neighborhood H of ξ such that, for each z in H, and for the above w, T,

$$\int_0^T N(\|z(t)\|)\,dt < \int_0^T N(\|x(t)\|)\,dt + \frac{\varepsilon}{2} \tag{3.54}$$

for the corresponding solution with $z(0) = z$, and such that also $\|z(T)\| < \delta$. By (3.53) and (3.18b) (continuity at 0), $V(z) < V(\xi) + \varepsilon$. This proves upper semicontinuity.

We are only left with establishing (2.4d). Let ξ be any state, $V(\xi)=R(\xi, w)$. Take any $z=x(t)$ in the trajectory, and let w' be the translation of w by $(-t)$, so in particular $\|w'\|\leqq\|w\|$. It follows that

$$V(z)\leqq R(z, w')\leqq\int_t^\infty N(\|x(s)\|)\,ds+\max\{\|w'\|-k, 0\}. \tag{3.55}$$

So

$$\lim_{t\to 0^+}\frac{1}{t}[V(z)-V(\xi)]<\lim_{t\to 0^+}\frac{1}{t}\left(-\int_0^t N(\|x(s)\|)\,ds\right)=-N(\|\xi\|)<0. \tag{3.56}$$

Further, $\|w\|$ is bounded by $\nu(\|\xi\|)$. This completes the proof of the theorem.

REFERENCES

Z. ARSTEIN [1978], *Relaxed controls and the dynamics of control systems*, this Journal, 16, pp. 689–701.

E. A. BARBASHIN [1970], *Introduction to the Theory of Stability*, Walters–Noordhoff, Groningen.

R. V. GAMKRELIDZE [1978], *Principles of Optimal Control Theory*, Plenum Press, New York.

W. HAHN [1978], *Stability of Motion*, Springer, Berlin.

D. H. JACOBSON [1977], *Extensions of Linear-quadratic Control, Optimization, and Matrix Theory*, Academic Press, New York.

S. LEFSCHETZ [1965], *Stability of Nonlinear Control Systems*, Academic Press, New York.

A. L. LETOV [1961], *Stability in Nonlinear Control Systems*, Princeton University Press, Princeton, NJ.

N. ROUCHE, P. HABETS AND M. LALOY [1977], *Stability Theory by Lyapunov's Direct Method*, Springer-Verlag, New York.

E. D. SONTAG AND H. J. SUSSMANN [1980], *Remarks on continuous feedback*, Proc. IEEE Conference Decision and Control, Albuquerque, NM, pp. 916–921.

E. D. SONTAG [1982], *Abstract regulation of nonlinear systems: Stabilization*, in Feedback Control of Linear and Nonlinear Systems, Lecture Notes in Control and Information Sciences, 39, D. Hinrichsen and A. Isidori, eds., Springer, Berlin, pp. 227–243.

H. TOKUMARU, N. ADACHI AND A. INOUE [1969], *On the controllability of non-linear control systems*, Memoirs Faculty of Engr. Kyoto, 31, pp. 402–424.

V. I. UTKIN [1977], *Variable structure systems with sliding mode: A survey*, IEEE Trans. Automat. Control., AC-22, pp. 212–222.

R. VINTER [1980], *A characterization of the reachable set for nonlinear control systems*, this Journal, 18, pp. 599–610.

J. WARGA [1972], *Optimal Control of Differential and Functional Equations*, Academic Press, New York.

V. I. ZUBOV [1964], *Methods of A. M. Lyapunov and their Application*, Noordhoff, Groningen.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0011$01.25/0

ON THE RELATION OF ZAKAI'S AND MORTENSEN'S EQUATIONS*

VÁCLAV E. BENEŠ† AND IOANNIS KARATZAS‡

Abstract. The problem of optimal control for partially observable diffusion processes is studied by the dynamic programming in function space approach first proposed by R. E. Mortensen. The density of the conditional distribution of the (unobservable) signal given past and present observations, which satisfies the Zakai equation of nonlinear filtering, is viewed as the new state of the system. A verification result is established for the corresponding Bellman–Hamilton–Jacobi equation, known as Mortensen's equation.

1. Introduction. For some time now, research in stochastic control has concentrated on understanding the case of partial or incomplete information, especially on extending the so-called separation principle to cases other than that of linear dynamics. Heuristically, the idea is to replace the system state by its conditional distribution (usually, density) given past observations, and to use the latter as a new state in a problem with complete information. The usual separation theorem really does this already, although its exegeses do not always make this fact clear. One is thus led at once to measure- (or density-)valued stochastic processes, and to stochastic control problems with an infinite-dimensional state space.

Credit for the original suggestion in this direction clearly belongs to R. E. Mortensen, although some of his ideas were presaged by Kushner [10] and Shiryaev [16]. In a brilliantly prophetic paper [14] published in 1966, he mapped out the structure of such a program for signal processes and observations obeying a pair of stochastic differential equations in Euclidean space forced by two independent Wiener processes, and in particular, he wrote down the appropriate nonlinear Bellman equation for the value function (equation (6.10) of the present paper). This is the equation we call Mortensen's, and it is a natural generalization of the classical Bellman–Hamilton–Jacobi equation used in the control of finite-dimensional systems with complete observations. In addition, he saw clearly that since the new state was to be the conditional density, the new dynamical equation would have to be (essentially) Zakai's equation (4.6) for the conditional density. Actually, he was the first to derive formally this equation. At this point (1966) his insights were blunted by lack of progress in nonlinear filtering; the only example that could be done was the well-known linear dynamics case leading to the Kalman–Bucy filter, and even this example was limited to Gaussian initial densities. Thus, Mortensen's equation has lain unused, if not unnoticed, for some fourteen years.

We are glad to report that recent progress in nonlinear filtering should place Mortensen's equation in the central position it deserves: it is the key to the separation principle. For years, many workers in stochastic control have entertained or advanced the idea that nonlinear filtering was the key to optimal control in the case of incomplete information. The problem was to find the lock into which the key fit. Here is a sketch of that lock; a parallel approach, constructing the nonlinear semigroup corresponding to Mortensen's equation, can be found in recent work by Fleming [6] and Davis–Kohlmann [3]:

* Received by the editors March 30, 1981, and in revised form April 30, 1982. This research was supported by the National Science Foundation under grants MCS-79-05774 and MCS-81-03435.

† Bell Laboratories, Murray Hill, New Jersey 07974.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. Current address: Department of Mathematical Statistics, Columbia University, New York, New York 10027.

(1) One can prove a verification principle for a version of Mortensen's equation, to the effect that a solution of it is a lower bound on the achievable cost. Such a principle was already envisaged by Mortensen; a proof is sketched in § 6.

(2) Zakai's equation is for an unnormalized density, which is a linear functional of the initial density. It is not hard to see that normalization induces a linear fractional dependence on the initial density (Theorem 4.1).

(3) Expected cost can be calculated by integration of its conditional expectation. But the dependence on p it inherits, under this integration, is actually linear. Thus, for any admissible control law u, the cost $V^u(\tau, p)$ of operating over $[0, \tau]$, starting in "state" p, i.e. with all information summarized in the density p, is linear in p (relation (5.1) in text) under expectation with respect to an appropriate probability measure.

(4) For a constant control u, the cost $V^u(\tau, p)$ satisfies the Hamilton–Jacobi form of a simplified Mortensen's equation, with the minimization left out and $u =$ constant put in. This is proved directly in § 5. The linear dependence makes it possible to calculate the first and second functional derivatives with respect to p that occur in the equation.

2. Formulation, assumptions and notation. We start with a probability space $(\Omega, \mathbf{F}, \tilde{P})$ and with an $(r+m)$-dimensional Wiener process $\{(w_t, y_t);\ 0 \leqq t \leqq T\}$ defined on it; in other words, $\{w_t;\ 0 \leqq t \leqq T\}$ and $\{y_t;\ 0 \leqq t \leqq T\}$ are independent Wiener processes, of dimensions r and m, respectively. We also consider an $\mathbf{R}^r$-valued random variable x_0 defined on the same space, with probability distribution having a density p in the space of functions

$$\mathbf{E}_k \triangleq \left\{ p \in L^1(\mathbf{R}^r);\ \|p\|_k = \int_{\mathbf{R}^r} (1+|z|^k)|p(z)|\,dz < \infty \right\} \tag{2.1}$$

for some $k \geqq 1$; under the indicated norm, $\mathbf{E}_k$ is a normed linear space. Let x_0 be independent of $\sigma\{(w_t, y_t);\ 0 \leqq t \leqq T\}$, and

$$\mathbf{F}_t \triangleq \sigma(x_0, w_s, y_s;\ 0 \leqq s \leqq t), \qquad 0 \leqq t \leqq T, \tag{2.2}$$

be the basic increasing family of σ-fields, satisfying the usual conditions (right continuity and completion by negligible sets), with $\mathbf{F} \equiv \mathbf{F}_T$.

We make the following assumptions:

Assumption A1. The function $f(t, z) : [0, T] \times \mathbf{R}^r \to \mathbf{R}^r$ is continuous in t, continuously differentiable in z, with gradient satisfying the condition

$$|\nabla f(t, z)| \leqq K, \qquad (t, z) \in [0, T] \times \mathbf{R}^r,$$

for some constant $K > 0$.

Assumption A2. The function $h(t, z) : [0, T] \times \mathbf{R}^r \to \mathbf{R}^m$ is in $C^{1,2}_{t,z}([0, T] \times \mathbf{R}^r)$, with gradient and Laplacian satisfying the boundedness condition

$$|\nabla h(t, z)| + |\Delta h(t, z)| \leqq K$$

and time derivative satisfying the linear growth condition

$$|h_1(t, z)| \leqq K(1+|z|)$$

for all $(t, z) \in [0, T] \times \mathbf{R}^r$.

Under Assumption A1, the stochastic differential equation

$$dx_t = f(t, x_t)\,dt + dw_t, \quad 0 \leqq t \leqq T, \qquad x(0) = x_0 \tag{2.3}$$

has a pathwise unique solution $\{x_t;\ 0 \leqq t \leqq T\}$ on $(\Omega, \mathbf{F}, \tilde{P}; \mathbf{F}_t)$ which is an $\mathbf{R}^r$-valued Markov process. If P_s (respectively, $P_{s,x}$) is the measure induced by the process

$\{x_t; s \leqq t \leqq T\}$ whenever x_s has the distribution of x_0 (respectively, whenever $x_s = x \in \mathbf{R}^r$), then we evidently have

$$P_s(A) = \int_{\mathbf{R}^r} P_{s,x}(A) p(x)\, dx, \quad \forall A \in \mathbf{F}, \quad \forall s \in [0, T). \tag{2.4}$$

Dependence on s is dropped when $s = 0$. Now consider a compact subset U of $\mathbf{R}^r$. We say that the U-valued "control" process $u = \{u_t; 0 \leqq t \leqq T\}$ is *admissible* if it is progressively measurable with respect to $\{\mathbf{F}_t^y; 0 \leqq t \leqq T\}$, with $\mathbf{F}_t^y \triangleq \sigma(y_s; 0 \leqq s \leqq t)$. The class of all admissible control processes is denoted by $\mathbf{A}$.

For any control process $u \in \mathbf{A}$, we define

$$L_s^t(u) \triangleq \exp\left[\int_s^t \{u_\theta'\, dw_\theta + h'(\theta, x_\theta)\, dy_\theta\} - \frac{1}{2}\int_s^t \{|u_\theta|^2 + |h(\theta, x_\theta)|^2\}\, d\theta\right], \tag{2.5}$$

$L_t(u) = L_0^t(u)$, and introduce a new measure P^u through

$$\frac{dP^u}{d\tilde{P}}(\omega) = L_T(u). \tag{2.6}$$

The Girsanov theorem (Girsanov [9], Beneš [1], [2, Appendix]) states that P^u is a probability measure on $(\Omega, \mathbf{F})$ and that

$$\begin{bmatrix} w_t^u \\ b_t \end{bmatrix} \triangleq \begin{bmatrix} w_t \\ y_t \end{bmatrix} - \begin{bmatrix} \int_0^t u_s\, ds \\ \int_0^t h(s, x_s)\, ds \end{bmatrix}, \qquad 0 \leqq t \leqq T, \tag{2.7}$$

is a Wiener process on $(\Omega, \mathbf{F}, P^u; \mathbf{F}_t)$. On this new probability space we have thus constructed a (weak) solution of the system of stochastic equations

$$dx_t = f(t, x_t)\, dt + u_t\, dt + dw_t^u, \quad 0 \leqq t \leqq T, \qquad x(0) = x_0, \tag{2.8}$$

$$dy_t = h(t, x_t)\, dt + db_t, \quad 0 \leqq t \leqq T, \qquad y_0 = 0, \tag{2.9}$$

as follows from (2.4), (2.7). Equations (2.8), (2.9) constitute our basic "system model": $X = \{x_t; 0 \leqq t \leqq T\}$ is the "signal" process and is not directly observable; the controller observes the process $Y = \{y_t; 0 \leqq t \leqq T\}$ which is a nonlinear transformation of X corrupted by additive white noise, and on the basis of his observations tries to control the evolution of the signal through the "control" process $u = \{u_t; 0 \leqq t \leqq T\}$. His objective is to minimize a cost functional of the form

$$V^u(T, p) = E^u\left\{\int_0^T \phi(u_t, x_t)\, dt + g(x_T)\right\}, \tag{2.10}$$

over all $u \in \mathbf{A}$, for suitable terminal and running cost functions $g(z)$ and $\phi(a, z)$, respectively.

For each $a \in U$, we consider the backward differential operator

$$\mathbf{L}_t^a \triangleq \tfrac{1}{2}\Delta + \{f + a\}' \cdot \nabla \tag{2.11}$$

along with its formal adjoint, the forward operator

$$\mathbf{L}_t^{*a} \triangleq \tfrac{1}{2}\Delta - \{f + a\}' \cdot \nabla - \operatorname{tr}[\nabla f]. \tag{2.12}$$

Corresponding to a control process $u \in \mathbf{A}$, we also consider the family of operators

$$\{\mathbf{L}_t^u = \mathbf{L}_t^{u_t}; 0 \leqq t \leqq T\}, \qquad \{\mathbf{L}_t^{*u} = \mathbf{L}_t^{*u_t}; 0 \leqq t \leqq T\}.$$

It is proved in § 4 that a proper unnormalized version $\rho_t(z)$ of the density of the conditional distribution $P^u(x_t \in B|\mathbf{F}_t^y)$, $B \in \text{Borel}^r$ exists and satisfies the stochastic partial differential equation

$$d\rho_t(z) = \mathbf{L}_t^{*u}\rho_t(z)\,dt + h(t,z)\rho_t(z)\,dy_t, \quad 0 \leqq t \leqq T, \qquad \rho_0(z) = p(z), \tag{2.13}$$

which is the basic equation of nonlinear filtering (see Liptser and Shiryaev [11, Chap. 8], and Zakai [20]).

Mortensen's suggestion [14] was that $\rho_t(z)$ should be viewed as the state of a new, infinite-dimensional but completely observable system governed by (2.13), and that the control problem should be formulated in the following equivalent form: to minimize

$$V^u(T,p) = E^u\left[\int_0^T \frac{(\phi(u_t,\cdot),\rho_t)}{(1,\rho_t)}\,dt + \frac{(g,\rho_T)}{(1,\rho_T)}\right] \tag{2.10'}$$

over all $u \in \mathbf{A}$, with the notational convention: $(\alpha,\beta) = \int_{\mathbf{R}^r} \alpha(z)\beta(z)\,dz$. We will refer to this as the *separated* formulation of the control problem.

Inspired by an obvious analogy with control problems in finite-dimensional state space, Mortensen also proposed the following form of a Bellman–Hamilton–Jacobi equation in function space:

$$V_1(\tau,p) = \tfrac{1}{2}V_{22}(\tau,p)[h_{T-\tau}p, h_{T-\tau}p] + \min_{a\in U}\{V_2(\tau,p)[\mathbf{L}_{T-\tau}^{*a}p] + (\phi(a,\cdot),p)\}, \tag{2.14}$$

$$V(0,p) = (g,p). \tag{2.15}$$

In the present paper we address the task of elucidating some of Mortensen's pioneering ideas. In the sequel, for reasons of notational simplicity, we confine ourselves to the case of one-dimensional observation processes: $m = 1$. Everything extends, however, to the general multidimensional case, upon replacement of products by scalar products and squares by squares of norms.

3. Summary. In § 4, we study the basic equation (2.13) of nonlinear filtering, first obtained formally by Mortensen and Duncan and rigorously by Zakai [20]. We employ the "pathwise" method of constructing a solution to the Zakai equation, first introduced by Rozovsky (see Liptser and Shiryaev [11, p. 327]), that reduces the stochastic equation to a nonstochastic partial differential equation, in which the observation sample path enters parametrically through the coefficients. The use of forward and backward equations was inspired by the work of Pardoux [15]. It is seen that the dependence of the unnormalized conditional density on the initial p is linear, whence a kind of linear fractional dependence on p for the normalized density (Theorem 4.1).

Using an observation of M. H. A. Davis, also made by the referee, one can show that the cost of a policy u is actually a *linear* function of p; this fact leads to simplifications, and is exploited in § 5 in proving that the cost $V^a(\tau,p)$ under a constant law $u_t \equiv a \in U (0 < t < T)$ satisfies a second-order equation with functional partial derivatives in p (equation (5.3)'), which is equivalent to the Hamilton–Jacobi form of the Mortensen equation, without the minimization, and free of the nonlocal terms present in (2.14).

Finally, in § 6 we establish a form of Ito's rule for functionals on density-valued processes. We use this rule in proving a verification principle for the simplified

Mortensen's equation, which states that a solution of the latter is a lower bound (in fact, the greatest lower bound) on the expected total cost during $[0, T]$ incurred under any admissible control process. This justifies statistical sufficiency of the density process for control.

We conclude in § 7 with some remarks and open questions. A principal open problem is to define the state and observation process when control depends only on the conditional density process.

4. Forward and backward partial differential equations associated with the nonlinear filtering problem. In this section we are concerned with the estimation (filtering) problem associated with the partially observable system (2.8)–(2.9), namely the problem of characterizing the conditional distribution $P^u(x_t \in B|\mathbf{F}_t^y)$, $B \in \text{Borel}^r$. It is shown that this distribution has a sufficiently smooth density on $(0, T]$, satisfying the basic equation (2.13), and that its dependence on the initial density p is linear fractional.

THEOREM 4.1 (linear fractional dependence of the conditional density on p). *With the assumptions of* § 2, *consider the probability space* $(\Omega, \mathbf{F}, P^u; \mathbf{F}_t)$ *and the solution of the system of stochastic equations* (2.8)–(2.9) *defined on it. Then for each* $s \in [0, T)$ *there exists a random function* $\{q(t, z; s, x);\ s < t \leqq T\}$, *with* $(z, x) \in \mathbf{R}^r \times \mathbf{R}^r$, *which is a fundamental solution of the Zakai equation*:

$$d_t q(t, z; s, x) = \mathbf{L}_t^{*u} q(t, z; s, x)\, dt + h(t, z) q(t, z; s, x)\, dy_t, \qquad s \leqq t \leqq T, \tag{4.1}$$

$$\lim_{t \downarrow s} q(t, z; s, x) = \delta(z - x) \tag{4.2}$$

a.s. P^u, *for any* $u \in \mathbf{A}$, *where* $\delta(\cdot)$ *is the Dirac delta function. If now* $q_t(z; x) = q(t, z; 0, x)$, *the random function*

$$p_t(z) = \frac{\int_{\mathbf{R}^r} q_t(z; x) p(x)\, dx}{\int_{\mathbf{R}^r} \int q_t(z; x) p(x)\, dx\, dz}, \qquad 0 < t \leqq T, \tag{4.3}$$

is a version of the density of the conditional distribution $P^u(x_t \in B|\mathbf{F}_t^y)$; $B \in \text{Borel}^r$, *i.e., for any bounded continuous function* $g(z): \mathbf{R}^r \to \mathbf{R}^1$ *with compact support,*

$$E^u[g(x_T)|\mathbf{F}_T^y] = \int_{\mathbf{R}^r} p_T(z) g(z)\, dz, \quad \text{a.s. } P^u \tag{4.4}$$

where E^u *denotes expectation with respect to measure* P^u.

Proof. Suppose $q(t, z; s, x)$ is a solution of (4.1), (4.2). Then it is easily checked that the random function

$$\rho_t(z) = \int_{\mathbf{R}^r} q_t(z; x) p(x)\, dx \tag{4.5}$$

satisfies the so-called *Zakai equation* [20] for the unnormalized conditional density:

$$d\rho_t(z) = \mathbf{L}_t^{*u} \rho_t(z)\, dt + h(t, z) \rho_t(z)\, dy_t, \qquad 0 < t \leqq T, \tag{4.6}$$

$$\rho_0(z) = p(z). \tag{4.7}$$

On the other hand, we have for any t, $0 \leqq t \leqq T$, the "Bayes" rule

$$E^u[g(x_t)|\mathbf{F}_t^y] = \frac{\tilde{E}[g(x_t) L_t(u)|\mathbf{F}_t^y]}{\tilde{E}[L_t(u)|\mathbf{F}_t^y]} \triangleq \frac{\pi_t(g)}{\pi_t(1)}, \quad \text{a.s. } \tilde{P} \text{ or } P^u,$$

and since the processes $\{x_t;\ 0\leqq t\leqq T\}$, $\{y_t;\ 0\leqq t\leqq T\}$ are independent under $\tilde{P}$, the path $\{x_t;\ 0\leqq t\leqq T\}$ can be "integrated out" to yield, a.s. $\tilde{P}$,

$$\pi_t(g)=\tilde{E}[g(x_t)L_t(u)|\mathbf{F}_t^y]=E[g(x_t)L_t(u)]=\int_{\mathbf{R}^r}E_z(g(x_t)L_t(u))p(z)\,dz$$

by virtue of (2.4).
Therefore, in order to prove (4.4), it suffices to establish

$$E[g(x_T)L_T(u)]=\int_{\mathbf{R}^r}\rho_T(z)g(z)\,dz,\quad \text{a.s. } \tilde{P}. \tag{4.8}$$

For a given observation sample path $Y\triangleq\{y_t;\ 0\leqq t\leqq T\}\in C[0,T]$, consider the random function $\{\psi(t,z);\ 0\leqq t\leqq T\}$ defined by

$$\psi(t,z)\triangleq\rho_t(z)\cdot\exp\{-h(t,z)y_t\},\qquad 0\leqq t\leqq T. \tag{4.9}$$

An application of Ito's rule, in conjunction with (4.6), (4.7), yields the (nonstochastic) *forward parabolic equation* for $\psi(t,z)$:

$$\frac{\partial}{\partial t}\psi=\overset{v}{\mathbf{L}}{}_t^{*u}\psi+e\cdot\psi,\qquad (t,z)\in(0,T]\times\mathbf{R}^r, \tag{4.10}$$

$$\psi(0,z)=p(z),\qquad z\in\mathbf{R}^r, \tag{4.11}$$

where

$$\overset{v}{\mathbf{L}}{}_t^{*u}\triangleq\tfrac{1}{2}\Delta+\{y_t\cdot\nabla h-(f+u_t)\}'\nabla+\{\tfrac{1}{2}y_t\Delta h-tr[\nabla f]\}, \tag{4.12}$$

$$e(t,z)\triangleq\tfrac{1}{2}y_t^2|\nabla h|^2-y_t(f+u_t)'\cdot\nabla h-\tfrac{1}{2}h^2-y_th_1. \tag{4.13}$$

Equation (4.10) is parabolic with constant diffusion, linearly growing drift and quadratically growing potential terms, in view of Assumptions A1–A2, with the sample path Y entering parametrically through the coefficients. Consequently, for any $y\in C[0,T]$ (not just in a subset of the full Wiener measure), equation (4.10) has a unique, positive fundamental solution $\Gamma(t,z;s,x)$ satisfying

$$|D^m\Gamma(t,z;s,x)|<c(t-s)^{-(r+|m|)/2}\exp\left[-\mu\frac{|z-x|^2}{t-s}\right],\qquad 0\leqq|m|\leqq 2, \tag{4.14}$$

for some positive constants c, μ depending on Y, while the Cauchy problem (4.10)–(4.11) has the classical $(C_{t,z}^{1,2}((0,T]\times\mathbf{R}^r))$ solution

$$\psi(t,z)=\int_{\mathbf{R}^r}\Gamma(t,z;0,x)p(x)\,dx, \tag{4.15}$$

which is positive (by the maximum principle) and unique among those satisfying an exponential growth condition in the space variable; see Eidel'man [4], Friedman [7]. It is now easy to verify that

$$q(t,z;x,s)\triangleq\exp[y_th(t,z)-y_sh(s,x)]\cdot\Gamma(t,z;s,x),\qquad 0\leqq s<t<T, \tag{4.16}$$

solves (4.1)–(4.2), a.s. P^u, and that $\rho_t(z)=\psi(t,z)\cdot\exp\{h(t,z)y_t\}$, $0\leqq t\leqq T$, is the a.s.-P^u unique solution of the Zakai equation (4.6) under (4.7).

Consider now the adjoint equation to (4.10), namely

$$\frac{\partial \zeta}{\partial t} + e \cdot \zeta + \overset{v}{\mathbf{L}}{}_t^u \zeta = 0, \qquad (t, z) \in [0, T) \times \mathbf{R}^r, \tag{4.17}$$

$$\zeta(T, z) = g(z) \exp(y_T h(T, z)), \qquad z \in \mathbf{R}^r, \tag{4.18}$$

with $g(\cdot)$ as in (4.4) and

$$\overset{v}{\mathbf{L}}{}_t^u \triangleq (\overset{v}{\mathbf{L}}{}_t^{*u})^* = \tfrac{1}{2}\Delta + \{f + u_t - y_t \cdot \nabla h\}' \cdot \nabla - \tfrac{1}{2} y_t \Delta h. \tag{4.19}$$

Equation (4.17) is a (nonstochastic) *backward parabolic equation*, with the observation sample path y again entering parametrically through the coefficients. We propose to show that, if the random function $v(t, z)$ is defined through

$$v(t, z) \triangleq E_{t,z}[g(x_T) L_t^T(u)], \tag{4.20}$$

then

$$\zeta(t, z) \triangleq v(t, z) \cdot \exp[y_t h(t, z)] \tag{4.21}$$

solves the Cauchy problem (4.17), (4.18). Indeed, with $\mathbf{L}_t^0$ the backward operator associated with (2.3):

$$\mathbf{L}_t^0 = \tfrac{1}{2}\Delta + f'(t, z) \cdot \nabla,$$

an application of Ito's rule gives (a.s. $P_{t,z}$)

$$y_t h(t, z) - y_T h(T, x_T) \equiv \int_t^T \{y_s \mathbf{L}_s^0 h(s, x_s) + y_s h_1(s, x_s)\}\, ds + y_s (\nabla h(s, x_s))'\, dw_s + h(s, x_s)\, dy_s,$$

and substitution into (4.21):

$$\zeta(t, z) = E_{t,z}\left[g(x_T) \exp(y_t h(T, x_T)) \cdot \exp\left\{ \int_t^T l(s, x_s)\, ds \right\} \cdot \overset{v}{\mathbf{L}}{}_t^T(u) \right] \tag{4.22}$$

where:

$$\overset{v}{\mathbf{L}}{}_t^T(u) \triangleq \exp\left[\int_t^T \{u_s - y_s \cdot \nabla h(s, x_s)\}'\, dw_s - \frac{1}{2} \int_t^T |u_s - y_s \cdot \nabla h(s, x_s)|^2\, ds \right],$$

$$l(t, z) \triangleq \tfrac{1}{2} y_t^2 |\nabla h(t, z)|^2 - y_t [\mathbf{L}_t^0 h(t, z) + h_1(t, z) + u_t' \cdot \nabla h(t, z)] - \tfrac{1}{2} h^2(t, z).$$

In (4.22) we recognize the familiar Feynman–Kac formula [8] for the solution of the backward partial differential equation

$$\frac{\partial}{\partial t} \zeta + \tilde{\mathbf{L}}_t^u \zeta + l\zeta = 0, \qquad (t, z) \in [0, T) \times \mathbf{R}^r,$$

$$\zeta(T, z) = g(z) \exp[y_T h(T, z)], \qquad z \in \mathbf{R}^r,$$

where $\tilde{\mathbf{L}}_t^u \triangleq \tfrac{1}{2}\Delta + \{f(t, z) + u_t - y_t \cdot \nabla h(t, z)\}' \cdot \nabla$; noting $\tilde{\mathbf{L}}_t^u + l(t, z) = \mathbf{L}_t^u + e(t, z)$, we verify that $\zeta(t, z)$, defined through (4.20), (4.21), satisfies the Cauchy problem (4.17), (4.18).

Now we check that the process $A_t = \int_{\mathbf{R}^r} \zeta(t, z) \psi(t, z)\, ds$ is constant on $[0, T]$, because by virtue of equations (4.11) and (4.17),

$$\frac{d}{dt} \int_{\mathbf{R}^r} \zeta(t, z) \psi(t, z)\, ds = \int_{\mathbf{R}^r} \{-\psi[\mathbf{L}_t^u \zeta + e\zeta] + \zeta[\mathbf{L}_t^{*u} \psi + e\psi]\}\, dz = 0$$

for any $0<t\leqq T$. But $A_0=\int_{\mathbf{R}^r}E_z[g(x_T)L_T(u)]p(z)\,dz=E[g(x_T)L_T(x)]$ and $A_T=\int_{\mathbf{R}^r}g(z)\rho_T(z)\,dz$, so (4.8) is actually true for *all* $y\in C[0,T]$, not merely on a subset of full Wiener measure.

PROPOSITION 4.1 (stochastic PDE for the normalized conditional density (Stratonovich [18], Kushner [10])). *The random function* $p_t(z)$ *introduced in* (4.3) *satisfies the stochastic partial differential equation a.s.* P^u:

$$dp_t(z)=\mathbf{L}_t^{*u}p_t(z)\,dt+[h-(h,p_t)]p_t(z)\,d\nu_t^u,\qquad 0<t\leqq T, \tag{4.23}$$

$$p_0(z)=p(z), \tag{4.24}$$

where $\nu_t^u\triangleq y_t-\int_0^t E^u[h(s,x_s)|\mathbf{F}_s^y]\,ds=y_t-\int_0^t(h,p_s)\,ds$, $0\leqq t\leqq T$, *is the "innovations process" and is Wiener under* P^u. *In addition, for all* $y\in C[0,T]$:

$$p_t(z)\in\mathbf{E}_k,\qquad \mathbf{L}_t^{u*}p_t(z),[h-(h,p_t)]p_t(z)\in\mathbf{E}_{k-1} \tag{4.25}$$

for all $0<t\leqq T$.

Proof. Equation (4.23) follows from (4.6), by an application of Ito's rule to $p_t(z)=\rho_t(z)\cdot(\int_{\mathbf{R}^r}\rho_t(z)\,dz)^{-1}$. (4.25) is a consequence of Assumptions A1–A2 and (4.24). Q.E.D.

Remark. Because of (4.25), the function $g(\cdot)\colon\mathbf{R}^r\times\boldsymbol{R}^1$ in Theorem 4.1 can be taken to be continuous and to satisfy the growth condition

$$|g(z)|\leqq K(1+|z|^l),\quad \text{all } z\in\mathbf{R}^r,\quad l\leqq k. \tag{4.26}$$

PROPOSITION 4.2. *With* $g(\cdot)$ *as above, the random function* $v(t,z)$ *defined by* (4.20):

(i) *admits the representation*

$$v(t,z)=\int_{\mathbf{R}^r}q(T,\xi;t,z)g(\xi)\,d\xi,\qquad (t,z)\in[0,T]\times\mathbf{R}^r, \tag{4.27}$$

for all $y\in C[0,T]$, *with* $q(t,z;s,x)$ *the fundamental solution of the Zakai equation*;

(ii) *satisfies the backward stochastic partial differential equation*

$$d_tv+\mathbf{L}_t^uv\,dt+hv\,dy_t=0,\qquad 0\leqq t<T, \tag{4.28}$$

$$v(T,z)=g(z) \tag{4.29}$$

a.s. $\tilde{P}$, *where the stochastic differential appearing in* (4.28) *has to be interpreted as the differential of a backward Ito integral, the process* $\{\tilde{y}_t=y_t-y_T;\ 0\leqq t\leqq T\}$ *being a backward Wiener process.*

Proof. Recall the fundamental solution $\Gamma(t,z;s,x)$, $0\leqq s<t\leqq T$, of the forward PDE (4.10); the adjoint equation (4.17) has the unique solution

$$\zeta(t,z)=\int_{\mathbf{R}^r}\Gamma(T,\xi;t,z)g(\xi)\exp[y_Th(T,\xi)]\,d\xi$$

and so (4.27) follows by virtue of definitions (4.16), (4.21).

On the other hand, the backward stochastic differentiation rule (McKean [13, p. 35]) gives

$$d(\exp[-y_th(t,z)])=-\exp[-y_th(t,z)]\cdot[\tfrac{1}{2}h^2(t,z)+y_th_1(t,z)\}\,dt+h(t,z)\,dy_t],$$

and when applied to $v(t,z)=\zeta(t,z)\cdot\exp[-y_th(t,z)]$ it eventually yields (4.28), after a bit of simple calculus.

5. Hamilton–Jacobi equations with functional partial derivatives associated with the nonlinear filtering problem. We now return to the cost functional $V^u(T,p)$ introduced in (2.10) and impose the following condition on the cost functions $g(z)$, $\phi(u,z)$:

Assumption A3. $g(z):\mathbf{R}^r\to\mathbf{R}^+$ and $\phi(a,x):U\times\mathbf{R}^r\to\mathbf{R}^+$ are continuous on their domains of definition and satisfy, for some positive numbers K and $l\leqq k-1$, the growth condition

$$g(z)+\phi(a,z)\leqq K(1+|z|^l)\quad\forall(a,z)\in U\times\mathbf{R}^r.$$

It can be checked easily from (4.8) that, with $0<t\leqq T$,

$$E^u g(x_t)=\tilde{E}\left[L_T(u)\frac{\pi_t(g)}{\pi_t(1)}\right]=\tilde{E}\left[L_t(u)\frac{\pi_t(g)}{\pi_t(1)}\right]=\tilde{E}\left[\tilde{E}\{L_t(u)|\mathbf{F}_t^y\}\frac{\pi_t(g)}{\pi_t(1)}\right]=\tilde{E}[\pi_t(g)].$$

Consequently, the cost functional in (2.10) depends linearly on the initial density p, when we integrate with respect to measure $\tilde{P}$:

$$V^u(T,p)=\tilde{E}\left[\int_0^T\left(\int\int_{\mathbf{R}^r}\phi(u_t,z)q_t(z;x)p(x)\,dx\,dz\right)dt\right.$$
$$\left.+\int\int_{\mathbf{R}^r}g(z)q_T(z;x)p(x)\,dx\,dz\right]. \tag{5.1}$$

Use of this form for the "value" of u leads to a form of Mortensen's equation free of nonlocal terms. Our aim in this section is to show directly that the functional $V^a(\tau,p):[0,T]\times\mathbf{E}_k\to\mathbf{R}^1$, representing the expected total cost incurred during the time interval $[T-\tau,T]$ under the initial condition $p_{T-\tau}(z)=p(z)\in\mathbf{E}_k$ and the *constant* control $u_t\equiv a\in U$ for all $0\leqq t\leqq T$, satisfies the Hamilton–Jacobi form (5.3′) below of the Mortensen equation, without the minimization.

We start with the case $\phi\equiv 0$:

THEOREM 5.1 (Hamiltonian–Jacobi equation for $V^a(\tau,p)(\phi=0)$). *Under the assumptions of* § 2, *consider the probability space* $(\Omega,\mathbf{F},P^a;\mathbf{F}_t)$ *corresponding to a constant control process* $u_t\equiv a\in U$, $T-\tau\leqq t\leqq T$, *and the solution to the system of equations* (2.8)–(2.9) *defined on it for the time interval* $[T-\tau,T]$. *Suppose also that* $x_{T-\tau}$ *has a distribution with density* $p\in\mathbf{E}_k$, *and define a continuous function* $V^a(\tau,p)$: $[0,T]\times\mathbf{E}_k\to\mathbf{R}^1$ *by setting*

$$V^a(\tau,p)\triangleq E^a_{T-\tau,p}g(x_T)=\tilde{E}\int\int_{\mathbf{R}^r}g(z)q(T,z;T-\tau,x)p(x)\,dx\,dz \tag{5.2}$$

where $g(\cdot)$ *satisfies Assumption* A3. *Then* $V^a(\tau,p)$ *is once continuously differentiable with respect to* τ, *twice continuously Fréchet differentiable with respect to* p, *and satisfies the following Cauchy problem for the Hamiltonian–Jacobi equation*:

$$V_1^a(\tau,p)=\tfrac{1}{2}V_{22}^a(\tau,p)[h_{T-\tau}p,h_{T-\tau}p]+(\mathbf{L}^a_{T-\tau}V_2^a(\tau,p),p),\qquad(\tau,p)\in(0,T]\times\mathbf{E}_k, \tag{5.3}$$

$$V^a(0,p)=(g,p),\qquad p\in\mathbf{E}_k. \tag{5.4}$$

Proof. Introducing the notation $N_t(x)\triangleq\int_{\mathbf{R}^r}g(z)q(T,z;t,x)\,dz$, $s\triangleq T-\tau$, we write $V^a(\tau,p)$ in the form $V^a(\tau,p)=\tilde{E}(N_s,p)$. Then we define the linear functional

$$V_2^a(\tau,p)[\eta]=\int_{\mathbf{R}^r}V_2^a(\tau,p)(x)\eta(x)\,dx,\qquad\eta(\cdot)\in\mathbf{E}_{k-1}, \tag{5.5}$$

with

$$V_2^a(\tau, p)(x) = \tilde{E}N_s(x), \tag{5.6}$$

and the bilinear form

$$V_{22}^a(\tau, p)[\eta, \theta] = \int\int_{\mathbf{R}^r} V_{22}^a(\tau, p)(x, x') \cdot \eta(x)\theta(x')\, dx\, dx';\ \eta(\cdot), \theta(\cdot) \in \mathbf{E}_{k-1}, \tag{5.7}$$

with

$$V_{22}^a(\tau, p)(x, x') \equiv 0 \tag{5.8}$$

as the first and second, respectively, Fréchet derivatives of the function $V^a(\tau, p)$ with respect to p; see, for the definitions, Lusternik and Sobolev [12].

Let us now evaluate the left-hand side of (5.3). According to Proposition 4.2, $N_t(x)$ satisfies the backward stochastic differential equation, a.s. $\tilde{P}$

$$d\phi + \mathbf{L}_t^a\phi\, dt + h\phi\, dy_t = 0, \qquad T - \tau \leqq t < T,$$

subject to the terminal condition $N_T(x) = g(x)$.

Consequently

$$(N_s, p) = (g, p) + \int_{T-\tau}^{T} (\mathbf{L}_t^a N_t, p)\, dt + \int_{T-\tau}^{T} (h_t N_t, p)\, dy_t,$$

and by taking expectations with respect to $\tilde{P}$, one gets

$$V^a(\tau, p) = (g, p) + \tilde{E}\int_{T-\tau}^{T} (\mathbf{L}_t^a N_t, p)\, dt.$$

The initial condition (5.4) is obviously satisfied; on the other hand, differentiation with respect to τ gives

$$V_1^a(\tau, p) = \tilde{E}(\mathbf{L}_s^a N_s, p).$$

Equation (5.3) follows by virtue of relations (5.6), (5.8) and the fact that, for a constant control $u_t = a \in U$, $T - \tau \leqq t \leqq T$,

$$(\mathbf{L}_s^a V_2(\tau, p), p) = \int_{\mathbf{R}^r} \mathbf{L}_s^a V_2(\tau, p)(x) \cdot p(x)\, dx = \tilde{E}(\mathbf{L}_s^a N_s, p). \qquad \text{Q.E.D.}$$

Similarly, one can do the general case:

THEOREM 5.2 (Hamilton–Jacobi equation for $V^a(\tau, p)$). *Under the same assumptions as in Theorem* 5.1, *the cost functional*

$$V^a(\tau, p) = E_{T-\tau, p}^a\left[\int_{T-\tau}^{T} \phi(a, x_t)\, dt + g(x_T)\right], \tag{5.2'}$$

with the functions $g(z)$, $\phi(u, z)$ *satisfying Assumption* A3, *solves the Hamilton–Jacobi equation*

$$\begin{aligned} V_1^a(\tau, p) = {} & \tfrac{1}{2}V_{22}(\tau, p)[h_{T-\tau}p, h_{T-\tau}p] + (\mathbf{L}_{T-\tau}^a V_2(\tau, p), p) \\ & + (\phi(a, \cdot), p), \qquad (\tau, p) \in (0, T] \times \mathbf{E}_k, \end{aligned} \tag{5.3'}$$

$$V^a(0, p) = (g, p), \qquad p \in \mathbf{E}_k, \tag{5.4'}$$

for any fixed $a \in U$.

6. A verification theorem for Mortensen's equation. In this section we establish a verification principle for Mortensen's equation, saying essentially that a solution to this equation is a lower bound on the achievable expected cost $V^u(\tau, p)$ under any admissible control process $u \in A$ (Theorem 6.2). To this end, we need a version of Ito's rule valid for functionals on density-valued processes, similar to the classical Ito change-of-variable formula (see e.g. Skorokhod [17]).

DEFINITION 6.1. Consider a probability space $(\Omega, \mathbf{F}, P; \mathbf{F}_t)$ and an $\mathbf{E}_k$-valued stochastic process $\{\xi_t(z); 0 \leqq t \leqq T\}$ adapted to the increasing family $\{\mathbf{F}_t; 0 \leqq t \leqq T\}$ of σ-fields, with $l \geqq 0$. If

$$E\int_0^T \left(\int_{\mathbf{R}^r} (1+|z|^l)|\xi_t(z)|\, dz \right)^j dt < \infty, \tag{6.1}$$

we say that $\xi_t(z)$ belongs to $M_{l,j}[\mathbf{F}_t]$.

THEOREM 6.1 (Ito's rule for functionals on density-valued processes). *Let the $\mathbf{E}_k$-valued process $p_t(z)$ satisfy the relation*

$$dp_t(z) = a_t(z)\, dt + m_t(z)\, dw_t, \qquad 0 \leqq t \leqq T, \tag{6.2}$$

where $\{w_t; 0 \leqq t \leqq T\}$ is a Wiener process on $(\Omega, \mathbf{F}, P; \mathbf{F}_t)$ and the $\mathbf{E}_l$-valued processes $a_t(z)$, $m_t(z)$ belong to

$$M_{l,1}[\mathbf{F}_t] \cap M_{l,2}[\mathbf{F}_t] \quad \text{and} \quad M_{l,2}[\mathbf{F}_t] \cap M_{l,4}[\mathbf{F}_t],$$

respectively.

Consider also a function $V(\tau, p): [0, T] \times \mathbf{E}_k \to \mathbf{R}^1$, possessing continuous first derivative $V_1(\tau, p)$ in τ and first and second Fréchet derivatives $V_2(\tau, p)[\cdot]$ and $V_{22}(\tau, p)[\cdot, \cdot]$ with respect to p in the form of a linear functional and a bilinear form, respectively:

$$V_2(\tau, p)[\eta] = \int_{\mathbf{R}^r} V_2(\tau, p)(x)\eta(x)\, dx, \qquad \eta(\cdot) \in \mathbf{E}_l,$$

$$V_{22}(\tau, p)[\eta, \theta] = \int_{\mathbf{R}^r} V_{22}(\tau, p)(x, x')\eta(x)\theta(x')\, dx\, dx', \qquad \eta(\cdot), \theta(\cdot) \in \mathbf{E}_l,$$

with kernels $V(\tau, p)(x)$, $V_{22}(\tau, p)(x, x')$ continuous in their arguments and satisfying the following conditions:

$$|V_2(\tau, p)(x)| \leqq \nu_1(\tau, \|p\|_l)(1+|x|^l), \tag{6.3}$$

$$|V_{22}(\tau, p)(x, x')| \leqq \nu_2(\tau, \|p\|_l)(1+|x|^l)(1+|x'|^l) \tag{6.4}$$

where ν_1 and ν_2 are continuous functions on $[0, T] \times \mathbf{R}^+$.

Then the process $\{V(t, p_t); 0 \leqq t \leqq T\}$ admits the stochastic differential:

$$\begin{aligned} dV(t, p_t) = \{&V_1(t, p_t) + V_2(t, p_t)[a_t] + \tfrac{1}{2}V_{22}(t, p_t)[m_t, m_t]\}\, dt \\ &+ V_2(t, p_t)[m_t] \cdot dw_t, \qquad 0 \leqq t \leqq T. \end{aligned} \tag{6.5}$$

Proof. From the assumptions on $a_t(z)$, $m_t(z)$ it is seen that $\|p_t\|_l$ is finite for all $t \in [0, T]$, a.s. P.

In order to prove (6.5) we have to show that for any $0 < t \leqq T$.

$$\begin{aligned} V(t, p_t) = V(0, p_0) &+ \int_0^t \{V_1(s, p_s) + V_2(s, p_s)[a_s] + \tfrac{1}{2}\, V_{22}(s, p_s)[m_s, m_s]\}\, ds \\ &+ \int_0^t V_2(s, p_s)[m_s] \cdot dw_s, \quad \text{a.s. } P, \end{aligned} \tag{6.6}$$

where $p_0 \in \mathbf{E}_l$. We take a partition $0 = \sigma_1 < \sigma_2 < \cdots < \sigma_{N+1} = t$ of the interval $[0, t]$, denote its mesh by $|\sigma| = \max_{1 \leqq n \leqq N} |\sigma_{n+1} - \sigma_n|$, and write

$$
\begin{aligned}
V(t, p_t) - V(0, p_0) &= \sum_{n=1}^{N} \{V(\sigma_{n+1}, p_{\sigma_{n+1}}) - V(\sigma_n, p_{\sigma_n})\} \\
&= \sum_{n=1}^{N} \{V(\sigma_{n+1}, p_{\sigma_{n+1}}) - V(\sigma_n, p_{\sigma_{n+1}})\} + \sum_{n=1}^{N} \{V(\sigma_n, p_{\sigma_{n+1}}) - V(\sigma_n, p_{\sigma_n})\} \\
&= \sum_{n=1}^{N} V_1(\sigma_n + \bar{\nu}_n(\sigma_{n+1} - \sigma_n), p_{\sigma_{n+1}}) \cdot (\sigma_{n+1} - \sigma_n) \qquad (6.7) \\
&\quad + \sum_{n=1}^{N} V_2(\sigma_n, p_{\sigma_n})[p_{\sigma_{n+1}} - p_{\sigma_n}] \\
&\quad + \frac{1}{2} \sum_{n=1}^{N} V_{22}(\sigma_n, p_{\sigma_n} + \nu_n(p_{\sigma_{n+1}} - p_{\sigma_n}))[p_{\sigma_{n+1}} - p_{\sigma_n}, p_{\sigma_{n+1}} - p_{\sigma_n}]
\end{aligned}
$$

by the Taylor–Volterra formula (Lusternik–Sobolev [12]), where ν_n and $\bar{\nu}_n$ are numbers in (0, 1). From the continuity of $V_1(t, p)$ it follows that the first sum in the last member of relation (6.7) converges a.s. to $\int_0^t V_1(s, p_s)\, ds$ as $|\sigma| \to 0$. As for the second term, we have

$$
\begin{aligned}
\sum_{n=1}^{N} V_2(\sigma_n, p_{\sigma_n})[p_{\sigma_{n+1}} - p_{\sigma_n}] &= \sum_{n=1}^{N} \int_{\mathbf{R}^r} V_2(\sigma_n, p_{\sigma_n})(z) \left(\int_{\sigma_n}^{\sigma_{n+1}} a_s(z)\, ds \right) dz \\
&\quad + \sum_{n=1}^{N} \int_{\mathbf{R}^r} V_2(\sigma_n, p_{\sigma_n})(z) \left(\int_{\sigma_n}^{\sigma_{n+1}} m_s(z)\, dw_s \right) dz \\
&= I_1 + I_2.
\end{aligned}
$$

By Fubini's theorem, since $\int_0^T \int_{\mathbf{R}^r} (1 + |z|^l)|a_s(z)|\, dz\, ds < \infty$, a.s. P by assumption, we have

$$
I_1 = \int_0^t f_N(s)\, ds,
$$

where

$$
f_N(s) \triangleq V_2(\sigma_n, p_{\sigma_n})[a_s] \cdot 1_{[\sigma_n, \sigma_{n+1})}(s).
$$

Now

$$
f_N(s) \xrightarrow[|\sigma| \downarrow 0]{\text{a.s.}} f(s) \triangleq V_2(s, p_s)[a_s]
$$

for almost all s in $[0, t]$, by continuity of $V_2(\tau, p)$ in its arguments and of $p_t(z)$ in t. On the other hand, by assumption,

$$
\int_0^T |f_N(s)|\, ds, \int_0^T |f(s)|\, ds \leqq \sup_{0 \leqq s \leqq T} \nu_1(s, \|p_s\|_l),
$$

and

$$
\int_0^T \int_{\mathbf{R}^r} (1 + |z|^l)|a_s(z)|\, dz\, ds < \infty
$$

a.s. P, so by the dominated convergence theorem,

$$
I_1 \xrightarrow[|\sigma| \downarrow 0]{\text{a.s.}} \int_0^t V_2(s, p_s)[a_s]\, ds.
$$

Similarly, by a Fubini-type theorem for stochastic integrals (Szpirglas [19]), we can write

$$I_2 = \int_0^t \tilde{f}_N(s)\, dw_s,$$

where $\tilde{f}_N(s) \triangleq V_2(\sigma_n, p_{\sigma_n})[m_s] \cdot 1_{[\sigma_n, \sigma_{n+1})}(s)$, since, by assumption,

$$E\int_0^T \left(\int_{\mathbf{R}^r} (1+|z|^l)|m_s(z)|\, dz\right)^2 ds < \infty.$$

Again, $\tilde{f}_N(s) \xrightarrow[|\sigma|\downarrow 0]{\text{a.s.}} \tilde{f}(s) \triangleq V_2(s, p_s)[m_s]$, for almost all s in $[0, t]$, and

$$\int_0^T |\tilde{f}_N(s)|^2\, ds, \int_0^T |\tilde{f}(s)|^2\, ds \leqq \sup_{0\leqq s\leqq T} \nu_1^2(s, \|p_s\|_l) \cdot \int_0^T \left(\int_{\mathbf{R}^r} (1+|z|^l)|m_s(z)|\, ds\right)^2 ds.$$

Therefore, by the Ito dominated convergence theorem (Skorokhod [17, p. 19], we have

$$I_2 \xrightarrow[|\sigma|\downarrow 0]{P} \int_0^t V_2(s, p_s)[m_s]\, dw_s.$$

Finally, the third term in (6.7) can be decomposed in the form

$$\tfrac{1}{2}(I_3 + I_4) + I_5,$$

with

$$I_3 = \sum_{n=1}^N \int\int_{\mathbf{R}^r} V_{22}(\sigma_n, p_{\sigma_n} + \nu_n(p_{\sigma_{n+1}} - p_{\sigma_n}))(z, z') \cdot \left(\int_{\sigma_n}^{\sigma_{n+1}} a_s(z)\, ds\right)\left(\int_{\sigma_n}^{\sigma_{n+1}} a_s(z')\, ds\right) dz\, dz',$$

$$I_4 = \sum_{n=1}^N \int\int_{\mathbf{R}^r} V_{22}(\sigma_n, p_{\sigma_n} + \nu_n(p_{\sigma_{n+1}} - p_{\sigma_n}))(z, z') \cdot \left(\int_{\sigma_n}^{\sigma_{n+1}} m_s(z)\, dw_s\right)\left(\int_{\sigma_n}^{\sigma_{n+1}} m_s(z')\, dw_s\right) dz\, dz',$$

$$I_5 = \sum_{n=1}^N \int\int_{\mathbf{R}^r} V_{22}(\sigma_n, p_{\sigma_n} + \nu_n(p_{\sigma_{n+1}} - p_{\sigma_n}))(z, z') \cdot \left(\int_{\sigma_n}^{\sigma_{n+1}} a_s(z)\, ds\right)\left(\int_{\sigma_n}^{\sigma_{n+1}} m_s(z')\, dw_s\right) dz\, dz'.$$

We see that

$$|I_3| \leqq \max_{0\leqq s\leqq t} \nu_2(s, \|p_s\|_l) \cdot \sum_{n=1}^N \left(\int_{\mathbf{R}^r} (1+|z|^l)\left(\int_{\sigma_n}^{\sigma_{n+1}} |a_s(z)|\, ds\right) dz\right)^2$$

$$\leqq |\sigma| \cdot \max_{0\leqq s\leqq t} \nu_2(s, \|p_s\|_l) \cdot \int_0^t \left(\int_{\mathbf{R}^r} (1+|z|^l)|a_s(z)|\, dz\right)^2 ds \xrightarrow[|\sigma|\downarrow 0]{P} 0,$$

by virtue of the Fubini theorem, the Cauchy inequality and the assumption

$$E\int_0^T \left(\int_{\mathbf{R}^r} (1+|z|^l)|a_s(z)|\, dz\right)^2 ds < \infty.$$

Similarly,

$$|I_5| \leqq \max_{0 \leqq s \leqq t} \nu_2(s, \|p_s\|_l) \sum_{n=1}^{N} \left\{ \int_{\mathbf{R}^r} (1+|z|^l) \left(\int_{\sigma_n}^{\sigma_{n+1}} |a_s(z)|\, ds \right) dz \right\}$$
$$\cdot \left\{ \int_{\mathbf{R}^r} (1+|z|^l) \left| \int_{\sigma_n}^{\sigma_{n+1}} m_s(z)\, dw_s \right| dz \right\},$$

and so

$$|I_5|^2 \leqq |\sigma| \cdot \max_{0 \leqq s \leqq t} \nu_2^2(s, \|p_s\|_l) \cdot \int_0^t \left(\int_{\mathbf{R}^r} (1+|z|^l) |a_s(z)|\, dz \right)^2 ds \cdot \mathbf{J} \xrightarrow[|\sigma| \downarrow 0]{P} 0,$$

because the term

$$\mathbf{J} = \sum_{n=1}^{N} \left\{ \int_{\mathbf{R}^r} (1+|z|^l) \left| \int_{\sigma_n}^{\sigma_{n+1}} m_s(z)\, dw_s \right| dz \right\}^2$$

is bounded in probability, uniformly over all subdivisions of $[0, t]$ by virtue of the assumption $m_t(z) \in M_{l,2}[\mathbf{F}_t]$.

Therefore, in order to complete the proof of Theorem 6.1, it remains to show that

$$I_4 \xrightarrow[|\sigma| \downarrow 0]{P} \int_0^t V_{22}(s, p_s)[m_s, m_s]\, ds.$$

However, as is easily seen,

$$\sum_{n=1}^{N} \int \int_{\mathbf{R}^r} V_{22}(\sigma_n, p_{\sigma_n} + \nu_n(p_{\sigma_{n+1}} - p_{\sigma_n}))(z, z') \left\{ \int_{\sigma_n}^{\sigma_{n+1}} m_s(z) m_s(z')\, ds \right\} dz\, dz'$$
$$\xrightarrow[|\sigma| \downarrow 0]{P} \int_0^t V_{22}(s, p_s)[m_s, m_s]\, ds,$$

and

$$\max_{1 \leqq n \leqq N} |V_{22}(\sigma_n, p_{\sigma_n} + \nu_n(p_{\sigma_{n+1}} - p_{\sigma_n}))(z, z') - V_{22}(\sigma_n, p_{\sigma_n})(z, z')| \xrightarrow[|\sigma| \downarrow 0]{P} 0,$$

uniformly on compact subsets of $\mathbf{R}^{2d}$, by continuity of $V_{22}(\tau, p)$ in p and of $p_s(\cdot)$ in s. Therefore, it suffices to establish the convergence

$$\sum_{n=1}^{N} \mathbf{J}_n \xrightarrow[|\sigma| \downarrow 0]{P} 0, \tag{6.8}$$

$$\mathbf{J}_n = \int \int_{\mathbf{R}^r} V_{22}(\sigma_n, p_{\sigma_n})(z, z') \left\{ \int_{\sigma_n}^{\sigma_{n+1}} m_s(z)\, dw_s \int_{\sigma_n}^{\sigma_{n+1}} m_s(z')\, dw_s \right.$$
$$\left. - \int_{\sigma_n}^{\sigma_{n+1}} m_s(z) m_s(z')\, ds \right\} dz\, dz'.$$

We introduce the process $\chi(t) \triangleq 1_{\{\sup_{s \leq t} \|p_s\|_l \leqq c\}}$, $c > 0$ arbitrary but fixed, and observe that

$$\left(\sum_{n=1}^{N} \mathbf{J}_n > \varepsilon \right) \leqq P\left[\sum_{n=1}^{N} \chi(\sigma_n) \mathbf{J}_n > \varepsilon \right] + P\left[\sup_{0 \leqq t \leqq T} \|p_t\|_l > c \right]. \tag{6.9}$$

The sum $\sum_{n=1}^{N} \chi(\sigma_n) \mathbf{J}_n$ converges in probability to zero, as $|\sigma| \to 0$; in fact, with $m < n$,

$$E[\chi(\sigma_m) \mathbf{J}_m \cdot \chi(\sigma_n) \mathbf{J}_n] = E[\chi(\sigma_m) \chi(\sigma_n) \mathbf{J}_m \cdot E(\mathbf{J}_n | \mathbf{F}_{\sigma_n})] = 0$$

since

$$E(\mathbf{J}_n|\mathbf{F}_{\sigma_n}) = \int\int_{\mathbf{R}^r} V_{22}(\sigma_n, p_{\sigma_n})(z, z') \cdot E\left\{\int_{\sigma_n}^{\sigma_{n+1}} m_s(z)\, dw_s \cdot \int_{\sigma_n}^{\sigma_{n+1}} m_s(z')\, dw_s - \int_{\sigma_n}^{\sigma_{n+1}} m_s(z) m_s(z')\, ds | \mathbf{F}_{\sigma_n}\right\} dz\, dz'$$

$$= 0 \quad \text{a.s. } P,$$

and therefore $E(\sum_{n=1}^{N} \chi(\sigma_n)\mathbf{J}_n)^2 = \sum_{n=1}^{N} E[\chi(\sigma_n)\mathbf{J}_n^2]$. By repeatedly applying Fubini's theorem we write $\mathbf{J}_n$ in the form

$$\mathbf{J}_n = \int_{\sigma_n}^{\sigma_{n+1}} \int_{\sigma_n}^{\sigma_{n+1}} g_N(s, \theta)\, dw_\theta\, dw_s - \int_{\sigma_n}^{\sigma_{n+1}} g_N(s, s)\, ds,$$

where

$$g_N(s, \theta) \triangleq V_{22}(\sigma_n, p_{\sigma_n})[m_s, m_\theta] \cdot 1_{[\sigma_n, \sigma_{n+1})}(s) \cdot 1_{[\sigma_n, \sigma_{n+1})}(\theta),$$

and it is easily checked that, with $L \triangleq \sup_{0<\tau<T, |x| \leqq c} \nu_2(\tau, x)$,

$$\chi(\sigma_n)|g_N(s, \theta)| \leqq L\left(\int_{\mathbf{R}^r} (1 + |z|^l)|m_s(z)|\, dz\right) \cdot \left(\int_{\mathbf{R}^r} (1 + |z|^l)|m_\theta(z)|\, dz\right) 1_{[\sigma_n, \sigma_{n+1})}(s) \cdot 1_{[\sigma_n, \sigma_{n+1})}(\theta).$$

So:

$$E[\chi(\sigma_n)\mathbf{J}_n^2] \leqq 2\int_{\sigma_n}^{\sigma_{n+1}} \int_{\sigma_n}^{\sigma_{n+1}} E[\chi(\sigma_n) g_N^2(s, \theta)]\, d\theta\, ds + 2E\left(\int_{\sigma_n}^{\sigma_{n+1}} \chi(\sigma_n) g_N(s, s)\, ds\right)^2$$

$$\leqq 2L^2 \cdot E\left\{\int_{\sigma_n}^{\sigma_{n+1}} \left(\int_{\mathbf{R}^r} (1 + |z|^l)|m_s(z)|\, dz\right)^2 ds\right\}^2 + 2|\sigma| \int_{\sigma_n}^{\sigma_{n+1}} E[\chi(\sigma_n) g_N^2(s, s)]\, ds,$$

and by using the Cauchy inequality once more, we get

$$\sum_{n=1}^{N} E[\chi(\sigma_n)\mathbf{J}_n^2] \leqq 4L^2 \cdot E\int_0^T \left(\int_{\mathbf{R}^r} (1 + |z|^l)|m_s(z)|\, dz\right)^4 ds \cdot |\sigma| \to 0$$

as $|\sigma| \downarrow 0$.

Consequently, (6.9) implies that, for any $\varepsilon > 0$, $c > 0$,

$$\overline{\lim_{|\sigma| \downarrow 0}} P\left(\sum_{n=1}^{N} \mathbf{J}_n > \varepsilon\right) \leqq P[\sup_{0 \leqq t \leqq T} \|p_t\|_l > c],$$

whence:

$$\lim_{|\sigma| \downarrow 0} P\left(\sum_{n=1}^{N} \mathbf{J}_n > \varepsilon\right) = 0 \quad \text{for any } \varepsilon > 0,$$

which establishes (6.8). Q.E.D.

We are now in a position to prove the basic result of this section, which elucidates Mortensen's "dynamic programming in function space" approach:

THEOREM 6.2 (a "verification lemma" for Mortensen's equation). *Consider any admissible control process* $u=\{u_t;\ 0\leqq t\leqq T\}$ *in* $\mathbf{A}$, *along with the corresponding weak solution* $(x_t,\ y_t,\ w_t^u,\ b_t;\ 0\leqq t\leqq T)$ *on the probability space* $(\Omega, \mathbf{F}, P^u; \mathbf{F}_t)$ *of the system of equations* (2.8)–(2.9), *under the assumptions of* § 2. *Consider also the cost functions* $g(z)$, $\phi(a, z)$, *introduced in* § 5 *and satisfying Assumption* A3, *for some number* l, $l\leqq k-1$, *fixed henceforth.*

If the Mortensen equation

$$\begin{aligned}(6.10)\qquad V_1(\tau,p)&=\tfrac{1}{2}V_{22}(\tau,p)[h_{T-\tau}p,\ h_{T-\tau}p]\\&\quad+\min_{a\in U}\{(\mathbf{L}^a_{T-\tau}V_2(\tau,p)(\cdot),p)+(\phi(a,\cdot),p)\},\qquad (\tau,p)\in(0,T]+\mathbf{E}_k,\end{aligned}$$

$$(6.11)\qquad V(0,p)=(g,p),\qquad p\in\mathbf{E}_k,$$

has a solution $V(\tau,p)$; $[0,T]\times\mathbf{E}_k\to\mathbf{R}$ *which satisfies the conditions of Theorem* 6.1, *then* $V(\tau,p)$ *is a lower bound on the achievable expected cost; i.e., for any* $u\in\mathbf{A}$,

$$(6.12)\qquad V^u(\tau,p)\triangleq E^u_{T-\tau,p}[g(x_T)+\int_{T-\tau}^{T}\phi(u_t,x_t)\,dt]\geqq V(\tau,p)$$

for any $(\tau,p)\in[0,T]\times\mathbf{E}_k$.

Proof. We apply Ito's rule (Theorem 6.1) to the process $V(T-t,\rho_t)$, where $\rho_t(z)$ is the unnormalized density of the conditional distribution, under P^u, of x_t given $\mathbf{F}_t^y$, $T-\tau\leqq t\leqq T$. According to Theorem 4.1, $\rho_t(z)$ solves the stochastic equation (4.6) for $T-\tau<t\leqq T$ under the initial condition $\rho_{T-t}=p$, so we have, a.s. $\tilde{P}$,

$$\begin{aligned}dV(T-t,\rho_t)&=\{-V_1(T-t,\rho_t)+V_2(T-t,\rho_t)\}[\mathbf{L}_t^{*u}\rho_t]+\tfrac{1}{2}V_{22}(T-t,\rho_t)[h_t\rho_t,\ h_t\rho_t]\,dt\\&\quad+V_2(T-t,\rho_t)[h_t\rho_t]\cdot dy_t\\&\geqq-(\phi(u_t,\cdot),\rho_t)\,dt+V_2(T-t,\rho_t)[h_t\rho_t]\cdot dy_t,\end{aligned}$$

by virtue of the minimization in (6.10). Upon integrating the above differential inequality over $[T-\tau,\ T]$ and taking expectations with respect to measure $\tilde{P}$, we get in conjunction with (5.1):

$$\begin{aligned}V(\tau,p)&\leqq\tilde{E}\left[\int_{T-\tau}^{T}\left(\int_{\mathbf{R}^r}\phi(u_t,z)\rho_t(z)\,dz\right)dt+\int_{\mathbf{R}^r}g(z)\rho_T(z)\,dz\right]\\&=\tilde{E}\left[\int_{T-\tau}^{T}\left(\int\int_{\mathbf{R}^r}\phi(u_t,z)q(t,z\,;T-\tau,x)p(x)\,dz\,dx\right)dt\right.\\&\qquad\left.+\int\int_{\mathbf{R}^r}g(z)q(T,z\,;T-\tau,x)p(x)\,dz\,dx\right]\\&=V^u(\tau,p).\end{aligned}$$

Q.E.D.

7. Concluding remarks. We have presented and explained some aspects of a "dynamic programming in function space" approach to the problem of control of partially observable diffusion processes. In particular, we have shown the relevance of the Mortensen equation to reducing the global optimization problem of choosing a law to a pointwise minimization.

Future progress along these lines will depend on establishing the existence and uniqueness for solutions of this equation in the class of functions satisfying the conditions of Theorem 6.1. Under proper assumptions on the running cost function $\phi(a, z)$, the minimization in (6.10) would then suggest a natural candidate for the optimal process in the *separated* form

$$\tilde{u} = \{\tilde{u}_t = \tilde{a}(T-t, p_t);\ 0 \leqq t \leqq T\} \tag{7.1}$$

where

$$\begin{aligned} \tilde{a}(\tau, p) &= \arg\min_{a \in U} \{\mathbf{L}^a_{T-\tau} V_2(\tau, p)(\cdot), p) + (\phi(a, \cdot), p)\} \\ &= \arg\min_{a \in U} \left\{ a' \cdot \int_{\mathbf{R}^r} \nabla V_2(\tau, p)(z) p(z)\, dz + \int_{\mathbf{R}^r} \phi(a, z) p(z)\, dz \right\}. \end{aligned} \tag{7.2}$$

Indeed, the verification Theorem 6.2 would establish the optimality of $\tilde{u}$ as in (7.1), (7.2), provided it is shown that the latter is *admissible*, i.e., that the Zakai equation (4.6) is strongly solvable for an $\mathbf{F}^y_t$-adapted random function $\rho_t(z)$, with $u_t \equiv \tilde{a}(T-t, \rho_t/(\rho_t, 1))$.

Let us define the class of *separated control processes* as consisting of those processes $\{u_t;\ 0 \leqq t \leqq T\}$ for which there exists a measurable function

$$\alpha(\tau, p); \quad [0, T] \times \mathbf{E}_k \to U$$

such that

$$P[u_t = \alpha(T-t, p_t);\ 0 \leqq t \leqq T] = 1.$$

If the question of admissibility for the separated control processes, or some subclass thereof, is answered to the positive, one might be able to establish the Hamilton–Jacobi form (5.3)′ of the Mortensen equation for such controls as well, not just for the constant ones.

A "separated" problem for partially observed, controlled diffusions has also been formulated recently by W. H. Fleming and E. Pardoux [5] and by W. H. Fleming [6]. They use a convexified setup in which an *admissible* control law amounts to a measure on (y, u) paths such that y projects to Wiener measure, and the past (y, u) is independent of the future of y. If u can be obtained from y by a causal functional, they call the corresponding law *strictly admissible*. This subclass coincides with our admissible control laws. Also, Davis and Kohlmann [3] have recently refined the nonlinear semigroup approach by using results from convex analysis, which allow them to deal directly with our class **A** of admissible controls. By using their approach, it may be shown that $V(\tau, p)$ is actually the greatest lower bound on the performance of laws in **A**.

Acknowledgments. The use of $\tilde{E}$ in (5.1) to get rid of the awkward normalizer $\pi_t(1)$ was suggested by M. H. A. Davis in personal correspondence, and also by the referee; we thank them both.

REFERENCES

[1] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this Journal, 9 (1971), pp. 446–472.

[2] ———, *Full "bang" to reduce predicted miss is optimal*, this Journal, 14 (1976), pp. 62–84.

[3] M. H. A. DAVIS AND M. KOHLMANN, *On the nonlinear semigroup of stochastic control under partial observations*, submitted to this Journal.

[4] S. D. EIDEL'MAN, *Parabolic Systems*, North-Holland, Amsterdam, 1969.
[5] W. H. FLEMING AND E. PARDOUX, *Existence of optimal controls for partially observed diffusions*, this Journal, 20 (1982), pp. 261–285.
[6] W. H. FLEMING, *Nonlinear semigroup for controlled partially observed diffusions*, this Journal, 20 (1982), pp. 286–301.
[7] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[8] ———, *Stochastic Differential Equations and Applications*, Vol. 1, Academic Press, New York, 1975.
[9] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Probab. Appl., 5 (1960), pp. 285–301.
[10] H. J. KUSHNER, *On the dynamical equations of conditional probability density functions, with applications to optimal stochastic control theory*, J. Math. Anal. Appl., 8 (1964), pp. 332–344.
[11] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes, Vol.* I, *General Theory*, Springer-Verlag, Berlin, 1977.
[12] L. A. LUSTERNIK AND V. J. SOBOLEV, *Elements of Functional Analysis*, Hindustan Publishing Corp., New Delhi, India, 1974.
[13] H. P. MCKEAN, JR, *Stochastic Integrals*, Academic Press, New York, 1969.
[14] R. E. MORTENSEN, *Stochastic optimal control with noisy observations*, Internat. J. Control., 4 (1966), pp. 455–464.
[15] E. PARDOUX, *Backward and forward stochastic differential equations associated with a nonlinear filtering problem*, in Proc. 18th IEEE Conference on Decision and Control, Fort Lauderdale, FL, December 1979, pp. 166–171.
[16] A. N. SHIRYAEV, *Some new results in the theory of controlled stochastic processes* in Trans. 4th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Czech. Academy of Sciences, Prague, 1967 (in Russian).
[17] A. V. SKOROKHOD, *Studies in the Theory of Random Processes* Addison-Wesley, Reading, MA, 1965.
[18] R. L. STRATONOVICH, *Conditional Markov processes*, Theory Probab. Appl., 5 (1960), pp. 156–178.
[19] J. SZPIRGLAS, *Sur l'équivalence d'équations différentielles stochastiques à valeurs mesures intervenant dans le filtrage markovien non linéaire*, Ann. Inst. H. Poincaré Sect. B, 145 (1978), pp. 33–59.
[20] M. ZAKAI, *On the optimal filtering of diffusion processes*, Z. Wahrsch. Verw. Gebiete, 11 (1969), pp. 230–243.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 3, May 1983

0363-0129/83/2103-0012 $01.25/0

VECTOR-VALUED DYNAMIC PROGRAMMING*

MORDECHAI I. HENIG†

Abstract. Dynamic programming models with vector-valued returns are investigated. The sets of (Pareto) maximal returns and (Pareto) maximal policies are defined. Monotonicity conditions are shown to be sufficient for the set of maximal policies to include a stationary policy, and for the set of maximal returns to be in the convex hull of returns of stationary policies. In particular, it is shown that these results hold for Markov decision processes.

Key words. dynamic programming, multicriterion decision-making, Markov decision processes, monotone operators, convex analysis

1. Introduction. We study here a vector-valued analogue of the dynamic programming model proposed by Denardo [2]. The standard maximization is replaced by maximization with respect to a cone. The results can be applied to problems with several objective functions.

The model consists of a countable number of stages at which a decision has to be made. Policies are introduced, and each policy has a return which is a vector-valued function. Of special interest are the stationary policies.

Our main concern is to characterize the set of maximal returns. We introduce a monotonicity condition which suffices for the set of maximal returns to include the return of a stationary policy. Under stronger conditions we show that each maximal return can be obtained as a convex combination of returns of stationary policies. This result leads to an algorithm which approximates the set of maximal returns.

Specifically, we show that Markov decision processes with several objectives fulfil these requirements and, thus, the set of maximal returns can be approximated by multicriteria mathematical programming.

Lately, dynamic programming models with multicriteria have been applied in several areas: in fishery management by Mendelssohn (cf. [8]), in hospitalization by Schmee, Hannan and Mirabile [12], in budgeting by Grinold, Hopkins and Massy [4] and in scheduling by Kao [7].

Apparently the earliest results for dynamic vector-valued models are those of Brown and Strauch [1] for a finite horizon model in which they showed that the principle of optimality holds with respect to Pareto maximal returns. Furukawa [3] obtained similar results for maximization with respect to a convex cone in the Markov decision process.

The special cases of Markov decision processes with a finite number of decisions were analyzed by Viswanathan, Aggarwal and Nair [15], Nair and Aggarwal [10], White and Kim [17], Hartley [5] and Shin [13]. They developed efficient algorithms for finding the maximal stationary policies.

Ordinal and lexicographic optimization are two concepts that are related to multicriterion maximization. Ordinal dynamic programming was analyzed by Mitten [9] and Sobel [14]. Denardo [2] commented on lexicographic maximization. In these cases an optimal policy among the stationary policies can be found. However, this policy need not be (Pareto) maximal, as Example 6.2 shows.

* Received by the editors January 8, 1982, and in revised form May 25, 1982. This research was supported by the National Science Foundation under grant Eng-76-15559.

† Faculty of Management, Tel-Aviv University, Ramat-Aviv, Israel.

The paper is structured as follows: In § 2 we formulate the decision process and introduce boundedness and contraction conditions. In § 3 we introduce the maximization concept. A monotonicity condition is assumed under which we are able to transform the return functions into the set of real functions. Thus properties of one-criterion dynamic programming are used to obtain some preliminary results. In § 4 the main theorem is presented. It characterizes the set of maximal returns in terms of the maximal returns of stationary policies. This is used in § 5 to approximate the set of maximal returns. Three examples, one of them a Markov decision process, are given in § 6.

2. The vector-valued dynamic programming model. Consider the contraction dynamic programming model introduced by Denardo [2]. Let S be a nonempty set of *states*. For each $s \in S$, let $D(s)$ be a nonempty *action* set. Let the *decision set* D be the Cartesian product of the action sets. A *Markov policy* is an infinite sequence $\pi = (\delta_1, \delta_2, \cdots)$ of decisions $\delta_t \in D$ for all t; the collection of all such policies is denoted by D^*. A *stationary policy* is a Markov policy where $\delta_t = \delta$ for every t; the collection of all stationary policies is denoted by D^∞.

Let $I = \{1, 2, \cdots, n\}$ where n is fixed. Let W be the set of all bounded real-valued functions on $S \times I$ with the supremum norm

$$\|w\| = \sup \{|w(s, i)|: s \in S, i \in I\}.$$

The elements of I can be interpreted as the *criteria*. For each $i \in I$ let h_i be a function that assigns a real number to each triplet (s, a, w) with $s \in S$, $a \in D(s)$ and $w \in W$. Call

$$h(s, a, w) = [h_1(s, a, w), \cdots, h_n(s, a, w)]$$

the *local return* in state $s \in S$, if action $a \in D(s)$ is taken and $w \in W$ is given.

For $\delta \in D$, let H_δ be defined by

$$[H_\delta(w)](s, i) = h_i(s, \delta(s), w).$$

Accordingly, let

$$[H_\delta(w)](s) = h(s, \delta(s), w).$$

Given a policy $\pi \in D^*$ and given $w \in W$, define $H_\pi^t w$ inductively by $H_\pi^0 w = w$ and $H_\pi^t w = H_\pi^{t-1}(H_{\delta_t} w)$.

The following assumption is sufficient to ensure that each $H_\pi^t w \in W$:

Boundedness. There exist constants k_1 and k_2 such that $\|H_\delta w\| \leqq k_1 + k_2 \|w\|$ for all $w \in W$ and $\delta \in D$.

The *return function* of $\pi \in D^*$ is defined as $w_\pi = \lim_{t \to \infty} H_\pi^t w$, when the limit is well defined and independent of $w \in W$. A sufficient condition for $w_\pi \in W$ is the following assumption:

Contraction. There exists a constant c, $0 \leqq c < 1$, such that for all $w, v \in W$ and $\delta \in D$, $\|H_\delta w - H_\delta v\| \leqq c\|w - v\|$.

The proof that $w_\pi \in W$ uses the arguments that the sequence $\{H_\pi^t w\}$ is Cauchy and that W is a complete metric space with respect to the sup norm. The proof is standard and can be found in [6].

Throughout the paper, we assume that the contraction and boundedness assumptions are satisfied. However, contraction can be replaced by T-contraction (see definition in [2]), and boundedness can be relaxed by introducing the weighted sup norm (see definition and analysis in [16]).

3. The optimization criterion and preliminary results. The standard maximization criterion is usually not valid in the vector-valued case. Instead, quite frequently, a Pareto optimal policy is sought; i.e., we are looking for a policy Δ such that, for every $s \in S$, if for some policy π $w_\pi(s, i) \geqq w_\Delta(s, i)$ for all $i \in I$ then $w_\pi(s) = w_\Delta(s)$.

The main goal here is to characterize the set of Pareto optimal policies. We preferred, however, to maximize with respect to a larger class of criteria, in the following sense.

DEFINITION 3.1. Let Y and Λ be subsets of R^n. The set of *maximal* elements in Y with respect to Λ is $M(Y|\Lambda) = \{y \in Y | y \notin x + \Lambda \forall x \in Y\}$. The subset Λ is called a *domination set* and we shall say that x dominates y if $y \in x + \Lambda$.

When $\Lambda = \{a = (\alpha_1, \cdots, \alpha_n) \in R^n | \alpha_i \leqq 0 \forall i, \alpha \neq \{0\}\}$ then $M(Y|\Lambda)$ is the Pareto maximal set in Y.

Given a state s, let $w_\pi(s) = [w_\pi(s, i), \cdots, w_\pi(s, n)]$ be the return when policy $\pi \in D^*$ is used. Let $V(s) = \{w_\pi(s) | \pi \in D^*\}$ and $V = \{w_\pi | \pi \in D^*\}$. Notice that $V(s)$ is a subset of R^n.

DEFINITION 3.2. A *policy* $\pi \in D^*$ is *maximal for* $s \in S$ (with respect to Λ) if $w_\pi(s) \in M(V(s)|\Lambda)$. A *policy* $\pi \in D^*$ is *maximal* if it is maximal for each $s \in S$.

Let U be the set of all bounded functions from S to the reals. The main idea in this section is to transform the returns which are elements of W into elements in U and then use properties of one-criterion dynamic programming.

Given $\alpha \in R^n$, let T_α be the linear transformation from W to U defined as:

$$[T_\alpha w](s) \equiv w(s)\alpha \equiv \sum \alpha_i w(s, i).$$

Given $\alpha \in R^n \backslash \{0\}$ let L_α be the linear transformation from U to W defined as:

$$[L_\alpha u](s, i) \equiv \alpha_i^* \frac{u(s)}{\sum |\alpha_i|}, \quad \text{where } \alpha_i^* = \frac{\alpha_i}{|\alpha_i|} \quad (\alpha_i^* = 0 \text{ when } \alpha_i = 0).$$

Notice that $[T_\alpha(L_\alpha u)](s) = u(s)$ so that $T_\alpha L_\alpha$ is the identity function. The following condition generalizes the monotonicity condition in Denardo [2].

DEFINITION 3.3. For $\alpha \in R^n$, H_δ is said to be an *α-monotone* operator if $T_\alpha(w_1) \geqq T_\alpha(w_2)$ implies $T_\alpha(H_\delta w_1) \geqq T_\alpha(H_\delta w_2)$ for any $w_1, w_2 \in W$. We shall say that α-monotonicity is satisfied if H_δ is α-monotone for every $\delta \in D$. For $A \subseteq R^n$, A-monotonicity is said to hold if α-monotonicity is satisfied for every $\alpha \in A$.

Notice that $T_{\mu\alpha} w = \mu T_\alpha w$ for every $\mu \in R$ and therefore A-monotonicity implies μA-monotonicity for every $\mu \in R$, i.e., monotonicity is always satisfied for a union of subspaces of R^n. In the case of $n = 1$, the monotonicity assumption introduced by Denardo [2] is equivalent to $\{1\}$-monotonicity which implies R^1-monotonicity.

In § 6 we shall show that Markov decision processes with multicriteria satisfy the R^n-monotonicity condition.

DEFINITION 3.4. For a fixed $\alpha \in R^n \backslash \{0\}$ and each $\delta \in D$ define the operator G_δ on U by

$$G_\delta u \equiv T_\alpha[H_\delta L_\alpha(u)].$$

For $\pi = (\delta_1, \delta_2, \cdots) \in D^*$ and $u \in U$ define G_π^t inductively by $G_\pi^0 u = u$ and $G_\pi^t u = G_\pi^{t-1}(G_{\delta_t} u)$.

Notice that G_δ depends on α, and $[G_\delta u](s)$ is a weighted sum of $[H_\delta(L_\alpha u)](s, i)$. For convenience, the subscript α will be suppressed if α is fixed. Also H (or G) will be used when a property of H_δ (or G_δ) is true for every $\delta \in D$.

LEMMA 3.1. *Given $\alpha \in R^n$, if H is an α-monotone operator then G is a monotone operator in U satisfying the boundedness and the contraction conditions.*

Proof. Observe that $Lu \in W$ and $\|Lu\| = \|u\| / \sum |\alpha_i|$ for every $u \in U$. Boundedness and contraction can be easily verified.

To prove monotonicity let $u_1, u_2 \in U$ with $u_1 \geqq u_2$. Since TL is the identity function, $TLu_1 \geqq TLu_2$ where $Lu_1, Lu_2 \in W$. By the α-monotonicity assumption $T(Lu_1) \geqq T(Lu_2)$ implies $T(H_\delta(Lu_1)) \geqq T(H_\delta(Lu_2))$ for every $\delta \in D$. By the definition of G_δ, $G_\delta u_1 \geqq G_\delta u_2$ for every $\delta \in D$. □

As a consequence, the fixed-point theorem in complete metric spaces can be used (see [2]) to obtain a unique $v_\delta \in U$ for every $\delta \in D$ such that $v_\delta = G_\delta v_\delta$.

DEFINITION 3.5. Let $f = \sup_{\delta \in D} v_\delta$ where the supremum is taken pointwise.

Denardo [2, Thm. 3] shows that f is the unique solution for $v = \sup_{\delta \in D} G_\delta v$, $v \in U$. Furthermore, f can be approximated by some return. These results are summarized as follows:

LEMMA 3.2. (i) *For every $\delta \in D$ there exists a unique $v_\delta \in U$ such that $v_\delta = G_\delta v_\delta = \lim_{t \to \infty} G_\Delta^t u \equiv v_\Delta$ where $\Delta = (\delta, \delta, \cdots) \in D^\infty$ and $u \in U$.*

(ii) $f = \sup_{\delta \in D} G_\delta f$.

(iii) *For every $\varepsilon > 0$ there exists $\Delta \in D^\infty$ with $\|v_\Delta - f\| < \varepsilon$.*

In the following lemma we show that $T(w_\pi)$ can be obtained by using the operators G_π^t.

LEMMA 3.3. *Given $\alpha \in R^n$, if H is an α-monotone operator, then for every $\pi \in D^*$, $u \in U$ and $t = 1, 2, \cdots$*

$$G_\pi^t u = T(H_\pi^t(Lu)) \quad \text{and} \quad \lim_{t \to \infty} G_\pi^t u = T(w_\pi).$$

Proof. Notice that $Tw = TL(Tw)$, and by the α-monotonicity assumption $T(H_\delta w) = T(H_\delta(LTw))$. Fix t. The proof of the first identity is by decreasing induction. By definition $G_{\delta_t} u = T(H_{\delta_t}(Lu))$. Suppose that for some τ, $1 < \tau \leqq t$,

$$G_{\delta_\tau} G_{\delta_{\tau+1}} \cdots G_{\delta_t} u = T(H_{\delta_\tau} H_{\delta_{\tau+1}} \cdots H_{\delta_t}(Lu)).$$

By definition and the above result,

$$G_{\delta_{\tau-1}}(G_{\delta_\tau} \cdots G_{\delta_t} u) = T(H_{\delta_{\tau-1}} L(T(H_{\delta_\tau} \cdots H_{\delta_t}(Lu)))) = T(H_{\delta_{\tau-1}} H_{\delta_\tau} \cdots H_{\delta_t}(Lu)).$$

Hence the argument is true for $\tau - 1$ and by induction the claim is true for $\tau = 1$, as required.

By the continuity of T

$$\lim_{t \to \infty} G_\pi^t u = \lim_{t \to \infty} T(H_\pi^t(Lu)) = T\left(\lim_{t \to \infty} H_\pi^t(Lu)\right) = T(w_\pi). \qquad \square$$

By combining the two lemmas, we can show now that a sequence of stationary policies can be found whose limit of transformed returns achieves f.

LEMMA 3.4. *Given $\alpha \in R^n$, if H is an α-monotone operator, then there exists a sequence $\{\Delta_j\}$ of stationary policies such that*

$$\lim_{j \to \infty} T(w_{\Delta_j}) = f.$$

Furthermore

$$\lim_{j \to \infty} T(w_{\Delta_j}) \geqq T(w_\pi) \quad \text{for any } \pi \in D^*.$$

Proof. By the previous lemmas, for every $\varepsilon > 0$ there exists $\Delta \in D^\infty$ such that

$$\|T(w_\Delta) - f\| = \|\lim_{t \to \infty} G_\Delta^t u - f\| = \|v_\Delta - f\| < \varepsilon.$$

By [2, Thm. 5]

$$f \geqq \lim_{t \to \infty} G_\pi^t u = T(w_\pi) \quad \text{for any } \pi \in D^*. \qquad \square$$

The limit can actually be attained in certain conditions, as in the following result which can be proved by standard methods.

COROLLARY 3.5. *Suppose for each $s \in S$ and $w \in W$ that $h(s, \cdot, w)$ is a continuous function in a topology for which $D(s)$ is compact; then $f = T(w_\Delta)$ for some stationary policy Δ.*

Remark. Under reasonable assumptions, as shown in Henig [6], Lemma 3.4 can be extended to include policies which are not in D^*, as history-remembering and randomized policies.

4. The sets of maximal returns and policies. We shall show when every maximal return can be expressed as a linear combination of returns obtained by stationary policies.

DEFINITIONS AND NOTATION. A set $\Lambda \subseteq R^n$ is a *cone* if $\lambda \Lambda \subseteq \Lambda$ for every $\lambda \in R$, $\lambda > 0$.

The *polar cone* of Y is $Y^0 = \{\alpha \mid y\alpha \leqq 0 \text{ for all } y \in Y\}$.

The *strict polar cone* of Y is $Y^* = \{\alpha \mid y\alpha < 0 \text{ for all } y \in Y \backslash \{0\}\}$.

The *recession cone* of Y is $0^+Y = \{\alpha \mid y + \lambda\alpha \in Y \text{ for every } y \in Y \text{ and } \lambda \geqq 0\}$.

A cone Λ is *strictly supported* if $\Lambda^* \neq \varnothing$.

Define $\operatorname{Ex}(Y \mid \alpha) = \{y \in Y \mid y\alpha > y_0\alpha \ \forall y_0 \in Y \backslash \{y\}\}$. This set is either empty or a singleton and is then called an *exposed point* (see Rockafellar [11]).

For $A \subseteq R^n$ define $\operatorname{Ex}(Y \mid A) = \bigcup_{\alpha \in A} \operatorname{Ex}(Y \mid \alpha)$.

For simplicity we assume that $V(s)$ is closed (or closure of $V(s)$ replaces $V(s)$). We also assume that $V(s)$ is convex (or convex hull of $V(s)$ replaces $V(s)$). This means that any convex combination of returns can be obtained by some policy (e.g., by initial randomization among policies).

When a state $s \in S$ is given we shall write $C \equiv V(s)$.

The first result will give sufficient conditions for the set of maximal policies to contain a stationary policy.

THEOREM 4.1. *Given a state $s \in S$, let Λ be a cone in R^n. Suppose A-monotonicity holds for some $A \subseteq R^n$ such that $A \cap \Lambda^* \neq \varnothing$. Then there exists a sequence $\{\Delta_j\}$ of stationary policies such that*

$$\lim_{j \to \infty} w_{\Delta_j}(s) \in M(C \mid \Lambda).$$

Proof. Pick any $\alpha \in A \cap \Lambda^*$. As a consequence of Lemma 3.4 there exists a sequence $\{\Delta_j\}$, $\Delta_j \in D^\infty$ such that

$$\lim_j w_{\Delta_j}(s)\alpha \geqq x\alpha \quad \forall x \in C.$$

Let $x_0 \equiv \lim_j w_{\Delta_j}(s)$ and suppose that $x_0 \notin M(C \mid \Lambda)$. Then there exists $x_1 \in C$ such that $x_0 - x_1 \in \Lambda$. By definition of the strict polar cone, $\alpha(x_0 - x_1) < 0$ and $\alpha x_0 < \alpha x_1$ which is a contradiction. Hence $\lim_j w_{\Delta_j}(s) \in M(C \mid \Lambda)$. □

We should remark that the theorem makes sense only if $\Lambda^* = \varnothing$ which is satisfied if and only if Λ is contained in a homogeneous open half space (excluding perhaps the origin).

In the last theorem, the sequence of stationary policies may depend on $s \in S$. However, if Λ is fixed for each $s \in S$, Lemma 3.4 can be used to conclude that for every $s \in S$,

$$\lim_{j \to \infty} w_{\Delta_j}(s) \in M(V(s) \mid \Lambda).$$

This proves the next result.

COROLLARY 4.2. *Under the assumptions of Theorem* 4.1, *if* Λ *is the same for every* $s \in S$, *then there exists a sequence* $\{\Delta_j\}$ *of stationary policies such that*

$$\lim_{j\to\infty} w_{\Delta_j} \in X_{s\in S} M(V(s)|\Lambda).$$

Moreover, there exists a stationary maximal policy if the conditions of Corollary 3.5 *are satisfied.*

Two lemmas will now precede the main result of the section. The first will show that every exposed point of C can be approximated by a stationary policy. The second will present the directions of recession of C.

LEMMA 4.3. *Let* $y \in \mathrm{Ex}(C|\alpha)$. *If the* α*-monotonicity assumption is satisfied, then there exists a sequence of stationary policies* $\{\Delta_j\} \in D^\infty$ *such that*

$$\lim_{j\to\infty} w_{\Delta_j}(s) = y.$$

Proof. Since $y \in C$, by Lemma 3.4 there exists $\{\Delta_j\}$, $\Delta_j \in D^\infty$, such that $\lim_{j\to\infty} w_{\Delta_j}(s)\alpha \geqq y\alpha$. Clearly, $\lim_{j\to\infty} w_{\Delta_j}(s) \in C$.

If $\lim_j w_{\Delta_j}(s) \neq y$ then, since y is an exposed point, $y\alpha > \lim_j w_{\Delta_j}(s)\alpha$ which is a contradiction. Hence $\lim_j w_{\Delta_j}(s) = y$. □

LEMMA 4.4. *Suppose* A*-monotonicity holds for some* $A \subseteq R^n$. *Then* $0^+C \subseteq (A)^0$.

Proof. Let $k \in 0^+C$ and choose any $x \in C$. By definition $x + \lambda k \in C$ for every $\lambda \geqq 0$. By Lemma 3.4 $\sup\{(x+\lambda k)\alpha \,|\, \lambda \geqq 0\} < \infty$ for every $\alpha \in A$. Since $x\alpha$ is fixed, then $\sup\{\lambda k\alpha \,|\, \lambda \geqq 0\} \leqq 0$ and $k\alpha \leqq 0$ for every $\alpha \in A$. Hence by definition $k \in A^0$. □

Notice that monotonicity is always satisfied for a union of subspaces of R^n. Therefore $\mathrm{co}(A) = R^m$ for some $m \leqq n$. By Lemma 4.4 if $\mathrm{co}(A) = R^n$ then $0^+C \subseteq (R^n)^0 = \{0\}$ and by [11, Thm. 8.4] C is compact.

Let $F(s)$ denote the elements of $V(s)$ which can be obtained by stationary policies. The theorem claims that if the appropriate monotonicity condition is valid then the following is true: For a given state $s \in S$ a return is maximal only if it is a maximal element in $\mathrm{co}(\bar{F}(s))$.

THEOREM 4.5. *Given* $s \in S$, *let* Λ *be a strictly supported convex cone. Let* $\Gamma \subseteq \Lambda \cup \{0\}$ *be a closed convex acute cone and suppose that* Γ^**-monotonicity is satisfied, then* $M(C|\Lambda) \subseteq M(\mathrm{co}\,\bar{F}(s)|\Lambda)$.

Proof. Consider the subset $C + \Gamma$. By Lemma 4.4 $0^+(C) \subseteq (\Gamma^*)^0 = \Gamma$ and since Γ is acute we get $0^+(C) \cap -\Gamma = \{0\}$. By [11, Corollary 9.1.2] $C + \Gamma$ is closed and convex and $0^+(C+\Gamma) = 0^+(C) + \Gamma = \Gamma$. Hence $0^+(C+\Gamma) \cap -0^+(C+\Gamma) \subseteq \Gamma \cap -\Gamma = \{0\}$ and $C + \Gamma$ contains no lines.

Let $x_0 \in M(C+\Gamma|\Lambda)$. By [11, Thm. 18.5] $x_0 = x_1 + d$ where $x_1 = \sum_{j=1}^{l} \lambda_j g_j$, $d = \sum_{j=l+1}^{k} \lambda_j d_j$, each g_j (d_j) is an extreme point (direction) of $C + \Gamma$, $\lambda_j > 0$ for every j and $\sum_{j=1}^{l} \lambda_j = 1$. Notice that $d \in 0^+(C+\Gamma) = \Gamma$ which implies $x_0 - x_1 \in \Gamma \subseteq \Lambda \cup \{0\}$. Since $x_0 \in M(C+\Gamma|\Lambda)$ it means that $x_0 = x_1$ and x_0 is in the convex hull of the extreme points of $C + \Gamma$.

By [11, Thm. 18.6] $M(C+\Gamma|\Lambda) \subseteq \mathrm{co}\,\overline{\mathrm{Ex}}(C+\Gamma|R^n)$. In matter of fact it can be shown that $M(C+\Gamma|\Lambda) \subseteq \mathrm{co}\,\overline{\mathrm{Ex}}(C+\Gamma|\Gamma^*)$ as follows. Let $\alpha \in R^n$ and $x_0 \in C + \Gamma$ be so that $x_0\alpha > x\alpha$ for all $x \in (C+\Gamma)\backslash\{x_0\}$. Choose any $\beta \in \Gamma\backslash\{0\}$. Clearly $x_0 + \beta \in C + \Gamma$ and $x_0\alpha > x_0\alpha + \beta\alpha$. Hence $\beta\alpha < 0$ for every $\beta \in \Gamma\backslash\{0\}$ and by definition $\alpha \in \Gamma^*$. It is easy to verify that $M(C+\Gamma|\Lambda) = M(C|\Lambda)$ and $\mathrm{Ex}(C+\Gamma|\Lambda) = \mathrm{Ex}(C|\Lambda)$ and so $M(C|\Lambda) \subseteq \mathrm{co}\,\overline{\mathrm{Ex}}(C|\Gamma^*)$. By Lemma 4.3 and the assumption that $M(C|\Lambda) \subseteq \mathrm{co}\,\bar{F}(s)$. Since $\mathrm{co}(\bar{F}(s)) \subseteq C$ we get $M(C|\Lambda) \subseteq M(\mathrm{co}\,\bar{F}(s)|\Lambda)$. □

The theorem suggests that the search for maximal elements in $V(s)$, for a fixed state s, can be limited to a smaller subset, namely $\mathrm{co}(\bar{F}(s))$.

Notice that for $\Lambda=\{0\}$ we get $M(Y|\Lambda)=Y$, so that the following corollary is immediate.

COROLLARY 4.6. *If R^n-monotonicity is satisfied then for every $s\in S$*

$$V(s)=\text{co}\,(\bar{F}(s)).$$

When $V(s)$ is not convex the theorem is not necessarily true. The theorem then states that a maximal return is dominated by a convex combination of stationary returns. The corollary states that $V(s)\subseteq \text{co}\,(\bar{F}(s))$.

5. Approximations for the sets of maximal returns and policies. A procedure to approximate $M(C|\Lambda)$ is now presented. Basically the procedure suggests that for each $\alpha\in\Lambda^*$, we must find a stationary policy, whose transformed return is maximal. Each of these returns can be approximated by an algorithm which prevails in one-criterion dynamic programming.

Let $Y\subseteq R^n$ and $\alpha\in R^n$. The support subset of Y with respect to α is defined as

$$\text{supp}\,(Y|\alpha)=\{y\in Y|\, y\alpha\geqq y_0\alpha\,\forall y_0\in Y\}.$$

For $A\subseteq R^n$, let

$$\text{supp}\,(Y|A)=\bigcup_{\alpha\in A}\text{supp}\,(Y|\alpha).$$

The main result is a corollary of Theorem 4.5.

COROLLARY 5.1. *With the same conditions presented in Theorem* 4.5

$$M(C|\Lambda)\subseteq \text{co}\,(\overline{\text{supp}}\,(\bar{F}(s)|\Gamma^*)).$$

In particular if Λ is closed then

$$M(C|\Lambda)\subseteq \text{co}\,(\overline{\text{supp}}\,(\bar{F}(s)|\Lambda^*)).$$

Proof. In the proof of Theorem 4.5 we showed that $M(C|\Lambda)\subseteq \text{co}\,\overline{\text{Ex}}\,(C|\Gamma^*)$. It is clear that $x_0\in\overline{\text{Ex}}\,(C|\Gamma^*)$ implies $x_0\in\bar{F}(s)\cap C$ and therefore $x_0\in\overline{\text{Ex}}\,(\bar{F}(s)|\Gamma^*)$. Since the support subset contains the exposed points the result is proven. □

The corollary suggests that supp $(\bar{F}(s)|\alpha)$ should be approximated for each $\alpha\in\Lambda^*$. Using now the notation of § 3, for fixed $\alpha\in\Lambda^*$, we can write

$$\text{supp}\,(\bar{F}(s)|\alpha)=\{w(s)\in\bar{F}(s)|\,(Tw)(s)\geqq(Tw_0)(s)\forall w_0\in V\}.$$

By Lemmas 3.2 and 3.3 we can write

$$\text{supp}\,(\bar{F}(s)|\alpha)=\left\{w(s)|\,T(w)=f=\sup_{\delta\in D}G_\delta f\right\}.$$

Since G is a contraction and monotone operator in U, the usual optimization schemes in dynamic programming will approximate $T(w)=f$. Moreover, these schemes also provide a stationary policy whose return approximates f. Thus, for every $\varepsilon>0$, a stationary policy, Δ, can be found such that

$$\|f-T(w_\Delta)\|<\varepsilon.$$

The following theorem uses this information to approximate the maximal set of returns and policies.

THEOREM 5.2. *Under the same conditions as in Theorem* 4.5, *let $\alpha\in\Lambda^*$, and suppose that*

$$\|f-T(w_\Delta)\|\leqq\varepsilon \quad \textit{for some } \Delta\in D^\infty \textit{ and } \varepsilon\geqq 0.$$

Then for a fixed $s \in S$ and any $\beta \in -\Lambda$

$$w_\Delta(s) + \beta\lambda \notin C \quad \textit{for all } \lambda > \frac{\varepsilon}{\beta\alpha}.$$

Proof. By definition of strict polar cone, $\beta\alpha > 0$ so $(\varepsilon/\beta\alpha)$ is well defined. Suppose $w_\Delta(s) + \beta\lambda \in C$ for some $\lambda > \varepsilon/\beta\alpha$. Then

$$f(s) \geqq (w_\Delta(s) + \beta\lambda)\alpha = w_\Delta(s)\alpha + \beta\lambda\alpha > w_\Delta(s)\alpha + \varepsilon,$$

which is a contradiction. Hence $w_\Delta(s) + \beta\lambda \notin C$. □

By this theorem, if a maximal element of C exists along the ray $w_\Delta(s) + \lambda\beta$ then its distance to $w_\Delta(s)$ is less than $\varepsilon/\beta\alpha$.

In certain processes, particularly when the sets of states and decisions are finite, we can find $\Delta \in D^\infty$ with $T(w_\Delta) = f$. It is clear then that $w_\Delta(s) \in M(V(s) \mid \Lambda)$ for every $s \in S.$, Otherwise $w_\Delta(s) + \beta = y$ for some $y \in V(s)$ and $\beta \in \Lambda$, which contradicts the theorem. This proves the following result:

COROLLARY 5.3. *If the assumptions of the theorem hold with $\varepsilon = 0$ then $w_\Delta(s) \in M(C \mid \Lambda)$.*

6. Examples. *Example* 6.1. A common model is the infinite-horizon discounted Markov decision process, with countable sets of states and decisions. The local return function for $w \in W$ and $\delta \in D$ is

$$H_\delta w = r_\delta + c \cdot P_\delta w, \quad \text{where } w \in W,\ r_\delta \in W,\ c \geqq 0,\ P_\delta \in R^{S \times S} \text{ and } P_\delta \geqq 0.$$

The boundedness and contraction assumptions are satisfied in several cases (e.g., $c < 1$ and P_δ is substochastic).

To show R^n-monotonicity, suppose that for a fixed $\alpha \in R^n$

$$T(w_1) \geqq T(w_2) \quad \text{for any } w_1, w_2 \in W.$$

Then

$$T(H_\delta w_1) = T(r_\delta + cP_\delta w_1) = T(r_\delta) + cP_\delta T(w_1) \geqq T(r_\delta) + cP_\delta T(w_2)$$
$$= T(H_\delta w_2) \quad \text{for every } \delta \in D.$$

This is true for any $\alpha \in R^n$; hence R^n-monotonicity is satisfied.

Consider $M(C \mid \Lambda)$ where $\Lambda = \{\beta \mid \beta_i \leqq 0 \forall i \in I,\ \beta \neq 0\}$.

Then $\Lambda^* = \{\alpha \mid \alpha_i > 0 \forall i)$. By Corollary 5.1 we should look for

$$\operatorname{supp}(\bar{F}(s) \mid \alpha) \quad \text{for every } \alpha \in \Lambda^*.$$

The operator $G_\delta : U \to U$ is defined as

$$G_\delta u \equiv T_\alpha[H_\delta L_\alpha(u)] = T_\alpha(r_\delta) + cP_\delta u.$$

A known method to approximate f is by mathematical programming (see Denardo [2]):

$$f = \min u \quad \text{such that } T_\alpha(r_\delta) + c \cdot P_\delta u \leqq u \quad \text{for every } \delta \in D.$$

When the sets of states and decisions are finite we have a linear programming model whose dual is:

$$\max\left(\sum_{\delta \in D} x_\delta T_\alpha(r_\delta)\right) \quad \text{such that}$$

$$x_\delta(I - cP_\delta) = 1 \quad \text{and} \quad x_\delta \geqq 0$$

for every $\delta \in D$, where $x_\delta \in R^{1 \times S}$ and $1 = (1, \cdots, 1)$.

This program, solved for every $\alpha \in \Lambda^*$, locates the set of all maximal solutions of multiple linear criteria defined over a polyhedron. Efficient ways to solve the dual program for each $\alpha \in \Lambda^*$ were suggested in Viswanathan, Aggarwal and Nair [15], Hartley [5] and Shin [13]. More recently, White and Kim [17] suggested successive approximations and policy iteration techniques to solve the finite decisions case.

Example 6.2. Let $n=2$, $S=\{1\}$, $D=\{\delta \mid 0 \leqq \delta \leqq 1\}$ and $H_\delta w = (\max[\delta, cw_1], \max[1-\delta, cw_2])$ where $w \in W = R^2$ and $0 \leqq c < 1$. It is easy to verify that this model satisfies the boundedness and contraction conditions.

The fixed point $w_\delta = H_\delta w_\delta$ is $w_\delta = (\delta, 1-\delta)$, $0 \leqq \delta \leqq 1$. Consider the nonstationary policies $\Gamma_1 = (\delta_1 = 0, \delta_t = 1 \forall t \geqq 2)$ and $\Gamma_2 = (\delta_1 = 1, \delta_t = 0, \forall t \geqq 2)$. It is easy to verify that $w_{\Gamma_1} = (1, c)$ and $w_{)_2} = (c, 1)$. Clearly, if $\Lambda = \{\beta \in R^2 \mid \beta_i \leqq 0, \beta \neq \{0\}\}$ then $(\delta, 1-\delta)$ is dominated by some linear combination of w_{Γ_1} and w_{Γ_2}, so that no stationary policy can be maximal. We shall note also that lexicographic maximization over the stationary policies will produce the policy $\delta_t = 1$ for every t, giving us a dominated policy.

It can be shown that A-monotonicity in this case is satisfied only for $A = \{\alpha \in R^2 \mid \alpha_1 \cdot \alpha_2 = 0\}$. Since $\Lambda^* = \{\alpha \in R^n \mid \alpha_i > 0 i = 1, 2\}$ we have $\Lambda^* \cap A = \varnothing$ which violates the condition in Theorem 4.1.

$$\text{Let } \Lambda = \{\beta \in R^2 \mid \beta_1 < 0\}, \text{ then } \Lambda^* = \{\alpha \in R^2 \mid \alpha_1 > 0, \alpha_2 = 0\}$$

and $A \cap \Lambda^* \neq \varnothing$ which assures the existence of a stationary policy maximizing the first criterion.

Example 6.3. Let $n=2$, $s=\{1\}$, $D=\{\delta_1, \delta_2\}$, $H_{\delta_1} w = (1,1) + \frac{1}{2}(w_2, w_1)$ and $H_{\delta_2} w = (1,2) + \frac{1}{2}(w_1, w_2)$.

Boundedness and contraction are satisfied, and it can be shown that A-monotonicity is satisfied for $A = \{\alpha \in R^2 \mid \alpha_1 = \alpha_2\}$. The fixed points of H_{Γ_1} and H_{Γ_2} are: $w_1 = (2, 2)$ and $w_2 = (2, 4)$. When $\Lambda = \{\beta \in R^2 \mid \beta_i \leqq 0, \beta \neq \{0\}\}$ then $A \cap \Lambda^* \neq \varnothing$ and Theorem 4.1 implies that w_2 is a maximal return. However Λ^*-monotonicity is not satisfied and the set of maximal policies cannot be expressed as a linear combination of the stationary policies.

Acknowledgments. This paper is part of the author's Ph.D. thesis done at Yale University under the supervision of Professor Mathew J. Sobel to whom the author is very grateful. He also wishes to acknowledge helpful comments from Eric V. Denardo, Martin Puterman and Uriel Rothblum.

REFERENCES

[1] T. A. BROWN AND R. E. STRAUCH, *Dynamic programming in multiplicative lattices*, J. Math Anal. Appl., 12 (1965), pp. 364–370.

[2] E. V. DENARDO, *Contraction mappings in the theory underlying dynamic programming*, SIAM Rev., 9 (1967), pp. 165–177.

[3] N. FURUKAWA, *Characterization of optimal policies in vector-valued Markovian decision processes*, Math. Oper. Res., 5 (2) (1980), pp. 271–279.

[4] R. GRINOLD, D. HOPKINS AND W. MASSY, *A model for long-range university budget planning under uncertainty*, Bell Economics J., 9 (2) (1978), pp. 396–420.

[5] R. HARTLEY, *Finite, Discounted, Vector Markov Decision Processes*, University of Manchester, 1979.

[6] M. HENIG, *Multicriteria dynamic programming*. Ph.D. Dissertation. Yale University, New Haven, CT 1978.

[7] E. P. C. KAO, *A multiple objective decision theoretic approach to one-machine scheduling problems*, Computers and Operations Research, 7 (1980), pp. 251–259.

[8] R. MENDELSSOHN, *Pareto policies for harvesting with multiple objectives*, Math. Biosci., 51 (3–4) (1980), pp. 213–224.

[9] L. G. MITTEN, *Preference order dynamic programming*, Management Sci., 21 (1974), pp. 43–46.
[10] K. P. NAIR AND V. AGGARWAL, *Stationarity property in multiple criteria Markov decision processes*, Operations Research, to appear.
[11] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton NJ, 1972.
[12] J. SCHMEE, E. HANNAN AND M. MIRABILE, *An examination of patient referral and discharge policies using a multiple objective semi-Markov decision process*, J. Opl. Res. Soc., 30 (2) (1979), pp. 121–129.
[13] M. SHIN, *Computational methods for Markov decision problems*, Ph.D. Dissertation, University of British Columbia, 1980.
[14] M. J. SOBEL, *Ordinal Dynamic Programming*, Management Sci., 21 (1975), pp. 967–975.
[15] B. VISWANATHAN, V. AGGARWAL AND K. NAIR, *Multiple criteria Markov decision processes*, TIMS Studies in Management Sciences, 6 (1977), pp. 263–272.
[16] J. WESSELS, *Markov programming by successive approximations with respect to weighted supremum norms*, J. Math. Anal. Appl., 58 (1977), pp. 326–335.
[17] C. WHITE AND K. KIM, *Solution procedures for vector criterion Markov decision processes.* Large Scale Systems, 1 (1980), pp. 129–140.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

0363-0129/83/2104-0001 $01.25/0

IDENTIFIABILITY OF LINEAR SYSTEMS IN HILBERT SPACES*

SHIN-ICHI NAKAGIRI†

Abstract. The identifiability problem is discussed for linear dynamical systems described by first and second order evolution equations in Hilbert spaces. The unknowns are the initial values and the operators appearing in the system equations and a number of identifiability conditions are established within the framework of linear operator theory. These are applied to various classes of partial differential equations on bounded and unbounded spatial domains to obtain the constant and spatially varying parameter identifiability conditions for such systems.

Key words. identifiability, distributed system, semigroup, eigenfunction expansion, system theory

1. Introduction. The present paper studies the identifiability problem for linear systems in Hilbert spaces. In order to determine unknown parameters or operators in a given system, the model reference method is usually employed, that is, the method of minimizing the difference between the system outputs and the model outputs. In the final stage of this process, there arises the problem of whether or not the parameters or operators in the model system coincide with those in the true system when the error between the outputs of both systems becomes zero. This problem, the so-called identifiability problem, is of considerable interest and importance in the field of control engineering and also gives rise to interesting questions in applied mathematics (see [3], [16], [23], [24], [27] and the references cited therein). The purpose here is to solve the problem in distributed systems in some abstract manner.

In recent years, there have appeared a number of papers which deal with the identifiability of (constant and spatially varying) parameters in specific classes of distributed systems of parabolic type [5], [16], [20], [21], [23], [26], [27]. In particular, Kitamura and the author [16] proposed a method of solving the identifiability problem which largely depends on the uniqueness of Dirichlet series. By applying the method to various systems, many identifiability conditions are established for constant parameters in [5], [16], [21], [23]. Pierce [23] has sharpened the method in a satisfactory way and has obtained an interesting identifiability result for spatially varying parameters with the aid of the Gel'fand–Levitan theory. Some complementary results to [23] are also obtained by Suzuki and Murayama [26] and Suzuki [27] without using the Gel'fand–Levitan theory directly. However, most results are limited to systems described by 1-dimensional parabolic partial differential equations on bounded intervals.

In this paper we consider the abstract systems represented by first and second order evolution equations in Hilbert spaces and provide a formulation of the identifiability problem by using their model systems. In the problem formulation, the unknowns to be identified are initial values and an operator, which is assumed to generate a semigroup or a cosine family, appearing in the system equation. We treat the problem within the framework of linear operator theory and give a number of abstract identifiability results under the basic assumption that the operator acting on the state is selfadjoint (more generally, normal) with compact resolvent. This assumption implies the existence of eigenvalues and eigenfunctions of such an operator, and this enables us to treat the problem considered in [16]. Because of the generality of

* Received by the editors July 15, 1980, and in final revised form June 20, 1982.

† Department of Applied Mathematics, Faculty of Engineering, Kobe University, Rokkodai, Nada, Kobe 657, Japan.

our treatment, our results can be applied to a wide variety of dynamical systems including those in [3], [16], [20], [21], [23], [26] (see examples in §§ 4 and 5).

We enumerate the contents of this paper. In § 2, the notation used throughout the paper and a formulation of the identifiability problem in Hilbert spaces are given. The representations of system states and preliminary results are given in § 3. Section 4 is concerned with the unique identification of eigenvalues. For a bounded observation, some identifiability conditions of all eigenvalues are established in Theorems 4.1 and 4.2 for first and second order systems, respectively. Theorems 4.3 and 4.4, in which the observation is not necessarily bounded, are for the identifiability of two constant parameters in operators and are deduced from the coincidence of two eigenvalues. Section 5 is concerned with the identifiability of the whole system, in other words, the unique identification of the initial values and the operator (i.e., its eigenvalues and eigenfunctions). To identify the system requires finite numbers of zero output errors. For the whole domain observation, the rank conditions for identifiability are established in Theorems 5.1–5.4. The rank conditions named the compatibility conditions are expressed in terms of the known (model) quantities only. In § 6, an extension of Theorem 5.1 is given, and the relations between identifiability and observability and controllability are discussed. Applications to identifiability for some parabolic and hyperbolic partial differential equations on bounded and unbounded domains are given in Examples 4.1–4.3 and 5.1, 5.2. Detailed investigations of the practical identifiability problems are made in these examples, requiring some additional knowledge of such equations. In Example 4.1 the identifiability of a spatially varying coefficient in some 1-dimensional hyperbolic equation (of normal type) is established via the Gel'fand–Levitan theory. Examples 4.2 and 4.3 give the constant parameter identifiability conditions of N-dimensional parabolic and hyperbolic equations, respectively. The identifiability of coefficients in a second order elliptic differential operator (on a bounded domain) by means of initial and forcing inputs is established in Example 5.1. Example 5.2 treats the identifiability of a potential in the Schrödinger equation.

2. Basic notation and statement of the problem. We use the following notation throughout this paper.

$R^+ = [0, \infty)$;

R^N, the Euclidean N-space;

$R = R^1$; C^N, the complex N-space;

X, the underlying complex separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ (the corresponding norm is denoted by $\| \cdot \|$);

Y, a Banach space;

$L(X, Y)$, the space of all bounded linear operators with domain X and range in Y;

$C^r(J; X)$, $J \subset R$, $r = 0, 1, 2, \cdots$, the space of all r-times (strongly) continuously differentiable functions from J into X;

$L_p(J; X)$, $J \subset R$, $p \geqq 1$, the space of all equivalent classes of (strongly) measurable functions from J into X which are p-Bochner integrable on J;

$C(R^+; X) = C^0(R^+; X)$;

$L_p^{\mathrm{loc}}(R^+; X) = \bigcap_{t>0} L_p([0, t]; X)$;

$L_p(J) = L_p(J; C^1)$, $J \subset R$;

$L_p^{\mathrm{loc}}(R^+) = L_p^{\mathrm{loc}}(R^+; C^1)$;

G, a (closed or open) domain in R^N;

$\bar{G}$, the closure of G;

∂G, the smooth boundary of G;

$C^r(G)(C^r(\partial G))$, $r=0, 1, 2, \cdots$, the space of r-times continuously differentiable functions on $G(\partial G)$ with values in C^1;

$C(G)=C^0(G)$;$C^\infty(G)=\bigcap_{r=0}^{\infty} C^r(G)$;

$C^\infty(\partial G)=\bigcap_{r=0}^{\infty} C^r(\partial G)$;

$C_0^\infty(G)(C_0^r(G))$, the set of all functions in $C^\infty(G)(C^r(G))$ having compact support in $\bar{G}-\partial G$;

$L_2(G)$, the complex Hilbert space of all square Lebesgue integrable functions on G (its inner product and norm are denoted by $\langle\cdot,\cdot\rangle_G$ and $\|\cdot\|_G$, respectively);

$H_m(G)$, the Sobolev space on G of order m.

The spaces $L_p([a,b])$, $C^r([a,b])$, etc. on $[a,b](a<b)$ are denoted simply by $L_p[a,b]$, $C^r[a,b]$, etc.

Consider the following first and second order linear systems on X:

$$S_1: \quad \dot{x}(t)=Ax(t)+f(t), \qquad x(0)=x_0; \tag{2.1}$$

$$S_2: \quad \ddot{x}(t)=Ax(t)+f(t), \qquad x(0)=x_0, \quad \dot{x}(0)=y_0, \tag{2.2}$$

where $x(t)\in X$, x_0, $y_0\in X$, $f(\cdot)\in L_2^{\mathrm{loc}}(R^+;X)$ and A is a closed linear operator with a dense domain $D(A)\subset X$. In the systems S_1 and S_2, the initial values x_0, y_0 and the forcing function $f(\cdot)$ are considered to be system inputs. The observation of the system state $x(t)$ of S_1 or S_2 is

$$y(t)=Bx(t), \qquad t\geqq 0, \tag{2.3}$$

where $B\in L(X, Y)$ is the observation operator. The function $y(t)$ is called the system output.

We assume throughout this paper that the following conditions are satisfied for the identifiability of the systems S_1 and S_2:

I. The operator A is unknown except that:
   (i) A in the system S_1 generates a strongly continuous semigroup $\{T(t): t\in R^+\}$, and
   (ii) A in the system S_2 generates a strongly continuous cosine family $\{C(t): t\in R\}$.

II. The operator B is a priori known.

III. The initial values x_0, $y_0\in X$ are unknown but the forcing input $f(\cdot)\in L_2^{\mathrm{loc}}(R^+;X)$ is known.

A and x_0, y_0 are unknown quantities to be determined.

From I and III (more generally, from the weaker condition that x_0, $y_0\in X$, $f(\cdot)\in L_1^{\mathrm{loc}}(R^+;X)$), it follows that the functions

$$x_1(t)=T(t)x_0+\int_0^t T(t-s)f(s)\,ds \tag{2.4}$$

and

$$x_2(t)=C(t)x_0+S(t)y_0+\int_0^t S(t-s)f(s)\,ds \tag{2.5}$$

make sense as the sums of a continuous function and a Bochner integral in X and are strongly continuous on R^+, i.e., $x_1(\cdot)$, $x_2(\cdot)\in C(R^+;X)$. Here the operator $S(t)$ is defined by

$$S(t)x=\int_0^t C(s)x\,ds, \qquad x\in X, \quad t\in R. \tag{2.6}$$

The function x_1 (resp. x_2) is called the mild solution of S_1 (resp. S_2). In this paper we treat the identifiability problem by means of the mild solution.

For the given systems S_1 and S_2, we also consider the model systems $S_1^{\mathscr{m}}$ and $S_2^{\mathscr{m}}$ which are given by replacing A, x_0 and A, x_0, y_0 in S_1 and S_2 by $A^{\mathscr{m}}$ satisfying I(i), $x_0^{\mathscr{m}} \in X$ and $A^{\mathscr{m}}$ satisfying I(ii), $x_0^{\mathscr{m}}$, $y_0^{\mathscr{m}} \in X$, respectively. The corresponding mild solution of $S_j^{\mathscr{m}}$ is denoted by $x_j^{\mathscr{m}}(t)$ $(j=1, 2)$. All quantities suffixed by $\mathscr{m}$ are assumed to be known.

Let J be an interval in R^+. The difference

$$e(S_j, S_j^{\mathscr{m}}; t) = y_j(t) - y_j^{\mathscr{m}}(t) = Bx_j(t) - Bx_j^{\mathscr{m}}(t), \qquad t \in J, \tag{2.7}$$

is called the output error between S_j and $S_j^{\mathscr{m}}$ on J.

Now the identifiability problem for S_1 (resp. S_2) can be stated as follows:

Under what conditions do $A = A^{\mathscr{m}}$ and/or $x_0 = x_0^{\mathscr{m}}$ (resp. $A = A^{\mathscr{m}}$ and/or $x_0 = x_0^{\mathscr{m}}$, $y_0 = y_0^{\mathscr{m}}$) follow from the zero output error on J

$$e(S_1, S_1^{\mathscr{m}}; t) = 0 \quad (\text{resp. } e(S_2, S_2^{\mathscr{m}}; t) = 0) \quad \text{in } Y, \quad t \in J?$$

Remark 2.1. The assumption I(ii) is equivalent to the existence of a unique *weak* solution of S_2. That is, for any $x_0, y_0 \in X$ and $f(\cdot) \in L_1^{\mathrm{loc}}(R^+; X)$, there exists a function $x(\cdot) \in C(R^+; X)$ such that for any $v \in D(A^*)$ (A^* denotes the adjoint of A), $\langle x(t), v\rangle$ is differentiable on $R^+ - \{0\}$, $(d/dt)\langle x(t), v\rangle$ is locally absolutely continuous on R^+ and $\langle x(t), v\rangle$ satisfies

$$\frac{d^2}{dt^2}\langle x(t), v\rangle = \langle x(t), A^*v\rangle + \langle f(t), v\rangle \quad \text{a.e. } t \in R^+,$$

$$x(0) = x_0, \qquad \frac{d}{dt}\langle x(0), v\rangle = \langle y_0, v\rangle.$$

Furthermore, the *weak* solution $x(t)$ is given by the right-hand side of (2.5). Recently a similar fact for the system S_1 in a general Banach space has been established by Ball [2].

Remark 2.2. Let A satisfy I(ii) and let $f(\cdot) \in C^1(R^+; X)$, $x_0 \in \{x \in X: C(\cdot)x \in C^2(R; X)\}$, $y_0 \in \{x \in X: C(\cdot)x \in C^1(R; X)\}$ be satisfied. Then the mild solution $x_2(t)$ defined by (2.5) is a unique *strong* solution of S_2, i.e., $x_2(\cdot) \in C^2(R^+; X)$, $x(t) \in D(A)$ for each $t \in R^+$, and $x_2(t)$ satisfies the equalities in (2.2) (cf. Travis and Webb [29]). See also the next section.

Remark 2.3. Under the assumption I(ii), A generates an analytic semigroup $T(t)$ on the half plane $\{t \in C: Re\, t > 0\}$ which is given explicitly by

$$T(t)x = \frac{1}{\sqrt{\pi t}} \int_0^\infty \exp\left(-\frac{s^2}{4t}\right) C(s)x \, ds, \qquad x \in X, \quad t > 0. \tag{2.8}$$

For more important properties of cosine families, we refer to the fundamental papers by Fattorini [8] and Sova [25].

3. Spectral theory and representations of mild solutions. To discuss the identifiability by using the general theory of spectral decompositions, we suppose that the operator A and its model operator $A^{\mathscr{m}}$ satisfy the following assumption (H) throughout the paper.

(H). A is selfadjoint with compact resolvent.

Under the assumption (H), the next general fact on the eigenfunction expansions for A holds (cf. Kato [14, p. 277], Triggiani [30, p. 323], [31, p. 857]).

There exists a set of eigenvalues and eigenfunctions $\{\lambda_n, \phi_{nj}: j = 1, \cdots, m_n, n = 1, 2, \cdots\}$ of A such that:

(a) The spectrum $\sigma(A) = \{\lambda_n: n = 1, 2, \cdots\} \subset R$ and

$$+\infty > C \geqq \lambda_1 > \lambda_2 > \cdots > \lambda_n > \cdots, \qquad \lim_{n\to\infty} \lambda_n = -\infty. \tag{3.1}$$

(b) The system $\{\phi_{nj}: j = 1, \cdots, m_n, n = 1, 2, \cdots\}$ is a complete orthonormal system in X, i.e., the following unique expansion holds:

$$x = \sum_{n=1}^{\infty} \sum_{j=1}^{m_n} \langle x, \phi_{nj}\rangle \phi_{nj}, \qquad x \in X.$$

(c) Ax, $x \in D(A)$ and $R(\lambda; A)y$, $y \in X$, $\lambda \notin \sigma(A)$ are given respectively by

$$Ax = \sum_{n=1}^{\infty} \lambda_n \sum_{j=1}^{m_n} \langle x, \phi_{nj}\rangle \phi_{nj}, \tag{3.2}$$

$$D(A) = \left\{x \in X: \sum_{n=1}^{\infty} |\lambda_n|^2 \sum_{j=1}^{m_n} |\langle x, \phi_{nj}\rangle|^2 < \infty\right\} \tag{3.3}$$

and

$$R(\lambda; A)y = (\lambda - A)^{-1} y = \sum_{n=1}^{\infty} \frac{1}{\lambda - \lambda_n} \sum_{j=1}^{m_n} \langle y, \phi_{nj}\rangle \phi_{nj}. \tag{3.4}$$

(d) If A satisfies I(i), then the semigroup $T(t)$ generated by A is given by

$$T(t)x = \sum_{n=1}^{\infty} e^{\lambda_n t} \sum_{j=1}^{m_n} \langle x, \phi_{nj}\rangle \phi_{nj}, \qquad x \in X, \quad t \in R^+. \tag{3.5}$$

(e) If A satisfies I(ii), then the cosine family $C(t)$ and the sine family $S(t)$ generated by A are given respectively by

$$C(t)x = \sum_{n=1}^{\infty} \cos \sqrt{-\lambda_n}\, t \sum_{j=1}^{m_n} \langle x, \phi_{nj}\rangle \phi_{nj}, \qquad x \in X, \quad t \in R, \tag{3.6}$$

and

$$S(t)x = \sum_{n=1}^{\infty} \frac{\sin \sqrt{-\lambda_n}\, t}{\sqrt{-\lambda_n}} \sum_{j=1}^{m_n} \langle x, \phi_{nj}\rangle \phi_{nj}, \qquad x \in X, \quad t \in R. \tag{3.7}$$

Here we remark that if $\lambda_n = 0$ for some n, the term $(\sin \sqrt{-\lambda_n}\, t)/\sqrt{-\lambda_n}$ in (3.7) is replaced by t.

The following propositions are evident.

PROPOSITION 3.1. *The mild solution* (2.4) *of* S_1 *is given by*

$$\begin{aligned} x_1(t) &= \sum_{n=1}^{\infty} e^{\lambda_n t} \sum_{j=1}^{m_n} \langle x_0, \phi_{nj}\rangle \phi_{nj} \\ &\quad + \sum_{n=1}^{\infty} \sum_{j=1}^{m_n} \left(\int_0^t e^{\lambda_n (t-s)} \langle f(s), \phi_{nj}\rangle \, ds \right) \phi_{nj}, \qquad t \in R^+. \end{aligned} \tag{3.8}$$

PROPOSITION 3.2. *The mild solution* (2.5) *of* S_2 *is given by*

$$x_2(t) = \sum_{n=1}^{\infty} \cos\sqrt{-\lambda_n}t \sum_{j=1}^{m_n} \langle x_0, \phi_{nj}\rangle \phi_{nj} + \sum_{n=1}^{\infty} \frac{\sin\sqrt{-\lambda_n}t}{\sqrt{-\lambda_n}} \sum_{j=1}^{m_n} \langle y_0, \phi_{nj}\rangle \phi_{nj}$$

$$+ \sum_{n=1}^{\infty} \sum_{j=1}^{m_n} \left(\int_0^t \frac{\sin\sqrt{-\lambda_n}(t-s)}{\sqrt{-\lambda_n}} \langle f(s), \phi_{nj}\rangle \, ds \right) \phi_{nj}, \qquad t \in R^+. \tag{3.9}$$

Similar representations hold for the mild solutions of models.

Let k be a nonnegative integer. Define the space $H_k^{\mathrm{loc}}(R^+; X)$ by

$$H_k^{\mathrm{loc}}(R^+; X) = \{f(\cdot) \in L_2^{\mathrm{loc}}(R^+; X) : f(t) \in D(A^k) \text{ for a.e. } t \in R^+ \text{ and } A^k f(\cdot) \in L_2^{\mathrm{loc}}(R^+; X)\}.$$

It is proved by direct calculations that if $x_0 \in D(A)$ (resp. x_0, $y_0 \in D(A)$) and $f(\cdot) \in H_1^{\mathrm{loc}}(R^+; X)$, then $x_1(t)$ in (3.8) (resp. $x_2(t)$ in (3.9)) is the *strong* solution of S_1 (resp. S_2).

By (3.1) and (3.8), it is verified that the mild solution $x_1(t)$ of S_1 belongs to $D(A^k)$ for each $t > 0$ if $x_0 \in X$ and $f(\cdot) \in H_k^{\mathrm{loc}}(R^+; X)$. In the case of S_2, the circumstances are slightly different. That is, the mild solution $x_2(t)$ of S_2 belongs to $D(A^k)$ for each $t \geqq 0$ if x_0, $y_0 \in D(A^k)$ and $f(\cdot) \in H_k^{\mathrm{loc}}(R^+; X)$.

Now we consider some restrictive, not necessary bounded, observation operator B. Assume that:

(3.10) B is a continuous linear functional on $D(A^k)$ (so that $Y = C^1$).

Here the topology of $D(A^k)$ is induced by the graph norm $\|\cdot\|_{D(A^k)}$;

$$\|x\|_{D(A^k)} = \|x\| + \|A^k x\| \quad \text{for } x \in D(A^k).$$

Then we have the following propositions which will be used for the case of pointwise observation in the next section.

PROPOSITION 3.3. *Let* $x_0 \in X$, $f(\cdot) \in H_k^{\mathrm{loc}}(R^+; X)$ *and let* B *satisfy* (3.10). *Then the output* $y_1(t)$ *of* S_1 *is given by*

$$y_1(t) = \sum_{n=1}^{\infty} \left(\sum_{j=1}^{m_n} \langle x_0, \phi_{nj}\rangle B\phi_{nj} \right) e^{\lambda_n t} + \sum_{n=1}^{\infty} \int_0^t e^{\lambda_n (t-s)} \left(\sum_{j=1}^{m_n} \langle f(s), \phi_{nj}\rangle B\phi_{nj} \right) ds \tag{3.11}$$

for $t > 0$. *Moreover, the first series in* (3.11) *converges uniformly on* $[\delta, \infty)$ *for any fixed* $\delta > 0$ *and the second series converges uniformly on any compact set in* R^+.

PROPOSITION 3.4. *Let* x_0, $y_0 \in D(A^k)$, $f(\cdot) \in H_k^{\mathrm{loc}}(R^+; X)$ *and let* B *satisfy* (3.10). *Then the output* $y_2(t)$ *of* S_2 *is given by*

$$y_2(t) = \sum_{n=1}^{\infty} \left(\sum_{j=1}^{m_n} \langle x_0, \phi_{nj}\rangle B\phi_{nj} \right) \cos\sqrt{-\lambda_n}t$$

$$+ \sum_{n=1}^{\infty} \left(\sum_{j=1}^{m_n} \langle y_0, \phi_{nj}\rangle B\phi_{nj} \right) \frac{\sin\sqrt{-\lambda_n}t}{\sqrt{-\lambda_n}}$$

$$+ \sum_{n=1}^{\infty} \int_0^t \frac{\sin\sqrt{-\lambda_n}(t-s)}{\sqrt{-\lambda_n}} \left(\sum_{j=1}^{m_n} \langle f(s), \phi_{nj}\rangle B\phi_{nj} \right) ds, \qquad t \in R^+. \tag{3.12}$$

Moreover, all series in (3.12) *converge uniformly on any compact set in* R^+.

4. Identifiability of eigenvalues. In this section we discuss the identifiability of eigenvalues of unknown operator A. Since the operator A and its model operator A^m satisfy the assumption (H), there exist the sets of eigenvalues and eigenfunctions of A and A^m satisfying (a), (b), (c) in § 3, which we denote by $\{\lambda_n, \phi_{nj}: j = 1, \cdots, m_n, n = 1, 2, \cdots\}$ and $\{\Lambda_n, \Phi_{nj}: j = 1, \cdots, k_n, n = 1, 2, \cdots\}$, respectively. The set $\{\Lambda_n, \Phi_{nj}\}$ is considered to be known. Let X_n (X_n^m) be the eigenmanifold corresponding to the eigenvalue λ_n (Λ_n) and let P_n (P_n^m) be the associated eigenprojector, i.e.,

$$X_n = \mathrm{Ker}\,(\lambda_n - A) \quad (X_n^m = \mathrm{Ker}\,(\Lambda_n - A^m)) \quad \text{and}$$

$$P_n x = \sum_{j=1}^{m_n} \langle x, \phi_{nj}\rangle \phi_{nj} \in X_n \quad \left(P_n^m x = \sum_{j=1}^{k_n} \langle x, \Phi_{nj}\rangle \Phi_{nj} \in X_n^m\right), \qquad x \in X.$$

Here $m_n = \dim X_n$ (resp. $k_n = \dim X_n^m$) and $m_\infty = \sup\{m_n: n = 1, 2, \cdots\}$ (resp. $k_\infty = \sup\{k_n: n = 1, 2, \cdots\}$) is called the multiplicity of A (resp. A^m).

DEFINITION 4.1. The eigenvalues $\{\lambda_n: n = 1, 2, \cdots\}$ of A in S_j are said to be *identifiable on* J if the zero output error $e(S_j, S_j^m; t) = 0$ in Y, $t \in J$, implies $\lambda_n = \Lambda_n$ for all $n = 1, 2, \cdots$ ($j = 1, 2$).

Let the space of observations Y be C^1. Then by Riesz's theorem, the observation operator B is given by

$$Bx = \langle w, x\rangle \quad \text{for some } w \in X \text{ and for all } x \in X. \tag{4.1}$$

For a less restrictive observation operator in specific cases we refer to the examples in this section.

From (3.8), (3.9) and the continuity of inner product (or from (3.11), (3.12) with $k = 0$) it follows that the output errors $e(S_1, S_1^m; t)$ and $e(S_2, S_2^m; t)$ are represented by

$$\begin{aligned} e(S_1, S_1^m; t) = {} & \sum_{n=1}^{\infty} \langle w, P_n x_0\rangle e^{\lambda_n t} - \sum_{n=1}^{\infty} \langle w, P_n^m x_0^m\rangle e^{\Lambda_n t} \\ & + \sum_{n=1}^{\infty} \int_0^t e^{\lambda_n (t-s)} \langle w, P_n f(s)\rangle\, ds - \sum_{n=1}^{\infty} \int_0^t e^{\Lambda_n (t-s)} \langle w, P_n^m f(s)\rangle\, ds, \qquad t \in R^+ \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} e(S_2, S_2^m; t) = {} & \sum_{n=1}^{\infty} \langle w, P_n x_0\rangle \cos\sqrt{-\lambda_n}\, t - \sum_{n=1}^{\infty} \langle w, P_n^m x_0^m\rangle \cos\sqrt{-\Lambda_n}\, t \\ & + \sum_{n=1}^{\infty} \langle w, P_n y_0\rangle \frac{\sin\sqrt{-\lambda_n}\, t}{\sqrt{-\lambda_n}} - \sum_{n=1}^{\infty} \langle w, P_n^m y_0^m\rangle \frac{\sin\sqrt{-\Lambda_n}\, t}{\sqrt{-\Lambda_n}} \\ & + \sum_{n=1}^{\infty} \int_0^t \frac{\sin\sqrt{-\lambda_n}(t-s)}{\sqrt{-\lambda_n}} \langle w, P_n f(s)\rangle\, ds \\ & - \sum_{n=1}^{\infty} \int_0^t \frac{\sin\sqrt{-\Lambda_n}(t-s)}{\sqrt{-\Lambda_n}} \langle w, P_n^m f(s)\rangle\, ds, \qquad t \in R^+, \end{aligned} \tag{4.3}$$

respectively. It is clear that $e(S_1, S_1^m; \cdot)$, $e(S_2, S_2^m; \cdot) \in C(R^+)$.

Using the expression (4.2) and the uniqueness of Dirichlet series, we obtain the following result which gives an abstract version of the result by Pierce [23, Thm. 1] to linear systems of first order. However, our result does not include Pierce's result as a special case, since ours is for distributed observation, while his is for pointwise observation.

THEOREM 4.1. *Let the observation operator B be given by* (4.1). *The eigenvalues* $\{\lambda_n: n=1,2,\cdots\}$ *of A in S_1 are identifiable on J in the following two cases*:

i) *If J is of positive measure, $f(\cdot)=0$ in $L_2^{\mathrm{loc}}(R^+;X)$ and if $\langle w, P_n x_0\rangle \neq 0$, $\langle w, P_n^{m} x_0^{m}\rangle \neq 0$ for all n.*

ii) *If $J=[0,T]$, $T>0$ (resp. $J=R^+$), $x_0=x_0^{m}=0$ in X, f has the form $z_0 g(t)$, $z_0\in X$, $g(\cdot)\in L_2^{\mathrm{loc}}(R^+)$ such that $g(\cdot)\neq 0$ in $L_2[0,T]$ (resp. $g(\cdot)\neq 0$ in $L_2^{\mathrm{loc}}(R^+)$) and if $\langle w, P_n z_0\rangle \neq 0$, $\langle w, P_n^{m} z_0\rangle \neq 0$ for all n.*

Proof. Case i). In this case the output error e on J is given by

$$e(S_1, S_1; t)=\sum_{n=1}^{\infty} c_n e^{\lambda_n t}-\sum_{n=1}^{\infty} c_n^{m} e^{\Lambda_n t}, \qquad t\in J,$$

where $c_n=\langle w, P_n x_0\rangle$, $c_n^{m}=\langle w, P_n^{m} x_0^{m}\rangle$. Assume that $e(S_1, S_1^{m}; t)=0$ for all t in J. Since J is of positive measure and both series of e converge for all t in R^+, we see by the analytic continuation that $e(S_1, S_1^{m}; t)=0$ for all t in R^+. Then if $c_n\neq 0$, $c_n^{m}\neq 0$ for all n, we have $\lambda_n=\Lambda_n$ and $c_n=c_n^{m}$ for all n by the unique expansion of Dirichlet series.

Case ii). Since $x_0=x_0^{m}=0$, the output error e on J is given by

$$\begin{aligned} e(S_1, S_1^{m}; t) &= \int_0^t \left(\sum_{n=1}^{\infty} h_n e^{\lambda_n (t-s)}-\sum_{n=1}^{\infty} h_n^{m} e^{\Lambda_n (t-s)}\right) g(s)\, ds \\ &= \int_0^t h(t-s) g(s)\, ds, \qquad t\in J, \end{aligned}$$

where $h_n=\langle w, P_n z_0\rangle$, $h_n^{m}=\langle w, P_n^{m} z_0\rangle$. Let $J=[0,T]$, $T>0$ and let $e(S_1, S_1^{m}; \cdot)=0$ in $C[0,T]$. It is clear that $h(t)$ is continuous on $[0,T]$ (and analytic on $(0,T]$) and $g(\cdot)\in L_2[0,T]\subset L_1[0,T]$. Then by the extended Titchmarsh's theorem [28, Thm. 151, pp. 324–325] there exist two nonnegative numbers t_1 and t_2 satisfying $t_1+t_2=T$ such that $h(t)=0$ for all $t\in[0,t_1]$ and $g(t)=0$ for almost every $t\in[0,t_2]$. Define a number $t_g\geqq 0$ by $t_g=\sup\{t: g(s)=0 \text{ for a.e. } s\in[0,t]\}$. Then the condition $t_g<T$ is equivalent to $g(\cdot)\neq 0$ in $L_2[0,T]$, and this implies $h(t)=0$ on $[0,T-t_g)$, so that $h(t)=0$ on R^+ by analyticity. Therefore, as in Case i) the identifiability of eigenvalues $\{\lambda_n: n=1,2,\cdots\}$ on $J=[0,T]$ follows from $h_n\neq 0$, $h_n^{m}\neq 0$ for all n. If $J=R^+$, then the conclusion follows immediately from Titchmarsh's theorem [28, Thm. 152, p. 325].

For linear systems of second order, we establish the next theorem.

THEOREM 4.2. *Let the observation operator B be given by* (4.1). *The eigenvalues* $\{\lambda_n: n=1,2,\cdots\}$ *of A in S_2 are identifiable on J in the following cases*:

i) *If J is of positive measure, $f(\cdot)=0$ in $L_2^{\mathrm{loc}}(R^+;X)$ and if*

$$\langle w, P_n x_0\rangle \neq 0, \qquad \langle w, P_n^{m} x_0^{m}\rangle \neq 0 \quad \text{for all } n \quad \text{or if}$$

$$\langle w, P_n y_0\rangle \neq 0, \qquad \langle w, P_n^{m} y_0^{m}\rangle \neq 0 \quad \text{for all } n.$$

ii) *If $J=[0,T]$, $T>0$ (resp. $J=R^+$), $x_0=x_0^{m}=y_0=y_0^{m}=0$ in X, f has the form $z_0 g(t)$, $z_0\in X$, $g(\cdot)\in L_2^{\mathrm{loc}}(R^+)$ such that $g(\cdot)\neq 0$ in $L_2[0,T]$ (resp. $g(\cdot)\neq 0$ in $L_2^{\mathrm{loc}}(R^+)$) and if*

$$\langle w, P_n z_0\rangle \neq 0, \qquad \langle w, P_n^{m} z_0\rangle \neq 0 \quad \text{for all } n.$$

Proof. Assume that for each t in J,

$$(4.4)\quad \sum_{n=1}^{\infty} c_n \cos\sqrt{-\lambda_n}t-\sum_{n=1}^{\infty} c_n^{m}\cos\sqrt{-\Lambda_n}t+\sum_{n=1}^{\infty} d_n\frac{\sin\sqrt{-\lambda_n}t}{\sqrt{-\lambda_n}}-\sum_{n=1}^{\infty} d_n^{m}\frac{\sin\sqrt{-\Lambda_n}t}{\sqrt{-\Lambda_n}}=0,$$

where $c_n=\langle w, P_n x_0\rangle$, $c_n^{m}=\langle w, P_n x_0^{m}\rangle$ and $d_n=\langle w, P_n y_0\rangle$, $d_n^{m}=\langle w, P_n^{m} y_0^{m}\rangle$. Assume also

that J is of positive measure. Then by analyticity of sine and cosine functions on R, the equality (4.4) holds for all t in R. Therefore, by using the same argument contained in [31, p. 857], we obtain that

$$c(t)=\sum_{n=1}^{\infty} c_n \cos\sqrt{-\lambda_n}t-\sum_{n=1}^{\infty} c_n''' \cos\sqrt{-\Lambda_n}t=0, \qquad t\geqq 0,$$

and

$$d(t)=\sum_{n=1}^{\infty} d_n\frac{\sin\sqrt{-\lambda_n}t}{\sqrt{-\lambda_n}}-\sum_{n=1}^{\infty} d_n'''\frac{\sin\sqrt{-\Lambda_n}t}{\sqrt{-\Lambda_n}}=0, \qquad t\geqq 0.$$

Since $c(t)$ is bounded by a function of exponential order (finite numbers of eigenvalues may be positive!), the integral transformation of $c(t)$ can be taken to obtain that

$$\frac{1}{\sqrt{\pi t}}\int_0^{\infty}\exp\left(-\frac{s^2}{4t}\right)c(s)\,ds=\sum_{n=1}^{\infty} c_n e^{\lambda_n t}-\sum_{n=1}^{\infty} c_n''' e^{\Lambda_n t}=0, \qquad t>0. \tag{4.5}$$

Next differentiating $d(t)$ and taking the same integral transformation, we have

$$\sum_{n=1}^{\infty} d_n e^{\lambda_n t}-\sum_{n=1}^{\infty} d_n''' e^{\Lambda_n t}=0, \qquad t>0. \tag{4.6}$$

The equalities (4.5) and (4.6) prove this theorem in Case i). The proof of Case ii) is similar to that of Theorem 4.1.

The above theorems require some information on unknown functions. For example, in Case i) of Theorem 4.1 the condition that $\langle w, P_n x_0\rangle \neq 0$ for all n, in which all $P_n x_0$ are unknown, is necessary for the identifiability of eigenvalues. If the condition is not satisfied, however, the coincidence of all eigenvalues does not follow. Actually, let $\langle w, P_n''' x_0'''\rangle \neq 0$ for all n and let $\langle w, P_n x_0\rangle = 0$ for some n (say n_0). Then the zero output error implies $\{\Lambda_n : n=1, 2, \cdots\}\subset\{\lambda_n : n=1, 2, \cdots\}$ and $\Lambda_{n_0}\neq\lambda_{n_0}$, which means the nonidentifiability of $\{\lambda_n : n=1, 2, \cdots\}$.

In some classes of linear systems described by 1-dimensional partial differential equations, such conditions (of unknown functions) are automatically satisfied (see [23, Thm. 2], [26], and the following example).

Example 4.1. Consider the system described by the hyperbolic partial differential equation

$$\frac{\partial^2 x}{\partial t^2}=\frac{\partial}{\partial\xi}\left(a(\xi)\frac{\partial x}{\partial\xi}\right)+b(\xi)x, \qquad t>0, \quad \xi\in(0,1), \tag{4.7}$$

with boundary and initial conditions

$$\alpha_0 x(t,0)-(1-\alpha_0)\frac{\partial x}{\partial\xi}(t,0)=\alpha_1 x(t,1)+(1-\alpha_1)\frac{\partial x}{\partial\xi}(t,1)=0, \tag{4.8}$$

$$x(0,\xi)=x_0(\xi), \quad \frac{\partial x}{\partial t}(0,\xi)=y_0(\xi), \quad \xi\in(0,1), \tag{4.9}$$

where $a(\xi)$, $b(\xi)$ are real functions and α_0, α_1 are constants such that $0\leqq\alpha_0, \alpha_1\leqq 1$. For the system (4.7)–(4.9) we assume the following conditions (which are weaker than those used in [16], [23], [26]):

(IV) the spatially varying coefficients $a(\xi)$ and $b(\xi)$ are unknown except that $a(\xi)>0$ for $\xi\in[0,1]$, $a(\cdot)\in C^1[0,1]$ and $b(\cdot)\in C[0,1]$;

(V) the boundary coefficients α_0, α_1 and the initial conditions $x_0(\cdot)$, $y_0(\cdot) \in L_2[0, 1]$ are unknown.

By the model we understand the system (4.7)–(4.9) in which $a(\xi)$, $b(\xi)$, α_0, α_1, $x_0(\xi)$ and $y_0(\xi)$ are replaced by $a'''(\xi)$, $b'''(\xi)$, α_0''', α_1''', $x_0'''(\xi)$ and $y_0'''(\xi)$ satisfying the same conditions as in (IV) and (V), respectively.

We denote by A the realization in $L_2[0, 1]$ of the Stürm–Liouville operator $(\partial/\partial\xi)(a(\xi)\partial/\partial\xi)+b(\xi)$ with the boundary condition (4.8), i.e., the operator A is given by

$$D(A)=\Big\{z \in L_2[0,1]: z \in H_1[0,1],\ a\frac{\partial z}{\partial \xi} \in H_1[0,1] \quad \text{and}$$

$$\alpha_0 z(0)-(1-\alpha_0)\frac{\partial z}{\partial \xi}(0)=\alpha_1 z(1)+(1-\alpha_1)\frac{\partial z}{\partial \xi}(1)=0\Big\}, \tag{4.10}$$

$$Ax=\frac{\partial}{\partial \xi}\Big(a(\xi)\frac{\partial}{\partial \xi}\Big)x+b(\xi)x \quad \text{for } x \in D(A).$$

Similarly the realization of model is denoted by A'''. Then the system (4.7)–(4.9) and its model can be written by the following evolution equations in $L_2[0, 1]$:

$$S: \ddot{x}(t)=Ax(t), \qquad x(0)=x_0, \quad \dot{x}(0)=y_0,$$

$$S''': \ddot{x}(t)=A'''x(t), \qquad x(0)=x_0''', \quad \dot{x}(0)=y_0'''.$$

Since A and A''' satisfy I(ii) and (H), there exist two sets of eigenvalues and eigenfunctions $\{\lambda_n, \phi_n: n=1, 2, \cdots\}$ of A and $\{\Lambda_n, \Phi_n: n=1, 2, \cdots\}$ of A''' (the multiplicity is 1) such that the mild solutions x of S and x''' of S''' are given respectively by

$$\begin{aligned} x(t,\cdot)=&\sum_{n=1}^{\infty}\langle x_0, \phi_n\rangle_{[0,1]}\cos\sqrt{-\lambda_n}t\phi_n(\cdot) \\ &+\sum_{n=1}^{\infty}\langle y_0, \phi_n\rangle_{[0,1]}\frac{\sin\sqrt{-\lambda_n}t}{\sqrt{-\lambda_n}}\phi_n(\cdot), \end{aligned} \tag{4.11}$$

$$\begin{aligned} x'''(t,\cdot)=&\sum_{n=1}^{\infty}\langle x_0''', \Phi_n\rangle_{[0,1]}\cos\sqrt{-\Lambda_n}t\Phi_n(\cdot) \\ &+\sum_{n=1}^{\infty}\langle y_0''', \Phi_n\rangle_{[0,1]}\frac{\sin\sqrt{-\Lambda_n}t}{\sqrt{-\Lambda_n}}\Phi_n(\cdot), \end{aligned} \tag{4.12}$$

where $\langle u, v\rangle_{[0,1]}=\int_0^1 u(\xi)v(\xi)\,d\xi$.

It is also known [11, pp. 270–273] that

$$\phi_n, \Phi_n,\ n=1, 2, \cdots, \text{ are uniformly bounded on } [0, 1] \tag{4.13}$$

and

$$\sqrt{-\lambda_n}=Cn+O\Big(\frac{1}{n}\Big), \qquad \sqrt{-\Lambda_n}=C'''n+O\Big(\frac{1}{n}\Big) \quad \text{as } n\to\infty, \tag{4.14}$$

where C and C''' are constants depending only on a and a''', respectively.

Then if $a=a'''$, we have $C=C'''$.

The substitution and elsewhere $x(t, \xi)$ of ξ in (4.11) has sense as a continuous function in t for almost every $\xi \in [0, 1]$. This function $x(t, \xi)$, which defines an element in $L_2^{\mathrm{loc}}(R^+; L_2[0, 1])$, is the *weak* solution of (4.7)–(4.9) in the sense of Ito [13, Chap.

IV]. It is verified that $x(t, \xi)$ satisfies (4.7) in the sense of distribution (cf. Lions and Magenes [19, pp. 292–294]). Next we consider the *strong* solution of (4.7)–(4.9). Let $x_0, y_0 \in D(A)$. Then by (4.13) and (4.14), we have

$$|\langle x_0, \phi_n\rangle_{[0,1]}\phi_n(\xi)| \leqq C_1\|Ax_0\|_{[0,1]}/n^2,$$

$$|\langle y_0, \phi_n\rangle_{[0,1]}\phi_n(\xi)| \leqq C_2\|Ay_0\|_{[0,1]}/n^2$$

for large n, where C_1, C_2 are constants not depending on n. Hence the substitution and elsewhere $x(t, \xi)$ of ξ in (4.11) has sense and converges uniformly on any compact set in R^+ for all $\xi \in [0, 1]$. In this case the mild solution $x(t, \cdot)$ is the *strong* solution of S (see Remark 2.2 and § 3). The function $x(t, \xi)$ on $R^+ \times [0, 1]$ is also called the *strong* solution of (4.7)–(4.9) (in the sense of Ito). It would be interesting to note that the *strong* solution $x(t, \xi)$ is the classical solution of (4.7)–(4.9) if $x_0, y_0 \in D(A^2)$. The analogous fact holds for the model solution.

We now consider the two different types of observation on $J \subset R^+$:

(i) *Observation by distributed measurement*:

$$y(t) = \langle w, x(t, \cdot)\rangle_{[0,1]}, \qquad t \in J, \quad w \in L_2[0, 1]. \tag{4.15}$$

(ii) *Observation by pointwise measurement*:

$$y(t) = x(t, \xi_p), \qquad t \in J, \quad \xi_p \in [0, 1]. \tag{4.16}$$

The type (i) observation is for the mild (or *weak*) solutions and type (ii) is for the *strong* solutions.

COROLLARY 4.1. *Let* $a = a^{mm}$.

In the system (4.7)–(4.9) *the eigenvalues* $\{\lambda_n : n = 1, 2, \cdots\}$ *of* A *given by* (4.10) *are identifiable on* J *of positive measure in the following cases*:

a) *If the observation is given by* (4.15) *and if*

$$\langle x_0^{mm}, \Phi_n\rangle_{[0,1]} \cdot \langle w, \Phi_n\rangle_{[0,1]} \neq 0 \quad \textit{for all } n$$

or

$$\langle y_0^{mm}, \Phi_n\rangle_{[0,1]} \cdot \langle w, \Phi_n\rangle_{[0,1]} \neq 0 \quad \textit{for all } n.$$

b) *If the observation is given by* (4.16), $x_0, y_0 \in D(A)$, $x_0^{mm}, y_0^{mm} \in D(A^{mm})$ *and if*

$$\langle x_0^{mm}, \Phi_n\rangle_{[0,1]}\Phi_n(\xi_p) \neq 0 \quad \textit{for all } n$$

or

$$\langle y_0^{mm}, \Phi_n\rangle_{[0,1]}\Phi_n(\xi_p) \neq 0 \quad \textit{for all } n.$$

Proof. In both cases the output error is bounded by some function of exponential order. Then, as in the proof of Theorem 4.2, the zero output error implies that

$$\sum_{n=1}^{\infty} c_n e^{\lambda_n t} = \sum_{n=1}^{\infty} c_n^{mm} e^{\Lambda_n t}, \qquad t > 0, \tag{4.17}$$

and

$$\sum_{n=1}^{\infty} d_n e^{\lambda_n t} = \sum_{n=1}^{\infty} d_n^{mm} e^{\Lambda_n t}, \qquad t > 0, \tag{4.18}$$

where $c_n = \langle x_0, \phi_n\rangle_{[0,1]} \cdot \langle w, \phi_n\rangle_{[0,1]}$, $c_n^{mm} = \langle x_0^{mm}, \Phi_n\rangle_{[0,1]} \cdot \langle w, \Phi_n\rangle_{[0,1]}$, $d_n = \langle y_0, \phi_n\rangle_{[0,1]} \cdot \langle w, \phi_n\rangle_{[0,1]}$, $d_n^{mm} = \langle y_0^{mm}, \Phi_n\rangle_{[0,1]} \cdot \langle w, \Phi_n\rangle_{[0,1]}$ in Case a), and $c_n = \langle x_0, \phi_n\rangle_{[0,1]}\phi_n(\xi_p)$, $c_n^{mm} = \langle x_0^{mm}, \Phi_n\rangle_{[0,1]}\Phi_n(\xi_p)$, $d_n = \langle y_0, \phi_n\rangle_{[0,1]}\phi_n(\xi_p)$, $d_n^{mm} = \langle y_0^{mm}, \Phi_n\rangle_{[0,1]}\Phi_n(\xi_p)$ in Case b).

If $c_n''' \neq 0$ for all n or $d_n''' \neq 0$ for all n, then by (4.17) or (4.18) we have $\{\Lambda_n : n = 1, 2, \cdots\} \subset \{\lambda_n : n = 1, 2, \cdots\}$.

This inclusion and the asymptotics (4.14) of λ_n and Λ_n mean that $\lambda_n = \Lambda_n$ for all n, since $C = C'''$.

In Case b) the assumption of unknowns that $x_0, y_0 \in D(A)$ is required. But the assumption is satisfied if $x_0, y_0 \in C_0^2[0, 1]$ (or, α_0, α_1 are known, $x_0, y_0 \in C^2[0, 1]$ and x_0, y_0 satisfy the boundary condition (4.8)).

In what follows we assume the following condition (VI).

(VI) $b, \alpha_0, \alpha_1, x_0, y_0$ are unknown but it is known that $a(\xi) = 1$ for all $\xi \in [0, 1]$, $b(\cdot) \in C^1[0, 1]$, $0 \leqq \alpha_0, \alpha_1 < 1$ and $x_0, y_0 \in D(A)$ (or $x_0, y_0 \in C_0^2[0, 1]$).

For the *strong* solution of (4.7)–(4.9) we consider the following observation on J of positive measure:

(iii). *Observation from boundary*:

$$y(t) = \{x(t, 0), x(t, 1)\}, \qquad t \in J. \tag{4.19}$$

Differently from the above corollary it is assumed that the eigenfunctions ϕ_n, Φ_n are normalized so that

$$\phi_n(0) = \Phi_n(0) = 1 \qquad (\text{since } \alpha_0, \alpha_1, \alpha_0''', \alpha_1''' < 1).$$

We put $\rho_n = \|\phi_n\|_{[0,1]}$, $\kappa_n = \|\Phi_n\|_{[0,1]}$, $n = 1, 2, \cdots$. Let $\langle x_0''', \Phi_n\rangle_{[0,1]} \neq 0$ for all n or $\langle y_0''', \Phi_n\rangle_{[0,1]} \neq 0$ for all n. Then as in Corollary 4.1 we see that the zero output error by the type (iii) observation implies that

$$\lambda_n = \Lambda_n, \quad \phi_n(1) = \Phi_n(1), \quad n = 1, 2, \cdots, \tag{4.20}$$

and

$$\langle x_0, \phi_n\rangle_{[0,1]}/\rho_n = \langle x_0''', \Phi_n\rangle_{[0,1]}/\kappa_n, \langle y_0, \phi_n\rangle_{[0,1]}/\rho_n = \langle y_0''', \Phi_n\rangle_{[0,1]}/\kappa_n, \quad n = 1, 2, \cdots. \tag{4.21}$$

As shown in Murayama [20] it follows from (4.20) that $\rho_n = \kappa_n$, $n = 1, 2, \cdots$. Since the spectral characteristics $\{\lambda_n, \rho_n : n = 1, 2, \cdots\}$ determine $b(\xi)$ and α_0, α_1 uniquely (Gel'fand–Levitan theory [9], [23, p. 498–499]), we have by (4.20) that

$$b(\xi) = b'''(\xi) \quad \text{for all } \xi \in [0, 1] \quad \text{and} \quad \alpha_0 = \alpha_0''', \quad \alpha_1 = \alpha_1'''.$$

Thus $\{\phi_n : n = 1, 2, \cdots\} = \{\Phi_n : n = 1, 2, \cdots\}$, and hence by (4.21),

$$x_0(\xi) = x_0'''(\xi), \quad y_0(\xi) = y_0'''(\xi) \quad \text{for all } \xi \in [0, 1].$$

Therefore we have the following corollary:

COROLLARY 4.2. *Let the condition* (VI) *be satisfied in the system* (4.7)–(4.9) *and let the observation be given by* (4.19). *Then the coefficient* $b(\xi)$, *the boundary coefficients* α_0, α_1 *and the initial values* $x_0(\xi)$, $y_0(\xi)$ *are identifiable on* J *of positive measure if* $\langle x_0''', \Phi_n\rangle_{[0,1]} \neq 0$ *for all* n *or* $\langle y_0''', \Phi_n\rangle_{[0,1]} \neq 0$ *for all* n.

We now consider the identifiability of constant parameters in operators. In the systems S_1 and S_2, it is assumed that the initial values x_0, y_0 are known and the unknown operator A has the form

$$A = aA_0 + b. \tag{4.22}$$

Here a and b are unknown real constants, but it is known that $a > 0$, $b \in R$ and A_0 is an a priori known operator satisfying (H). We denote the set of eigenvalues and eigenfunctions of A_0 by $\{\tau_n, \psi_{nj} : j = 1, \cdots, r_n, n = 1, 2, \cdots\}$. It is also assumed that

A_0 in S_j satisfies I(i) if $j=1$ and I(ii) if $j=2$, so that A in S_1 satisfies I(i) and A in S_2 satisfies I(ii).

The new feature here is that we can test a priori whether or not the system is identifiable, because we already know A_0 and thus its projections P_n^0 (see Theorems 4.2, 4.3 and Example 4.2 below). This is not the case for the previous theorems.

In the following we consider some restrictive class of observation operators. Let the observation operator B satisfy (3.10) with $A=A_0$ (we write this by $B\in LF(D(A_0^k))$). For any $(a,b)\in(R^+-\{0\})\times R$ and fixed $x_0\in X$ (resp. $x_0,\ y_0\in D(A_0^k)$), $f(\cdot)\in H_k^{\mathrm{loc}}(R^+;X)$, the output $y_1(a,b;t)=Bx_1(a,b;t)$ (resp. $y_2(a,b;t)=Bx_2(a,b;t)$) defines a continuous function on $R^+-\{0\}$ (resp. on R^+), where $x_j(a,b;t)$ $(j=1,2)$ is the mild solution of S_j with $A=aA_0+b$. For any interval J we define the nonlinear map

$$K_J^j\colon (R^+-\{0\})\times R\to C(J) \quad \text{by}$$

$$K_J^j(a,b)(s)=y_j(a,b;s), \qquad s\in J \quad (j=1,2).$$

The following definition does not require any information on the model systems.

DEFINITION 4.2. The pair of parameters (a,b) in S_j is said to be *identifiable* on J if K_J^j is an injection $(j=1,2)$.

The above definition implies that the parameters a, b in S_j are identifiable on J if and only if $a=a'$ in $R^+-\{0\}$ and $b=b'$ in R follows from

$$e(S_j,S_j';t)=K_J^j(a,b)(t)-K_J^j(a',b')(t)=0, \qquad t\in J.$$

Let $f=0$ in $L_2^{\mathrm{loc}}(R^+;X)$ and $B\in LF(D(A_0^k))$. Then it follows from Propositions 3.3 and 3.4 and (4.22) that

$$e(S_1,S_1';t)=\sum_{n=1}^{\infty}B(P_n^0x_0)(e^{\mu_n t}-e^{\mu'_n t}), \qquad t>0, \tag{4.23}$$

and

$$\begin{aligned} e(S_2,S_2';t)=&\sum_{n=1}^{\infty}B(P_n^0x_0)(\cos\sqrt{-\mu_n}t-\cos\sqrt{-\mu'_n}t)\\ &+\sum_{n=1}^{\infty}B(P_n^0y_0)\left(\frac{\sin\sqrt{-\mu_n}t}{\sqrt{-\mu_n}}-\frac{\sin\sqrt{-\mu'_n}t}{\sqrt{-\mu'_n}}\right), \qquad t\in R^+, \end{aligned} \tag{4.24}$$

where $\mu_n=a\tau_n+b$, $\mu'_n=a'\tau_n+b'$ and $P_n^0x=\sum_{j=1}^{r_n}\langle x,\psi_{nj}\rangle\psi_{nj}$, $x\in X$, $n=1,2,\cdots$. In (4.24) it is assumed that $x_0,\ y_0\in D(A_0^k)$, so that $\sum_{n=1}^{\infty}|B(P_n^0x_0)|<\infty$, $\sum_{n=1}^{\infty}|B(P_n^0y_0)|<\infty$.

Concerning the output errors (4.23), (4.24), we have the following propositions:

PROPOSITION 4.1. *Let $\{\mu_n\}_{n=1}^{\infty}$, $\{\mu'_n\}_{n=1}^{\infty}$ be strictly monotone decreasing sequences, and let $\sum_{n=1}^{\infty}c_ne^{\mu_n t}$, $\sum_{n=1}^{\infty}c_ne^{\mu'_n t}$ converge uniformly on $[\delta,\infty)$ for each $\delta>0$. Then*

$$\sum_{n=1}^{\infty}c_n(e^{\mu_n t}-e^{\mu'_n t})=0, \qquad t>0, \tag{4.25}$$

if and only if

$$c_n(\mu_n-\mu'_n)=0, \qquad n=1,2,\cdots. \tag{4.26}$$

PROPOSITION 4.2. *Let* $\sum_{n=1}^{\infty}|c_n|<\infty$ *and* $\sum_{n=1}^{\infty}|d_n|<\infty$ *and let* $\{\mu_n\}_{n=1}^{\infty}$ *and* $\{\mu_n'\}_{n=1}^{\infty}$ *be strictly monotone decreasing sequences. Then*

$$\sum_{n=1}^{\infty} c_n(\cos\sqrt{-\mu_n}t-\cos\sqrt{-\mu_n'}t) + \sum_{n=1}^{\infty} d_n\left(\frac{\sin\sqrt{-\mu_n}t}{\sqrt{-\mu_n}}-\frac{\sin\sqrt{-\mu_n'}t}{\sqrt{-\mu_n'}}\right)=0, \qquad t\geqq 0, \tag{4.27}$$

if and only if

$$c_n(\mu_n-\mu_n')=d_n(\mu_n-\mu_n')=0, \qquad n=1,2,\cdots. \tag{4.28}$$

Proof of Proposition 4.1. This proposition is first stated in [16, Lemma 3] but the proof is not given there. We give here a simple proof which is similar to [30, Rem. 3.5]. It is sufficient to show that (4.25) implies (4.26). First we prove that $c_1(\mu_1-\mu_1')=0$ by contradiction. Suppose on the contrary that $c_1\neq 0$ and $\mu_1\neq\mu_1'$ (we let $\mu_1>\mu_1'$). Multiplying both sides of (4.25) by $e^{-\mu_1 t}$, we obtain

$$c_1-c_1e^{(\mu_1'-\mu_1)t}+\sum_{n=2}^{\infty} c_n(e^{(\mu_n-\mu_1)t}-e^{(\mu_n'-\mu_1)t})=0. \tag{4.29}$$

The second term of (4.29) converges to 0 as $t\to\infty$. Since the third term of (4.29) converges uniformly on $[1,\infty)$ and each subterm of the third term converges to 0 because of $\mu_n<\mu_1$ for all $n\geqq 2$ and $\mu_n'<\mu_1$ for all $n\geqq 1$, then the third term itself converges to 0 as $t\to\infty$. Then we have $c_1=0$, which contradicts the assumption. Next suppose that $c_1(\mu_1-\mu_1')=\cdots=c_m(\mu_m-\mu_m')=0$, $m\geqq 2$. Then the sum of first m terms of (4.25) equals zero, and hence by similar arguments as above we see that $c_{m+1}(\mu_{m+1}-\mu_{m+1}')=0$. This proves (4.26) by mathematical induction.

Proof of Proposition 4.2. Since $\sum_{n=1}^{\infty}|c_n|<\infty$, the first trigonometric series in (4.27) is bounded by a function of exponential order. Then as in the proof of Theorem 4.2, we see that (4.27) implies

$$\sum_{n=1}^{\infty} c_n(e^{\mu_n t}-e^{\mu_n' t})=0, \quad \sum_{n=1}^{\infty} d_n(e^{\mu_n t}-e^{\mu_n' t})=0, \quad t>0. \tag{4.30}$$

Hence as shown in Proposition 4.1, (4.28) follows from (4.30).

For any x, $y\in X$, we define $\operatorname{rank}\{B(P_n^0x)\colon n=1,2,\cdots\}$ and $\operatorname{rank}\{B(P_n^0x), B(P_n^0y)\colon n=1,2,\cdots\}$ by $\#\{n\colon B(P_n^0x)\neq 0\}$ and $\#\{n\colon B(P_n^0x)\neq 0$ or $B(P_n^0y)\neq 0\}$, respectively. Here $\#\{\cdots\}$ denotes the potency of the set $\{\cdots\}$, i.e., the number of elements in $\{\cdots\}$ (this permits ∞).

The following Theorem 4.3 extends our previous results [16, § 4][1] to abstract systems of first order.

THEOREM 4.3. *Let* A *in* S_1 *have the form* (4.22) *and let* $B\in LF(D(A_0^k))$.

i) *If* $x_0\in X$ *is known and* $f(\cdot)=0$ *in* $L_2^{\mathrm{loc}}(R^+;X)$, *then the pair of parameters* (a,b) *in* S_1 *is identifiable on* J *of positive measure if and only if*

$$\operatorname{rank}\{B(P_n^0x_0)\colon n=1,2,\cdots\}\geqq 2. \tag{4.31}$$

ii) *If* $x_0=0$ *in* X *and* f *has the form* $z_0g(t)$, $z_0\in D(A_0^k)$, $g(\cdot)\in L_2^{\mathrm{loc}}(R^+)$ *such that* $g(\cdot)\neq 0$ *in* $L_2[0,T]$, $T>0$ (*resp.* $g(\cdot)\neq 0$ *in* $L_2^{\mathrm{loc}}(R^+)$), *then the pair of parameters*

[1] [16, Result 12, p. 798] is incorrect. A corrected statement of the result was given by Courdesses [33].

(a, b) in S_1 is identifiable on $[0, T]$ (resp. on R^+) if and only if

$$\text{rank}\,\{B(P_n^0 z_0)\colon n = 1, 2, \cdots\} \geqq 2. \tag{4.32}$$

Proof. Case i): We shall show the "if" part. Let (4.31) be satisfied. It then follows that there exist two distinct numbers n_1, n_2 such that $B(P_{n_1}^0 x_0) \neq 0$, $B(P_{n_2}^0 x_0) \neq 0$. Assume that $e(S_1, S_1'; t) = 0$, $t \in J$. Then by (4.23), meas $(J) > 0$ and Proposition 4.1, we have $\mu_{n_1} = \mu'_{n_1}$ and $\mu_{n_2} = \mu'_{n_2}$. Since $\tau_{n_1} \neq \tau_{n_2}$, these equalities imply $(a, b) = (a', b')$. This shows the identifiability of the pair (a, b). The proof of the "only if" part is the same as given in [16, Result 10].

THEOREM 4.4. *Let A in S_2 have the form* (4.22) *and let $B \in LF(D(A_0^k))$.*

i) *If x_0, $y_0 \in D(A_0^k)$ are known and $f(\cdot) = 0$ in $L_2^{\text{loc}}(R^+; X)$, then the pair of parameters (a, b) in S_2 is identifiable on J of positive measure if and only if*

$$\text{rank}\,\{B(P_n^0 x_0), B(P_n^0 y_0)\colon n = 1, 2, \cdots\} \geqq 2; \tag{4.33}$$

ii) *If $x_0 = y_0 = 0$ in X and f satisfies the same condition as in Case* ii) *of Theorem* 4.3, *then the same conclusion holds for the system S_2.*

This theorem follows from Proposition 4.2 and the representation (4.24) of the output error $e(S_2, S_2'; t)$.

It is worth noting that the finiteness of the multiplicity of A is not necessarily required for the identifiability in Theorems 4.1–4.4 as distinct from the case for finite controllability as studied in Fattorini [7].

Remark 4.1. Consider the identifiability of $l+1$ constant parameters a_0, $a_1, \cdots, a_l$ in the operator $A = -\sum_{i=0}^{l} a_i(-A_0)^{k_i}$. Assuming that $(a_0, a_1, \cdots, a_l) \in R \times (R^+)^{l-1} \times (R^+ - \{0\})$, A_0 is negative and satisfies (H) and I(i) or I(ii), $0 = k_0 < k_1 < \cdots < k_l$ are known constants, Theorems 4.3 and 4.4, in which the pair (a, b) and the (rank) number 2 in (4.31), (4.32), (4.33) are replaced by $(a_0, a_1, \cdots, a_l)$ and $l+1$, are also true.

Example 4.2. Let G be a bounded domain in R^N with smooth boundary ∂G. We consider, on the domain G, the following system described by the initial boundary value problem of a parabolic partial differential equation:

$$\frac{\partial x}{\partial t}(t, \xi) = a\Delta x(t, \xi) + bx(t, \xi), \qquad \xi \in G, \quad t > 0, \quad (a > 0, b\text{: constants}), \tag{4.34}$$

$$\alpha(\eta)x(t, \eta) + (1 - \alpha(\eta))\frac{\partial x}{\partial n}(t, \eta) = 0, \qquad \eta \in \partial G, \quad t > 0, \tag{4.35}$$

$$x(0, \xi) = x_0(\xi), \qquad \xi \in G. \tag{4.36}$$

Here Δ is the Laplace operator and $\partial/\partial n$ denotes the outward normal derivative. It is assumed that $\alpha(\cdot) \in C^\infty(\partial G)$ and $0 \leqq \alpha \leqq 1$ everywhere on ∂G. For the system (4.34)–(4.36) we assume that the initial value $x_0 \in L_2(G)$ and the boundary coefficient $\alpha(\eta)$ are a priori known while the diffusion coefficient $a > 0$ and the radiation coefficient $b \in R$ (or the size $b \in R$ of a uniform feedback $Fx = bx$) are unknown. We denote by A_0 the realization in $L_2(G)$ of Δ under the boundary condition (4.35). Since A_0 satisfies (H) and I(i) (also, I(ii)), there exists a set of eigenvalues and eigenfunctions $\{\tau_n, \psi_{nj}\colon j = 1, \cdots, r_n, n = 1, 2, \cdots\}$ (known quantities!) of A_0. That $\psi_{nj} \in C^\infty(\bar{G})$ for all n, j is also known [1], [12], [13]. The function

$$x(t, \xi) = \sum_{n=1}^{\infty} e^{(a\tau_n + b)t} \sum_{j=1}^{r_n} \langle x_0, \psi_{nj}\rangle_G \psi_{nj}(\xi), \qquad t > 0, \quad \xi \in \bar{G}, \tag{4.37}$$

is a *weak* solution of (4.34)–(4.36) in the sense of Ito [13] or Ladyženskaya, Solonnikov and Ural'ceva [17]. It is shown in [13, Chap. II] that $x(t, \xi)$ in (4.37) converges uniformly on $[\delta, \infty) \times \bar{G}$ for each $\delta > 0$ and satisfies (4.34) and (4.35) in classical sense. By considering this *weak* solution $x(t, \xi)$ as a function $x(t, \cdot) \in L_2(G)$ in $t > 0$, we see easily that $x(t) = x(t, \cdot)$ is the mild solution of the following evolution equation in $L_2(G)$:

$$\dot{x}(t) = aA_0 x(t) + bx(t), \qquad x(0) = x_0.$$

Here parameters a, b are to be determined uniquely. It would be evident that $x(t) \in \bigcap_{n=1}^{\infty} D(A_0^n)$ for all $t > 0$.

We now consider the following two types of observation on J of positive measure as in Example 4.1.

(i) *Observation by distributed measurement*:

$$y(t) = \langle w_0, x(t, \cdot) \rangle_G, \qquad t \in J \subset R^+, \quad w_0 \in L_2(G).$$

(ii) *Observation by pointwise measurement*:

$$y(t) = x(t, \xi_p), \qquad t \in J \subset R^+ - \{0\}, \quad \xi_p \in \bar{G}.$$

Directly from Theorem 4.3 i) with $k = 0$, we obtain the following result for the type (i) observation.

The identifiability condition of the constant parameters a and b on J for the type (i) observation is that

C(i) There exist two distinct n_1 and n_2 such that

$$\sum_{j=1}^{r_{n1}} \langle x_0, \psi_{n_1 j} \rangle_G \cdot \langle w_0, \psi_{n_1 j} \rangle_G \neq 0 \quad \text{and} \quad \sum_{j=1}^{r_{n2}} \langle x_0, \psi_{n_2 j} \rangle_G \cdot \langle w_0, \psi_{n_2 j} \rangle_G \neq 0.$$

To give the identifiability condition of parameters a, b for the type (ii) observation, we need some preparation. First we cite an important inequality from [13, Chap. IV]:

For any $x(\cdot) \in D(A_0^k)$, $k \geqq [N/4] + 1$, $x(\cdot) \in C(\bar{G})$ and

$$\sup_{\xi \in \bar{G}} |x(\xi)| \leqq \begin{cases} C_1 \|A_0^k x\|_G & \text{if } \alpha(\xi) \not\equiv 0, \\ C_2(\|x\|_G + \|A_0^k x\|_G) & \text{if } \alpha(\xi) \equiv 0, \end{cases} \tag{4.38}$$

where [ ] denotes Gauss's symbol and C_1, C_2 are constants depending only on k, N, α and G.

The inequality (4.38) is proved by using the estimates of the Green function of Δ ($\alpha(\xi) \not\equiv 0$) and the Neumann function of Δ ($\alpha(\xi) \equiv 0$). By (4.38) and $A_0^k \psi_{nj} = \tau_n^k \psi_{nj}$, we have

$$\sup_{\xi \in \bar{G}} |\psi_{nj}(\xi)| \leqq C_3(1 + |\tau_n|^k) \quad \text{for } k \geqq \left[\frac{N}{4}\right] + 1,$$

where $C_3 = \max\{C_1, C_2\}$. Then

$$\sum_{n=1}^{\infty} e^{(a\tau_n + b)t} \cdot \sum_{j=1}^{r_n} |\langle x_0, \psi_{nj} \rangle_G| \cdot |\psi_{nj}(\xi)| \leqq C_3 \|x_0\|_G \cdot \sum_{n=1}^{\infty} (1 + |\tau_n|^k) e^{(a\tau_n + b)t}.$$

This shows that the series in (4.37) converges uniformly on $[\delta, \infty)$ for each $\delta > 0$. By the way from (4.38), the observation operator B given by

$$Bx(t, \cdot) = x(t, \xi_p), \qquad t > 0, \quad \xi_p \in \bar{G}$$

is a continuous linear functional on $D(A_0^k)$. Notice that the output $x(t, \xi_p)$ makes sense for all $\xi_p \in \bar{G}$ (not for a.e. $\xi_p \in \bar{G}$).

Applying Theorem 4.3 i) with $k \geqq [N/4]+1$, we conclude that the identifiability condition of the parameters a, b for the type (ii) observation is that:

C(ii) There exist two distinct n_1 and n_2 such that

$$\sum_{j=1}^{r_{n1}} \langle x_0, \psi_{n1j}\rangle_G \psi_{n1j}(\xi_p) \neq 0 \quad \text{and} \quad \sum_{j=1}^{r_{n2}} \langle x_0, \psi_{n2j}\rangle_G \psi_{n2j}(\xi_p) \neq 0.$$

Consider the following special system. Let G be the circle $\{\xi: |\xi|<1\}$ and let the boundary data $\alpha = 1$ everywhere on $\{\eta: |\eta| = 1\}$ (Dirichlet type). It is assumed that the initial value x_0 is known and belongs to $L_2(G)$. Denote by A_0 the realization of the Laplace operator Δ in $L_2(G)$ under the Dirichlet boundary condition. In this system, the eigenfunctions of A_0 are given by

$$\psi_{m,n}(r, \theta) = c_{m,n} J_m(\lambda_{m,n} r) \begin{cases} \cos m\theta, \\ \sin m\theta, \end{cases} \qquad m = 0, 1, 2, \cdots, \quad n = 1, 2, \cdots,$$

in polar coordinates and the corresponding eigenvalues are $\lambda_{m,n}$, $m = 0, 1, 2, \cdots$, $n = 1, 2, \cdots$, where $J_m(r)$ is the Bessel function of order m and $\lambda_{m,n}$ is its nth positive zero and $c_{m,n}$ are constants for normalization. Note that $\lambda_{m,n}$ is a double eigenvalue for $m \geqq 1$.

Then the conditions C(i) and C(ii) are given respectively by

$$\int_0^{2\pi}\int_0^1 r x_0(r, \theta) J_m(\lambda_{m,n} r) \cos m\theta \, dr \, d\theta \cdot \int_0^{2\pi}\int_0^1 r w_0(r, \theta) J_m(\lambda_{m,n} r) \cos m\theta \, dr \, d\theta$$
$$+ \int_0^{2\pi}\int_0^1 r x_0(r, \theta) J_m(\lambda_{m,n} r) \sin m\theta \, dr \, d\theta \cdot \int_0^{2\pi}\int_0^1 r w_0(r, \theta) J_m(\lambda_{m,n} r) \sin m\theta \, dr \, d\theta \neq 0$$

for two distinct pairs of (m, n)

and

$$J_m(\lambda_{m,n} r_0) \cos m\theta_0 \cdot \int_0^{2\pi}\int_0^1 r x_0(r, \theta) J_m(\lambda_{m,n} r) \cos m\theta \, dr \, d\theta$$
$$+ J_m(\lambda_{m,n} r_0) \sin m\theta_0 \cdot \int_0^{2\pi}\int_0^1 r x_0(r, \theta) J_m(\lambda_{m,n} r) \sin m\theta \, dr \, d\theta \neq 0$$

for two distinct pairs of (m, n),

where $\xi_p = (r_0, \theta_0)$. The verification of the above conditions for given data $\{x_0, w_0\}$ or $\{x_0, (r_0, \theta_0)\}$ will be easy.

Remark 4.2. In Example 4.2, Theorem 4.3 can apply to obtain analogous identifiability conditions for other types of observations such as

$$y(t) = \int_G w_0(\xi) D_\xi^\beta x(t, \xi) \, d\xi, \qquad t \geqq 0, \quad w_0 \in L_2(G),$$

and

$$y(t) = D_\xi^\beta x(t, \xi_p), \qquad t > 0, \quad \xi_p \in \bar{G},$$

where $\beta = (\beta_1, \cdots, \beta_N)$ and $D_\xi^\beta = \partial^{\beta_1+\cdots+\beta_N}/(\partial\xi_1)^{\beta_1} \cdots (\partial\xi_N)^{\beta_N}$.

Example 4.3. On the domain G as in Example 4.2, we consider the following hyperbolic initial boundary value problem:

$$\frac{\partial^2 x}{\partial t^2}(t,\xi) = a\Delta x(t,\xi) + bx(t,\xi), \quad \xi \in G, \quad t>0, \quad (a>0, b: \text{constants}), \tag{4.39}$$

$$\alpha(\eta)x(t,\eta) + (1-\alpha(\eta))\frac{\partial x}{\partial n}(t,\eta) = 0, \qquad \eta \in \partial G, \quad t>0, \tag{4.40}$$

$$x(0,\xi) = x_0(\xi), \quad \frac{\partial x}{\partial t}(0,\xi) = y_0(\xi), \quad \xi \in G. \tag{4.41}$$

The unknowns are constants a, b and the situation is quite the same as in Example 4.2. The system (4.39)–(4.41) is represented by the second order evolution equation in $L_2(G)$;

$$\ddot{x}(t) = aA_0x(t) + bx(t), \qquad x(0) = x_0, \quad \dot{x}(0) = y_0. \tag{4.42}$$

Let x_0, $y_0 \in D(A_0^k)$ for $k \geqq [N/4]+1$. Then as seen in Example 4.2, the type (ii) observation is possible for the mild solution of (4.42). Applying Theorem 4.4 i), we have that the parameters a, b are identifiable on J of positive measure for the type (ii) observation if and only if at least one of the following conditions is satisfied:

1. $\sum_{j=1}^{r_n} \langle x_0, \psi_{nj}\rangle_G \psi_{nj}(\xi_p) \neq 0$ for two distinct n's;

2. $\sum_{j=1}^{r_n} \langle y_0, \psi_{nj}\rangle_G \psi_{nj}(\xi_p) \neq 0$ for two distinct n's;

3. there exist two distinct n_1 and n_2 such that

$$\sum_{j=1}^{r_{n_1}} \langle x_0, \psi_{nj}\rangle_G \psi_{nj}(\xi_p) \neq 0 \quad \text{and} \quad \sum_{j=1}^{r_{n_2}} \langle y_0, \psi_{nj}\rangle_G \psi_{nj}(\xi_p) \neq 0.$$

The similar identifiability condition of a, b is easily derived for the type (i) observation under the weaker assumption that x_0, $y_0 \in L_2(G)$.

5. Identifiability of operators and initial values. In this section we discuss the identifiability of the whole system S_j, i.e., of the operator A and the initial value(s) x_0 or x_0, y_0 in S_j, by using its model system S_j^{mm} $(j=1,2)$. Assuming that all quantities appearing in the model are known a priori, we derive some identifiability conditions of S_j in terms of known quantities such as the model eigenfunctions and the model initial values. To solve the problem in our general setting, we suppose that $B=I$, the identity operator on X $(Y=X)$. The assumption $B=I$ is not practical in application, but it is impossible in general that this assumption is weakened to that such as $Y=X$ and B is a compact operator. For certain restrictive classes of linear systems, however, the assumption can be weakened to the type (ii) or (iii) observation by using the theory of Gel'fand–Levitan as seen in Example 4.1 (see also [23, Thm. 2]).

Let $f=0$ in $L_2^{\text{loc}}(R^+; X)$. Since the output error $e(S_1, S_1^{mm}; t) = x_1(t) - x_1^{mm}(t)$ (resp. $e(S_2, S_2^{mm}; t) = x_2(t) - x_2^{mm}(t)$) depends on the initial values x_0 of S_1 and x_0^{mm} of S_1^{mm} (resp. the two pairs of initial values (x_0, y_0) of S_2 and (x_0^{mm}, y_0^{mm}) of S_2^{mm}), we denote this error by $e_1(x_0, x_0^{mm}; t)$ (resp. by $e_2(x_0, y_0, x_0^{mm}, y_0^{mm}; t)$).

The number of initial values as inputs has a close relation with the identifiability of A. This will be clarified by Theorems 5.1 and 5.2 given later.

Let $E_0=\{x_{0,1}, \cdots, x_{0,k}\}$, $E_0^m=\{x_{0,1}^m, \cdots, x_{0,k}^m\}\subset X$ (resp. $F_0=\{(x_{0,1}, y_{0,1}), \cdots, (x_{0,k}, y_{0,k})\}$, $F_0^m=\{(x_{0,1}^m, y_{0,1}^m), \cdots, (x_{0,k}^m, y_{0,k}^m)\}\subset X^2$) be the sets of initial values of S_1, S_1^m (resp. S_2, S_2^m). For these sets and $J\subset R^+$ we define the following conditions $C_J^k(E_0, E_0^m)$ and $C_J^k(F_0, F_0^m)$.

$$C_J^k(E_0, E_0^m): A=A^m \quad \text{and} \quad x_{0,1}=x_{0,1}^m, \cdots, x_{0,k}=x_{0,k}^m \quad \text{in } X$$

follow from

$$e_1(x_{0,1}, x_{0,1}^m; t)=0, \cdots, e_1(x_{0,k}, x_{0,k}^m; t)=0 \quad \text{in } X, \quad t\in J.$$

$$C_J^k(F_0, F_0^m): A=A^m \quad \text{and} \quad x_{0,1}=x_{0,1}^m, y_{0,1}=y_{0,1}^m, \cdots, x_{0,k}=x_{0,k}^m, y_{0,k}=y_{0,k}^m \quad \text{in } X$$

follow from

$$e_2(x_{0,1}, y_{0,1}, x_{0,1}^m, y_{0,1}^m; t)=0, \cdots, e_2(x_{0,k}, y_{0,k}, x_{0,k}^m, y_{0,k}^m; t)=0 \quad \text{in } X, \quad t\in J.$$

DEFINITION 5.1. The system S_1 (resp. S_2) is said to be *k-identifiable on J* with respect to the set E_0^m (resp. F_0^m) if $C_J^k(E_0, E_0^m)$ (resp. $C_J^k(F_0, F_0^m)$) is satisfied. If $C_J^k(E_0, E_0^m)$ (resp. $C_J^k(F_0, F_0^m)$) is satisfied for some sets of initial values E_0, E_0^m (resp. F_0, F_0^m), the system S_1 (resp. S_2) is said to be *identifiable* on J (more precisely, identifiable on J by using the model system S_1^m (resp. S_2^m)).

As in § 4 we denote the sets of eigenvalues and eigenfunctions of A and A^m by $\{\lambda_n, \phi_{nj}: j=1, \cdots, m_n, n=1, 2, \cdots\}$ and $\{\Lambda_n, \Phi_{nj}: j=1, \cdots, k_n, n=1, 2, \cdots\}$, respectively. Next we give the following concept.

DEFINITION 5.2. The set $E=\{x_1, \cdots, x_k\}$ in X is said to be *compatible* with respect to A^m if

$$\operatorname{rank} M_n=k_n \quad \text{for all } n=1, 2, \cdots.$$

The set $F=\{(x_1, y_1), \cdots, (x_k, y_k)\}$ in X^2 is said to be compatible with respect to A^m if

$$\operatorname{rank} M_n=k_n \quad \text{or} \quad \operatorname{rank} L_n=k_n \quad \text{for all } n=1, 2, \cdots.$$

Here the matrices M_n, L_n, $n=1, 2, \cdots$, are given by

$$M_n=\begin{pmatrix} \langle x_1, \Phi_{n1}\rangle & \cdots & \langle x_1, \Phi_{nk_n}\rangle \\ \langle x_2, \Phi_{n1}\rangle & \cdots & \langle x_2, \Phi_{nk_n}\rangle \\ \vdots & & \vdots \\ \langle x_k, \Phi_{n1}\rangle & \cdots & \langle x_k, \Phi_{nk_n}\rangle \end{pmatrix} \quad \text{and} \quad L_n=\begin{pmatrix} \langle y_1, \Phi_{n1}\rangle & \cdots & \langle y_1, \Phi_{nk_n}\rangle \\ \langle y_2, \Phi_{n1}\rangle & \cdots & \langle y_2, \Phi_{nk_n}\rangle \\ \vdots & \cdots & \vdots \\ \langle y_k, \Phi_{n1}\rangle & \cdots & \langle y_k, \Phi_{nk_n}\rangle \end{pmatrix}.$$

If E or F is compatible, then $k\geqq k_\infty=\sup\{k_n: n=1, 2, \cdots\}$. Conversely if the multiplicity k_∞ of A^m is finite, then we can choose a finite set E in X or F in X^2 such that E or F is compatible. From a practical point of view the number of initial values must be finite, but countably many zero output errors are needed to obtain the identifiability conditions for the case $k_\infty=\infty$.

THEOREM 5.1. *Let $f=0$ in $L_2^{\mathrm{loc}}(R^+; X)$ and let $J\subset R^+$ be of positive measure. If the multiplicity k_∞ of A^m is finite, then the system S_1 is identifiable on J (by using the model system S_1^m). And if the set $E_0^m=\{x_{0,1}^m, \cdots, x_{0,k}^m\}$ of k numbers of model initial values is compatible with respect to A^m, then the system S_1 is k-identifiable on J with respect to E_0^m.*

Proof. It is sufficient to show the latter half of the theorem. The output errors $e^i(t)=e_1(x_{0,i}, x_{0,i}^m; t)$, $i=1, \cdots, k$, are given by

$$e^i(t)=\sum_{n=1}^{\infty} e^{\lambda_n t} \sum_{j=1}^{m_n} \langle x_{0,i}, \phi_{nj}\rangle \phi_{nj} - \sum_{n=1}^{\infty} e^{\Lambda_n t} \sum_{j=1}^{k_n} \langle x_{0,i}^m, \Phi_{nj}\rangle \Phi_{nj}, \qquad i=1, \cdots, k.$$

Let $e^i(t)=0$ in X for all $t\in J$ and $i=1,\cdots,k$. Then for any fixed NJ, we have

$$\langle e^i(t),\phi_{NJ}\rangle = e^{\lambda_N t}\langle x_{0,i},\phi_{NJ}\rangle - \sum_{n=1}^{\infty} e^{\Lambda_n t}\sum_{j=1}^{k_n}\langle x_{0,i}^{m},\Phi_{nj}\rangle\langle\Phi_{nj},\phi_{NJ}\rangle = 0, \qquad t\in J. \tag{5.1}$$

Put $\langle x_{0,i},\phi_{NJ}\rangle = c^i_{NJ}$ and $\sum_{j=1}^{k_n}\langle x_{0,i}^{m},\Phi_{nj}\rangle\langle\Phi_{nj},\phi_{NJ}\rangle = c^i_{NJ}(n)$. Since J is of positive measure and the series in (5.1) is analytic on $R^+-\{0\}$, we have that

$$c^i_{NJ}e^{\lambda_N t} - \sum_{n=1}^{\infty} c^i_{NJ}(n)e^{\Lambda_n t} = 0, \qquad t>0, \quad i=1,\cdots,k. \tag{5.2}$$

If $\lambda_N\notin\{\Lambda_n\}_{n=1}^{\infty}$, then we see from (5.2) that

$$c^i_{NJ}(n) = \sum_{j=1}^{k_n}\langle x_{0,i}^{m},\Phi_{nj}\rangle\langle\Phi_{nj},\phi_{NJ}\rangle = 0 \quad \text{for all } n=1,2,\cdots, \quad i=1,\cdots,k. \tag{5.3}$$

The equalities (5.3) can be written by the following matrix equations

$$\begin{pmatrix}\langle x_{0,1}^{m},\Phi_{n1}\rangle & \cdots & \langle x_{0,1}^{m},\Phi_{nk_n}\rangle\\ \langle x_{0,2}^{m},\Phi_{n1}\rangle & \cdots & \langle x_{0,2}^{m},\Phi_{nk_n}\rangle\\ \vdots & & \vdots\\ \langle x_{0,k}^{m},\Phi_{n1}\rangle & \cdots & \langle x_{0,k}^{m},\Phi_{nk_n}\rangle\end{pmatrix}\begin{pmatrix}\langle\Phi_{n1},\phi_{NJ}\rangle\\ \langle\Phi_{n2},\phi_{NJ}\rangle\\ \vdots\\ \langle\Phi_{nk_n},\phi_{NJ}\rangle\end{pmatrix} = \begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}, \quad n=1,2,\cdots. \tag{5.4}$$

Since E_0^{m} is compatible with respect to A^{m}, the equations (5.4) imply that $0=\langle\phi_{NJ},\Phi_{nj}\rangle=\overline{\langle\Phi_{nj},\phi_{NJ}\rangle}$ (the complex conjugate) for all n,j. This contradicts that $\phi_{NJ}\neq 0$ in X, and this shows $\lambda_N\in\{\Lambda_n\}_{n=1}^{\infty}$. Hence it follows from (5.2) that there exists a natural number N^* (for fixed N) such that

$$\lambda_N = \Lambda_{N^*}, \tag{5.5}$$

$$c^i_{NJ}(n)=0 \quad \text{for all } n\neq N^* \text{ and } i=1,\cdots,k. \tag{5.6}$$

It is easy to see that (5.6) is equivalent to $\phi_{NJ}\in X_{N^*}^{m}=\operatorname{span}\{\Phi_{N^*1},\cdots,\Phi_{N^*k_{N^*}}\}$. Since NJ is arbitrary, we have from (5.5) and (5.6) that $\{\lambda_n\}_{n=1}^{\infty}\subset\{\Lambda_n\}_{n=1}^{\infty}$ and the inclusions $X_n\subset X_{n^*}^{m}$ hold for all n. We can assume that the correspondence $n\mapsto n^*$ is monotone increasing by changing the order of $\{n^*\}$ if necessary. The completeness of $\{\phi_{nj}: j=1,\cdots,m_n,\ n=1,2,\cdots\}$ implies $X=\oplus_{n=1}^{\infty}X_n$ (the direct sum), and hence $X=\oplus_{n=1}^{\infty}X_{n^*}^{m}$. Since $X=\oplus_{n=1}^{\infty}X_n^{m}$ is obvious, we have $\{n\}=\{n^*\}$, so that $\lambda_n=\Lambda_n$ and $X_n=X_n^{m}$ for all n. Then

$$D(A)=\left\{x\in X:\sum_{n=1}^{\infty}|\lambda_n|^2\sum_{j=1}^{m_n}|\langle x,\phi_{nj}\rangle|^2<\infty\right\}$$

$$=\left\{x\in X:\sum_{n=1}^{\infty}|\Lambda_n|^2\sum_{j=1}^{k_n}|\langle x,\Phi_{nj}\rangle|^2<\infty\right\}=D(A^{m}).$$

This follows since $\lambda_n=\Lambda_n$, $m_n=k_n$ and $\sum_{p=1}^{k_n}\langle\phi_{nj},\Phi_{nj}\rangle\langle\phi_{ni},\Phi_{np}\rangle=\delta_{j,i}$ (Kronecker's delta), $j,i=1,\cdots,k_n$, $n=1,2,\cdots$. Furthermore since Φ_{nj} can be expanded as $\Phi_{nj}=\sum_{p=1}^{k_n}\langle\Phi_{nj},\phi_{np}\rangle\phi_{np}$, we have $A\Phi_{nj}=\sum_{p=1}^{m_n}\langle\Phi_{nj},\phi_{np}\rangle A\phi_{np}=\Lambda_n\Phi_{nj}$. Therefore $(A-A^{m})\Phi_{nj}=0$ in X for all n,j, which means that $A=A^{m}$. Now we can assume without loss of generality that $\phi_{nj}=\Phi_{nj}$ for all n,j. Then the zero output errors $e^i(t)=0$ in X, $t>0$, $i=1,\cdots,k$, imply $\langle x_{0,i}-x_{0,i}^{m},\Phi_{nj}\rangle=0$ for all n,j and $i=1,\cdots,k$. This shows that $x_{0,1}=x_{0,1}^{m},\cdots,x_{0,k}=x_{0,k}^{m}$ in X. This completes the proof.

If the coefficient matrix in (5.4) is the zero matrix for some n_0, then the non-identifiability of S_1 on $R^+-\{0\}$ follows. To show this let the operator A be given by

$$Ax = \sum_{n=1}^{\infty} \sum_{j=1}^{k_n} \mu_n \langle x, \Phi_{nj}\rangle \Phi_{nj}, \qquad x \in D(A),$$

$$D(A) = \left\{ x \in X: \sum_{n=1}^{\infty} |\mu_n|^2 \sum_{j=1}^{k_n} |\langle x, \Phi_{nj}\rangle|^2 < \infty \right\},$$

where $\mu_n = \Lambda_n$ if $n \neq n_0$ and $\mu_{n_0} = \frac{1}{2}(\Lambda_{n_0-1} + \Lambda_{n_0})$. A satisfies (H) and I(i), since μ_n are reals, $1/|\lambda - \mu_n| \to 0$ as $n \to \infty$ if $\lambda \neq \mu_n$ for all n and A generates the semigroup

$$T(t)x = \sum_{n=1}^{\infty} e^{\mu_n t} \sum_{j=1}^{k_n} \langle x, \Phi_{nj}\rangle \Phi_{nj}, \qquad x \in X.$$

Clearly $A \neq A'''$. Let $x_{0,1} = x'''_{0,1}, \cdots, x_{0,k} = x'''_{0,k}$. Then the condition $C_J^k(E_0, E_0''')$ with $J = R^+ - \{0\}$ does not hold. Thus the system S_1 is not identifiable on any $J \subset R^+ - \{0\}$. Therefore when $k_\infty = 1$, the necessary and sufficient condition for 1-identifiability of S_1 is that $\langle x_0''', \Phi_0\rangle \neq 0$ for all n (here we write x_0''', Φ_n instead of $x'''_{0,1}$, Φ_{n1}).

It is an open question, however, whether the compatibility of E_0''' is necessary or not for k-identifiability ($k \geqq 2$) of S_1 when $k_\infty \geqq 2$.

THEOREM 5.2. *Let $f = 0$ in $L_2^{\mathrm{loc}}(R^+; X)$ and let $J \subset R^+$ be of positive measure. If the multiplicity k_∞ of A''' is finite, then the system S_2 is identifiable on J (by using the model system S_2'''). And if the set $F_0''' = \{(x'''_{0,1}, y'''_{0,1}), \cdots, (x'''_{0,k}, y'''_{0,k})\}$ of k pairs of initial values is compatible with respect to A''', then the system S_2 is k-identifiable on J with respect to F_0'''.*

Proof. The proof is omitted.

Remark 5.1. Theorems 5.1 and 5.2 hold even under the condition that A and A''' satisfy the next assumption (H_0) which is weaker than (H):

(H_0). A has a compact normal resolvent.

This can be verified easily by a slight modification of arguments in their proofs (cf. [30], [31]).

Example 5.1. Let G be a bounded domain with smooth (of C^∞-class) boundary ∂G. We consider, on G, the parabolic distributed system described by the equations

$$\frac{\partial x}{\partial t}(t, \xi) = \mathscr{L}x(t, \xi) + f_i(t, \xi), \qquad \xi \in G, \quad t > 0 \quad (i = 1, \cdots, k), \tag{5.7}$$

$$\alpha(\eta)x(t, \eta) + (1 - \alpha(\eta))\frac{\partial x}{\partial \nu}(t, \eta) = 0, \qquad \eta \in \partial G, \quad t > 0, \tag{5.8}$$

$$x(0, \xi) = x_{0,i}(\xi), \qquad \xi \in G \quad (i = 1, \cdots, k). \tag{5.9}$$

Here $\mathscr{L}$ is a second order, formally selfadjoint differential operator of the form

$$\mathscr{L} = \sum_{i,j=1}^{N} \frac{\partial}{\partial \xi_i}\left(a_{ij}(\xi)\frac{\partial}{\partial \xi_j}\right) + c(\xi), \qquad \xi = (\xi_1, \cdots, \xi_N) \in G \tag{5.10}$$

with $a_{ij}(\cdot), c(\cdot) \in C^\infty(\bar{G})$ and $a_{ij} = a_{ji}$ for all i, j, and $\mathscr{L}$ is uniformly elliptic, i.e., there exists a constant $c_0 > 0$ such that

$$\sum_{i,j=1}^{N} a_{ij}(\xi) d_i d_j \geqq c_0 \sum_{i=1}^{N} d_i^2 \quad \text{for all } d = (d_1, \cdots, d_N) \in R^N \text{ and } \xi \in G. \tag{5.11}$$

In (5.8), $0 \leqq \alpha(\eta) \leqq 1$ everywhere in ∂G and $\partial/\partial\nu$ denotes the conormal derivative given by

$$\frac{\partial}{\partial \nu} = \sum_{i,j=1}^{N} a_{ij}(\eta) n_j \frac{\partial}{\partial n_i}, \qquad \eta = (\eta_i, \cdots, \eta_N) \in \partial G, \tag{5.12}$$

where $n = (n_1, \cdots, n_N)$ is the outer unit normal vector to ∂G.

In the system (5.7)–(5.9) we assume that:

(VII) The spatially varying coefficients $a_{ij}(\xi)$, $i, j = 1, \cdots, N$, and $c(\xi)$ in $\mathscr{L}$ are unknown except that $a_{ij}(\cdot)$, $c(\cdot) \in C^{\infty}(\bar{G})$ and $a_{ij}(\xi)$ satisfy (5.11).

(VIII) The boundary coefficient $\alpha(\eta)$ and the initial values $x_{0,1}(\xi), \cdots, x_{0,k}(\xi)$ are unknown except that $\alpha(\cdot) \in C^3(\partial G)$ and $x_{0,1}(\cdot), \cdots, x_{0,k}(\cdot) \in L_2(G)$.

(IX) The forced input functions $f_i(t, \xi)$, $i = 1, \cdots, k$, can be controlled (so these are known functions) and $f_i \in L_2^{\mathrm{loc}}(R^+; L_2(G))$, i.e.,

$$\int_0^t \int_G |f_i(t, \xi)|^2 \, d\xi \, dt < \infty \quad \text{for each } t > 0, \quad i = 1, \cdots, k.$$

Under the conditions VII–IX, there exists a unique *weak* solution $x(x_{0,i}, f_i; t, \xi)$ in $L_2^{\mathrm{loc}}(R^+; L_2(G))$ of the system corresponding to the initial value $x_{0,i}$ and the forced input f_i for each $i = 1, \cdots, k$ (cf. [12], [13], [17], [19]). The representation of these *weak* solutions will be given later. The state $x(t, \xi)$ in (5.7)–(5.9) is understood in this sense.

By the model we understand the system (5.7)–(5.9) in which $a_{ij}(\xi)$, $i, j = 1, \cdots, N$, $c(\xi)$ in $\mathscr{L}$, $\alpha(\eta)$, $\partial/\partial\nu$ and $x_{0,i}(\xi)$, $i = 1, \cdots, k$, are replaced by $a_{ij}^{m}(\xi)$, $i, j = 1, \cdots, N$, $c^{m}(\xi)$ in $\mathscr{L}$, $\alpha^{m}(\eta)$, $\partial/\partial\nu^{m}$ and $x_{0,i}^{m}(\xi)$, $i = 1, \cdots, k$, respectively. Here $\partial/\partial\nu^{m}$ is given by (5.12) in which a_{ij} is replaced by a_{ij}^{m}. The differential operator given by the model coefficients satisfying (5.11) is denoted by $\mathscr{L}^{m}$. The corresponding model solution is denoted by $x^{m}(x_{0,i}^{m}, f_i; t, \xi)$ for $i = 1, \cdots, k$. As usual the quantities suffixed by m are assumed to be known.

We shall say that the coefficients $a_{ij}(\xi)$, $i, j = 1, \cdots, N$ and $c(\xi)$ and/or the set of initial values $\{x_{0,1}, \cdots, x_{0,k}\}$ are identifiable on J if

$$a_{ij}(\xi) = a_{ij}^{m}(\xi), \quad i, j = 1, \cdots, N, \quad \text{and} \quad c(\xi) = c^{m}(\xi) \quad \text{for all } \xi \in \bar{G} \tag{5.13}$$

and/or

$$x_{0,1}(\xi) = x_{0,1}^{m}(\xi), \cdots, x_{0,k}(\xi) = x_{0,k}^{m}(\xi) \quad \text{for a.e. } \xi \in G \tag{5.14}$$

follow from the relations

$$e_i(t, \xi) = x(x_{0,i}, f_i; t, \xi) - x^{m}(x_{0,i}^{m}, f_i; t, \xi) = 0 \quad \text{for a.e. } \xi \in G \text{ and } t \in J, \quad i = 1, \cdots, k. \tag{5.15}$$

The identifiability of the boundary coefficient $\alpha(\eta)$ can be defined similarly. If all data are analytic, (5.15) follows from the zero output errors on some open set contained in G.

By considering the functions x, x^{m}, f_i, $x_{0,i}$, $x_{0,i}^{m}$ as elements in $L_2(G)$, the system (5.7)–(5.9) and its model can be written as follows:

$$S_0: \begin{cases} \dot{x}(t) = A_0 x(t) + f_i(t), \\ x(0) = x_{0,i}, \quad (i = 1, \cdots, k), \end{cases} \qquad S_0^{m}: \begin{cases} \dot{x}^{m}(t) = A_0^{m} x^{m}(t) + f_i(t), \\ x^{m}(0) = x_{0,i}^{m}, \quad (i = 1, \cdots, k). \end{cases}$$

Here the operator A_0 denotes the realization in $L_2(G)$ under the boundary condition (5.8). The operator $A_0^{\sim}$ denotes the similar realization of model. Since G is bounded, A_0 and $A_0^{\sim}$ satisfy (H) and I(i). We denote the sets of eigenvalues and eigenfunctions of A_0 and $A_0^{\sim}$ by $\{\tau_n, \psi_{nj}: j=1, \cdots, r_n, n=1, 2, \cdots\}$ and $\{\Lambda_n, \Psi_{nj}: j=1, \cdots, k_n, n=1, 2, \cdots\}$, respectively. It is known that $\psi_{nj}, \Psi_{nj} \in C^2(\bar{G})$ for all n, j.

Now we can give the representation of the *weak* solutions $x(x_{0,i}, f_i; t, \xi)$:

$$x(x_{0,i}, f_i; t, \xi) = \sum_{n=1}^{\infty} e^{\tau_n t} \sum_{j=1}^{r_n} \langle x_{0,i}, \psi_{nj}\rangle_G \psi_{nj}(\xi)$$

$$+ \sum_{n=1}^{\infty} \sum_{j=1}^{r_n} \left(\int_0^t e^{\tau_n (t-s)} \langle f_i(s, \cdot), \psi_{nj}\rangle_G \, ds \right) \psi_{nj}(\xi) \tag{5.16}$$

for all $t \geqq 0$ and a.e. $\xi \in G$, $i = 1, \cdots, k$.

A similar formula holds for the *weak* solutions $x^{\sim}(x_{0,i}^{\sim}, f_i; t, \xi)$ of the model. We don't give the representation here but we number it by $(5.16)_{\sim}$. The first series in (5.16) converges uniformly on $[\delta, \infty) \times \bar{G}$ for each $\delta > 0$ and satisfies (5.7) with $f_i(t, \xi) \equiv 0$ in the classical sense. The second series satisfies (5.7) in the sense of distributions and converges on R^+ as a function of t for almost every $\xi \in G$ (cf. [13, Chap. II]). The equations (5.16) mean that the concept of mild solution coincides with that of *weak* solution as a function in $L_2^{\mathrm{loc}}(R^+; L_2(G))$. Let $f_i(t, \xi) \equiv 0$, $i = 1, \cdots, k$ in (5.7). In this case the *weak* solutions make sense as functions of (t, ξ) in $(R^+ - \{0\}) \times \bar{G}$, so the condition (5.15) can be replaced by

(5.15*) $\quad e_i(t, \xi) = 0$ for all $\xi \in \bar{G}$ and $t \in J \subset R^+ - \{0\}$, $i = 1, \cdots, k$.

Let $f_i(t) = 0$, $i = 1, \cdots, k$ in S_0 and $S_0^{\sim}$ and put the observation characteristic $B = I$ by (5.15) or (5.15*). Then applying Theorem 5.1 to the system S_0 (and its model $S_0^{\sim}$), we see that a sufficient condition for the identifiability of $a_{ij}(\xi)$, $i, j = 1, \cdots, N$, $c(\xi)$ and $\{x_{0,1}, \cdots, x_{0,k}\}$ on J of positive measure is that

$$\operatorname{rank} \begin{pmatrix} \langle x_{0,1}^{\sim}, \Psi_{n1}\rangle_G & \cdots & \langle x_{0,1}^{\sim}, \Psi_{nk_n}\rangle_G \\ \langle x_{0,2}^{\sim}, \Psi_{n1}\rangle_G & \cdots & \langle x_{0,2}^{\sim}, \Psi_{nk_n}\rangle_G \\ \vdots & & \vdots \\ \langle x_{0,k}^{\sim}, \Psi_{n1}\rangle_G & \cdots & \langle x_{0,k}^{\sim}, \Psi_{nk_n}\rangle_G \end{pmatrix} = k_n \quad \text{for all } n = 1, 2, \cdots. \tag{5.17}$$

This is verified easily in the following way. By Theorem 5.1, we have that (5.15) (or (5.15*)) and (5.17) imply that

$$A_0 x = A_0^{\sim} x \quad \text{for all } x \in D(A_0^{\sim}) \quad \text{and} \tag{5.18}$$

$$x_{0,1} = x_{0,1}^{\sim}, \cdots, x_{0,k} = x_{0,k}^{\sim} \quad \text{in } L_2(G). \tag{5.19}$$

Since $C_0^{\infty}(G) \subset D(A_0^{\sim})$, it follows from (5.18) that the equalities in (5.13) hold for almost every $\xi \in G$. By the continuity of those coefficients (5.13) follows. It is obvious that (5.19) is equivalent to (5.14). We note that if all $x_{0,i}$, $x_{0,i}^{\sim}$ are continuous on $\bar{G}$, then all equalities in (5.14) hold for all $\xi \in \bar{G}$.

We now consider the identifiability of $\alpha(\eta)$. Without loss of generality we can assume that

$$\psi_{nj}(\xi) = \Psi_{nj}(\xi), \qquad \xi \in \bar{G}, \quad \text{for all } n, j. \tag{5.20}$$

Then by taking the difference of boundary conditions of ψ_{nj} and Ψ_{nj}, we have

$$(\alpha(\eta) - \alpha^{\sim}(\eta)) \left(\Psi_{nj}(\eta) - \frac{\partial \Psi_{nj}}{\partial \nu^{\sim}}(\eta) \right) = 0, \qquad \eta \in \partial G, \quad \text{for all } n, j. \tag{5.21}$$

Define the following sets:

$$\Gamma_{nj}=\left\{\eta\in\partial G:\ \Psi_{nj}(\eta)\neq\frac{\partial\Psi_{nj}}{\partial\nu^{\#}}(\eta)\right\},\qquad \Gamma=\bigcup_{n=1}^{\infty}\bigcup_{j=1}^{k_n}\Gamma_{nj}. \tag{5.22}$$

Notice that Γ_{nj} has a positive measure on ∂G for any n, j, i.e., meas $(\Gamma_{nj})=\int_{\Gamma_{nj}}d\eta>0$. Indeed, if meas $(\Gamma_{nj})=0$, then by the continuity of Ψ_{nj}, $\partial\Psi_{nj}/\partial\nu^{\#}$ on ∂G and the boundary condition it follows that $\Psi_{nj}(\eta)=\partial\Psi_{nj}/\partial\nu^{\#}(\eta)=0$ for all $\eta\in\partial G$. Since Ψ_{nj} is an eigenfunction of $A_0^{\#}$, this means that $\Psi_{nj}(\eta)$ must be identically zero, hence a contradiction follows. Then meas $(\Gamma_{nj})>0$. If $\eta\in\Gamma_{nj}$, then by (5.21), (5.22) we have $\alpha(\eta)=\alpha^{\#}(\eta)$. Hence if the closure $\bar{\Gamma}=\partial G$ (e.g., in the case where α, $\alpha^{\#}$ are analytic), then the boundary coefficient $\alpha(\eta)$ is identifiable. We remark that $\bar{\Gamma}\neq\partial G$ in general.

Next we consider the identifiability by means of the forcing inputs $f_i(t,\xi)$. Let $x_{0,1}=x_{0,1}^{\#}=\cdots=x_{0,k}=x_{0,k}^{\#}=0$ in $L_2(G)$ and let $f_i(t,\xi)$, $i=1,\cdots,k$ have the form $z_i(\xi)h_i(t)$, $z_i(\cdot)\in L_2(G)$, $h_i\in L_2^{\mathrm{loc}}(R^+)$. Then by (5.16), (5.16)$_{\#}$, the output error e_i in (5.15) is represented by

$$e_i(t,\xi)=\int_0^t g_i(\xi;t-s)h_i(s)\,ds\quad\text{for a.e. }\xi\in G\text{ and }t\in J, \tag{5.23}$$

where

$$g_i(\xi;t)=\sum_{n=1}^{\infty}e^{\tau_n t}\sum_{j=1}^{r_n}\langle z_i,\psi_{nj}\rangle_G\psi_{nj}(\xi)-\sum_{n=1}^{\infty}e^{\Lambda_n t}\sum_{j=1}^{k_n}\langle z_i,\Psi_{nj}\rangle_G\Psi_{nj}(\xi).$$

It is clear that $g_i(\xi;\cdot)\in C(R^+)$ for almost every $\xi\in G$. Let $J=[0,T]$, $T>0$ (resp. $J=R^+$) and (5.15) be satisfied. If $h_i(\cdot)\neq 0$ in $L_2[0,T]$ (resp. $h_i(\cdot)\neq 0$ in $L_2(R^+)$) for all $i=1,\cdots,k$, then by the same argument as in the proof of Theorem 4.1 ii), we see from (5.23) that

$$\begin{aligned}&g_i(\xi;t)=0\quad\text{for a.e. }\xi\in G\quad\text{and}\\ &\qquad t\in[0,T_i],\quad 0<T_i<T\quad(\text{resp. }t\in R^+),\quad i=1,\cdots,k.\end{aligned} \tag{5.24}$$

The nonnegative number T_i depends only on h_i $(i=1,\cdots,k)$. Considering (5.24) as equations in $L_2(G)$ implies

$$\sum_{n=1}^{\infty}e^{\tau_n t}\sum_{j=1}^{r_n}\langle z_i,\psi_{nj}\rangle_G\psi_{nj}=\sum_{n=1}^{\infty}e^{\Lambda_n t}\sum_{j=1}^{k_n}\langle z_i,\Psi_{nj}\rangle_G\Psi_{nj}\quad\text{in }L_2(G)$$
$$\text{for } t>0 \text{ and } i=1,\cdots,k,$$

by analyticity. Hence using the same method as in the proof of Theorem 5.1, we have that a sufficient condition for the identifiability of $a_{ij}(\xi)$, $i,j=1,\cdots,N$, and $c(\xi)$ on $J=[0,T]$, $T<0$ (resp. $J=R^+$), is that

$$h_i(\cdot)\neq 0\quad\text{in }L_2[0,T]\quad(\text{resp. }h_i(\cdot)\neq 0\text{ in }L_2^{\mathrm{loc}}(R^+))\qquad\text{for all } i=1,\cdots,k \tag{5.25}$$

and

$$\operatorname{rank}\begin{pmatrix}\langle z_1,\Psi_{n1}\rangle_G & \cdots & \langle z_1,\Psi_{nk_n}\rangle_G\\ \langle z_2,\Psi_{n1}\rangle_G & \cdots & \langle z_2,\Psi_{nk_n}\rangle_G\\ \vdots & & \vdots\\ \langle z_k,\Psi_{n1}\rangle_G & \cdots & \langle z_k,\Psi_{nk_n}\rangle_G\end{pmatrix}=k_n\quad\text{for all } n=1,2,\cdots. \tag{5.26}$$

Therefore summing up the above arguments, we obtain the following result.

COROLLARY 5.1. *In the system* (5.7)–(5.9) *the coefficients* $a_{ij}(\xi)$, $i, j = 1, \cdots, N$ *and* $c(\xi)$ *are identifiable on* J *in the following two cases*:

i) *If* J *is of positive measure,* $f_i(t, \xi) \equiv 0$, $i = 1, \cdots, k$ *and if* (5.17) *is satisfied,*

ii) *If* $J = [0, T]$, $T > 0$ (*resp.* $J = R^+$), $x_{0,1} = x_{0,1}^{(m)} = \cdots = x_{0,k} = x_{0,k}^{(m)} = 0$ *in* $L_2(G)$, f_i, $i = 1, \cdots, k$, *have the form* $z_i(\xi)h_i(t)$, $z_i \in L_2(G)$, $h_i(\cdot) \in L_2^{\mathrm{loc}}(R^+)$ *and if* (5.25) *and* (5.26) *are satisfied.*

In Case i) *the set of initial values* $\{x_{0,1}, \cdots, x_{0,k}\}$ *is also identifiable on* J. *Furthermore in both cases if* $\bar{\Gamma} = \partial G$, *then the boundary coefficient* $\alpha(\eta)$ *is identifiable on* J.

Compare this corollary with Theorem 4.1. It would be easy to give an analogous result for the corresponding hyperbolic system.

We remark that an identifiability result similar to Corollary 5.1 holds for the unknown coefficients $a_\beta(\xi)$, $|\beta| \leqq 2m$, in a strongly elliptic differential operator $A_0 = \sum_{|\beta| \leqq 2m} a_\beta(\xi) D_\xi^\beta$ of order $2m$ in a bounded smooth domain G with the Dirichlet boundary condition $(\partial/\partial n)^j x(t, \eta) = 0$, $\eta \in \partial G$, $j = 0, 1, \cdots, m-1$ (cf. [1], [19]).

Example 5.2. Let $\mathscr{L}$ be the Schrödinger operator

$$\mathscr{L} = -\Delta + q(\xi), \qquad \xi \in R^N, \tag{5.27}$$

with the unknown potential $q(\xi) \geqq 0$, where Δ is the Laplace operator in R^N. We consider the following time dependent Schrödinger equation:

$$i\frac{\partial x}{\partial t}(t, \xi) = \mathscr{L}x(t, \xi), \qquad t > 0, \quad \xi \in R^N \tag{5.28}$$

with initial conditions

$$x(0, \xi) = x_{0,i}(\xi), \qquad \xi \in R^N, \quad i = 1, \cdots, k. \tag{5.29}$$

We assume in the system (5.28), (5.29) that:

(X) The potential $q(\xi)$ in $\mathscr{L}$ is unknown except that $q(\cdot) \in L_2^{\mathrm{loc}}(R^N)$, $q(\xi) \geqq 0$ for almost every ξ in R^N and $q(\xi) \to \infty$ as $|\xi| \to \infty$.

(XI) The initial values $x_{0,1}(\xi), \cdots, x_{0,k}(\xi)$ are unknown except that $x_{0,1}(\cdot), \cdots, x_{0,k}(\cdot) \in L_2(R^N)$.

To represent the system (5.28), (5.29) as evolution equations in $L_2(R^N)$, we make some preparations. We denote by A_∞ the realization of $\mathscr{L}$ on the domain $D(A_\infty) = C_0^\infty(R^N)$. It is obvious that A_∞ is symmetric and bounded from below (with a bound $q_0 = \operatorname{ess\,inf}_{\xi \in R^N} q(\xi)$) in $L_2(R^N)$. Then by the well-known theorem (Kato [14, p. 323]), A_∞ has the Friedrich's extension A_q. It is known that A_q is essentially selfadjoint, i.e., A_∞ has only one selfadjoint extension A_q (see Reed and Simon [22, Vol. 2, p. 184] and Kato [14]). Hence the system (5.28), (5.29) can be expressed by the following evolution equations in $L_2(R^N)$:

$$S_q: \begin{aligned} &\dot{x}(t) = -iA_q x(t), \\ &x(0) = x_{0,i}, \quad i = 1, \cdots, k. \end{aligned}$$

Since A_q satisfies (H) ([22, Vol. 4, p. 249]), there exists the set of eigenvalues and eigenfunctions $\{\tau_n, \psi_{nj} : j = 1, \cdots, m_n, \ n = 1, 2, \cdots\}$ of A_q. Then the mild solutions $x(x_{0,i}; t)$, $i = 1, \cdots, k$, of S_q are given by

$$x(x_{0,i}; t) = \sum_{n=1}^{\infty} e^{-i\tau_n t} \sum_{j=1}^{r_n} \langle x_{0,i}, \psi_{nj} \rangle_{R^N} \psi_{nj}, \qquad t \geqq 0, \quad i = 1, \cdots, k, \tag{5.30}$$

where $\tau_n \geqq 0$ (notice that $-iA_q$ is normal and satisfies (H_0)). Relating to the expressions (5.30), we say that the function $x(x_{0,i}; t, \xi)$, $i = 1, \cdots, k$, defined by (5.31) below is

the solution of the system (5.28), (5.29) corresponding to the initial value $x_{0,i}$:

$$x(x_{0,i}; t, \xi) = \sum_{m=1}^{\infty} e^{-i\tau_n t} \sum_{j=1}^{r_n} \langle x_{0,i}, \psi_{nj} \rangle_{R^N} \psi_{nj}(\xi), \qquad i = 1, \cdots, k. \tag{5.31}$$

For given $x_{0,i}, \cdots, x_{0,k}$ in $L_2(R^N)$ and all $t \geqq 0$, the series in (5.31) converge for almost every $\xi \in R^N$. Then the solutions (5.31) have sense almost everywhere in R^N and $x(x_{0,i}; t, \cdot) = x(x_{0,i}; t)$ in $L_2(R^N)$, $i = 1, \cdots, k$.

We now consider the model which means the system (5.28), (5.29) in which $q(\xi)$ in $\mathscr{L}$ and $x_{0,i}(\xi)$, $i = 1, \cdots, k$, are replaced by $q^{m}(\xi)$ in $\mathscr{L}$ and $x_{0,i}^{m}(\xi)$, $i = 1, \cdots, k$, respectively. The model quantities such as its solution, Friedrich's extension, etc. are suffixed by m. Then the model is represented as

$$S_q^{m}: \quad \begin{aligned} &\dot{x}^{m}(t) = -iA_q^{m} x^{m}(t), \\ &x^{m}(0) = x_{0,i}^{m}, \quad i = 1, \cdots, k. \end{aligned}$$

We denote the set of eigenvalues and eigenfunctions of A_q^{m} by $\{\Lambda_n, \Psi_{nj}: j = 1, \cdots, k_n, n = 1, 2, \cdots\}$.

As in Example 5.1 we shall say that the potential $q(\xi)$ and the set of initial values $\{x_{0,1}, \cdots, x_{0,k}\}$ are identifiable on $J \subset R^+ - \{0\}$ if

$$q(\xi) = q^{m}(\xi) \quad \text{for a.e. } \xi \in R^N,$$
$$x_{0,1}(\xi) = x_{0,1}^{m}(\xi), \cdots, x_{0,k}(\xi) = x_{0,k}^{m}(\xi) \quad \text{for a.e. } \xi \in R^N$$

follow from

$$e_i(t, \xi) = x(x_{0,i}; t, \xi) - x^{m}(x_{0,i}^{m}; t, \xi) = 0, \qquad i = 1, \cdots, k, \tag{5.32}$$
$$\text{for all } t \in J \text{ and a.e. } \xi \in R^N.$$

We now consider the abstract systems S_q and S_q^{m} and set the observation characteristic $B = I$ by (5.32). Then applying Theorem 5.1 with $X = L_2(R^N)$, we obtain (as in Example 5.1) that if

$$\operatorname{rank} \begin{pmatrix} \langle x_{0,1}^{m}, \Psi_{n1} \rangle_{R^N} & \cdots & \langle x_{0,1}^{m}, \Psi_{nk_n} \rangle_{R^N} \\ \langle x_{0,2}^{m}, \Psi_{n1} \rangle_{R^N} & \cdots & \langle x_{0,2}^{m}, \Psi_{nk_n} \rangle_{R^N} \\ \vdots & & \vdots \\ \langle x_{0,k}^{m}, \Psi_{n1} \rangle_{R^N} & \cdots & \langle x_{0,k}^{m}, \Psi_{nk_n} \rangle_{R^N} \end{pmatrix} = k_n \quad \text{for all } n = 1, 2, \cdots, \tag{5.33}$$

then the potential $q(\xi)$ and the set of initial values $\{x_{0,1}, \cdots, x_{0,k}\}$ are identifiable on J of positive measure.

Let $q(\xi) = |\xi|^2 = \xi_1^2 + \cdots + \xi_N^2$. In this case the phenomenon governed by (5.27), (5.28) is called the harmonic oscillator and has a fundamental importance in quantum mechanics (especially, in the case when $N = 3$). Since the potential $q(\xi) = |\xi|^2$ satisfies the conditions in (X), the selfadjoint operator A_H in $L_2(R^N)$ can be determined by (5.27) with $q(\xi) = |\xi|^2$. When $N = 1$, we denote the operator A_H by H (which is known as the Hamiltonian of the linear harmonic oscillator). Since $L_2(R^N)$ is isometric to the tensor product $L_2(R^1) \otimes L_2(R^1) \otimes \cdots \otimes L_2(R^1)$, n times, we can identify

$$A_H = H \otimes I \otimes \cdots \otimes I + I \otimes H \otimes I \otimes \cdots \otimes I + \cdots + I \otimes \cdots \otimes I \otimes H,$$

where I is the identity operator on $L_2(R^1)$ (cf. Weichman [32, Chap. 8]). It is known

that the eigenvalues and eigenfunctions of H are given by

$$\tau_n = 2n+1, \qquad n = 0, 1, 2, \cdots,$$
$$\psi_n(\eta) = c_n H_n(\eta)\, e^{-\eta^2/2}, \qquad \eta \in R^1,$$

where $H_n(\eta)$ is the Hermite polynomial of order n and c_n are constants for normalization. For details, see Courant and Hilbert [4, Chap. 5] and Landau and Lifschitz [18, Chap. 3]. Therefore the eigenfunctions of A_H are given by

$$\psi_{n_1 \cdots n_N}(\xi) = H_{n_1}(\xi_1) \cdots H_{n_N}(\xi_N) \cdot e^{-|\xi|^2/2}, \qquad n_1, \cdots, n_N = 0, 1, 2, \cdots,$$

and the corresponding eigenvalues are

$$\tau_{n_1 \cdots n_N} = \sum_{i=1}^{N} \tau_{n_i} = 2 \sum_{n=1}^{N} n_i + N, \qquad n_1, \cdots, n_N = 0, 1, 2, \cdots.$$

Since the multiplicity of A_H equals 1 when $N = 1$ and equals ∞ when $N \geqq 2$, the above system is not suitable as the model for the identifiability of the potential $q(\xi)$ when $N \geqq 2$. But for any dimension N, the system can be used as the model for the K-mode identifiability, i.e., the identifiability of finite numbers of eigenvalues and eigenfunctions (see the next section).

Theorems 5.1 and 5.2 are for the case when $f = 0$ in $L_2^{\mathrm{loc}}(R^+; X)$. For the case where the initial values vanish but the forcing functions $f_1, \cdots, f_k$ as inputs do not vanish, we can establish the similar identifiability conditions for the systems S_1 and S_2 as in Example 5.1. Here the definition of identifiability of S_j, $j = 1, 2$, with respect to the set of forcing functions $\{f_1, \cdots, f_k\}$ is similar to that given in Definition 5.1.

THEOREM 5.3. *Let all initial values of S_1 and S_1^{m} be 0 in X and let $G = \{f_1, \cdots, f_k\} \subset L_2^{\mathrm{loc}}(R^+; X)$ be the form $\{z_1 g_1(t), \cdots, z_k g_k(t)\}$, where $Z = \{z_1, \cdots, z_k\} \subset X$ and $g_1(\cdot), \cdots, g_k(\cdot) \in L_2^{\mathrm{loc}}(R^+)$. If the set Z is compatible with respect to A^{m} and $g_i(\cdot) \neq 0$ in $L_2[0, T]$ (resp. $g_i(\cdot) \neq 0$ in $L_2^{\mathrm{loc}}(R^+)$) for all $i = 1, \cdots, k$, then the system S_1 is k-identifiable on $[0, T]$ (resp. on R^+) with respect to G (by using the model S_1).*

THEOREM 5.4. *Let all initial values of S_2 and S_2^{m} be 0 in X and let G be the same set as given in Theorem 5.3. Then the same conclusion as in Theorem 5.3 holds for the system S_2.*

6. Relations between identifiability and observability and controllability. This section is devoted to study the relations between identifiability and observability and controllability. Let A generate a semigroup $T(t)$ on X and let $B \in L(X, Y)$. Let U be a complex, separable Banach space (of controls) and let $E \in L(U, X)$. We denote the dual Banach space of U and the dual operator of E by U^* and E^*, respectively. To make our standpoint clear, we assume that A satisfies (H). Then there exists the set of eigenvalues and eigenfunctions $\{\lambda_n, \psi_{nj} : j = 1, \cdots, m_n, n = 1, 2, \cdots\}$ of A satisfying (a), (b), (c), (d) in § 3. The nth eigenmanifold and the associated nth eigenprojector are denoted by X_n and P_n ($n = 1, 2, \cdots$). We now give the following definition:

DEFINITION 6.1. The system $\{A, E\}$ is said to be

(i) *approximately controllable* if $E_\infty = \bigcup_{t \geqq 0} T(t)EU = X$;

(ii) *K-mode controllable* if $E_\infty \supset \bigoplus_{n=1}^{K} X_n$.

The system $\{A, B\}$ is said to be

(iii) *approximately observable* if $B_\infty = \bigcap_{t \geqq 0} \mathrm{Ker}\, BT(t) = \{0\}$;

(iv) *K-mode observable* if $B_\infty \subset \bigoplus_{n=K+1}^{\infty} X_n$.

The condition in (i) is equivalent to:

$$E^* T(t) x = 0 \text{ in } U^* \text{ for } t \geqq 0 \text{ implies } x = 0 \text{ in } X,$$

which gives the contraposition of Proposition 1 in [7] for complete controllability (in the terminology of Fattorini). For other equivalent statements to (i), we refer to Curtain and Pritchard [6] and Triggiani [30]. The condition in (iv) is equivalent to:

$$BT(t)x = 0 \text{ in } Y \text{ for } t \geqq 0 \text{ implies } P_1 x = \cdots = P_K x = 0 \text{ in } X.$$

Hence the statement (iv) is a refinement of the definition of N-mode observability given by Goodson and Klein [10] (in which forced and boundary inputs are assumed to be known!).

Corresponding to the K-mode observability, we shall define the K-mode identifiability. In what follows we use the same notation as in § 5.

DEFINITION 6.1. (1) The system S_1 is said to be *K-mode identifiable* on $J \subset \mathrm{R}^+$ with respect to the set $\{x^{m}_{0,1}, \cdots, x^{m}_{0,k}\}$ of model initial values if

$$\lambda_n = \Lambda_n, \quad X_n = X^{m}_n, \quad n = 1, \cdots, K, \quad \text{and}$$

$$\left(\sum_{n=1}^{K} P^{m}_n\right)(x_{0,i} - x^{m}_{0,i}) = 0 \quad \text{in } X, \quad i = 1, \cdots, k,$$

follow from

$$e_1(x_{0,1}, x^{m}_{0,1}; t) = 0, \cdots, e_1(x_{0,k}, x^{m}_{0,k}; t) = 0 \quad \text{in } X, \quad t \in J.$$

(2) The set $E = \{x_1, \cdots, x_k\}$ is said to be *K-mode compatible* with respect to A^{m} if rank $M_n = k_n$ for $n = 1, \cdots, K$.

Using the same argument as in the proof of Theorem 5.1, we have the following corollary:

COROLLARY 6.1. *Let $f = 0$ in $L_2^{\mathrm{loc}}(R^+; X)$ and $J \subset R^+$ be of positive measure. If the set $E^{m}_0 = \{x^{m}_{0,1}, \cdots, x^{m}_{0,k}\}$ is K-mode compatible with respect to A^{m}, then the system S_1 is K-mode identifiable on J.*

In Example 5.2, if the Hamiltonian A_H of an N-dimensional harmonic oscillator is used as the model operator, then it requires $(K+1)\cdots(K+N-1)/(N-1)!$ numbers of initial values for K-mode identifiability.

The compatibility condition can be identified with the rank condition for controllability or observability in the following sense (cf. [30, § 3.2], [7]).

The set $E = \{x_1, \cdots, x_k\}$ is compatible (resp. K-mode compatible) with respect to A^{m} if and only if:

(a) $\{A^{m}, (x_1, \cdots, x_k)\}$ is approximately controllable (resp. K-mode controllable), where the k-tuple $(x_1, \cdots, x_k)$ is identified with the operator $E^c: C^k \to X$ given by

$$E^c(u_1, \cdots, u_k) = \sum_{i=1}^{k} x_i u_i \in X,$$

or

(b) $\{A^{m}, (x_1, \cdots, x_k)\}$ is approximately observable (resp. K-mode observable), where the k-tuple $(x_1, \cdots, x_k)$ is identified with the operator

$$E^o: X \to C^k \quad \text{given by} \quad E^o x = (\langle x_1, x\rangle, \cdots, \langle x_k, x\rangle) \in C^k.$$

We shall say as in [7] that A is finitely controllable if for some $E \in L(C^k, X)$, $\{A, E\}$ is approximately controllable. Then from [7, Thm. 3.2], [30, Thm. 3.6], Theorem 5.1 and Corollary 6.1, follows the next corollary which shows the connection between identifiability and controllability.

COROLLARY 6.2. *Let $f = 0$ in $L_2^{\mathrm{loc}}(R^+; X)$ and $J \subset R^+$ be of positive measure. If A^{m} is finitely controllable, then S_1 is identifiable on J. More precisely, if*

$\{A^m, (x_{0,1}^m, \cdots, x_{0,k}^m)\}$ *is approximately controllable (resp. K-mode controllable), then* S_1 *is identifiable (resp. K-mode identifiable) on J with respect to* $E_0^m = \{x_{0,1}^m, \cdots, x_{0,k}^m\}$.

In Theorem 4.3 i), it is assumed that the initial value x_0 is a priori known. But from a practical point of view this assumption cannot be accepted, since the observability of the system S_1 is not established yet. To apply the theorem, we must determine x_0 uniquely. For this purpose we use the K-mode observable operator $B_K \in L(X, C^{n_K})$ for $K \geqq 2$, which is given by

$$B_K x = (\langle w_1, x\rangle, \cdots, \langle w_{n_K}, x\rangle) \in C^{n_K} \quad \text{for all } x \in X,$$

where $n_K = \max\{m_1, \cdots, m_K\}$, $w_1, \cdots, w_{n_K} \in X$ and the set $\{w_1, \cdots, w_{n_K}\}$ is K-mode compatible with respect to A_0. It is clear that $\{A_0, B_K\}$ is K-mode observable. Hence it can be assumed that $P_1^0 x_0, \cdots, P_K^0 x_0$ are all known exactly. Let K be fixed. If $P_n^0 x_0 \neq 0$ in X for two distinct n's, we can choose an operator B such that $B(P_n^0 x_0) \neq 0$ for such n's. For the observation operator B, Theorem 4.3 i) can apply to identify the parameters a, b in S_1. If not, abandon x_0 and take another initial value x_0' satisfying such a condition and then apply the theorem. For any initial value x_0 such that $P_n^0 x_0 = 0$ except for at most one n, the parameters a, b in S_1 are not identifiable on $R^+ - \{0\}$ for any observation.

It would be easy to give results analogous to the above for the second order system S_2.

Acknowledgments. I would like to express my sincere gratitude to Professor H. Tanabe for his fruitful suggestions which have helped in improving the presentation. The introduction of the graph norm in pointwise observations is due to his idea. I am grateful to Professor H. Murakami for his constant encouragement and to Doctors S. Kitamura and T. Suzuki for their kind discussions. I also extend my gratitude to Professors H. T. Banks, M. Courdesses and the referees for their helpful comments and suggestions.

REFERENCES

[1] S. AGMON, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, NJ, 1965.
[2] J. BALL, *Strongly continuous semi-groups, weak solutions and the variation of constants formula*, Proc. AMS, 63 (1977), pp. 370–373.
[3] M. G. CHAVENT, *Analyse fonctionnelle et identification des coefficients répartis dans les équations aux dérivées partielles*, Thèse, Faculté des Sciences de Paris, 1971.
[4] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics* I, Interscience, New York, 1953.
[5] M. COURDESSES, M. P. POLIS AND M. AMOUROUX, *On identifiability of parameters in a class of parabolic distributed systems*, IEEE Trans. Automat. Control (to appear).
[6] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear System Theory*, Lecture Notes in Control and Information Sciences 8, Springer-Verlag, Berlin, 1978.
[7] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
[8] ———, *Ordinary differential equations in linear topological spaces*, I and II, J. Differential Equations, 5 (1968), pp. 72–105, and 6 (1969), pp. 50–70.
[9] I. M. GEL'FAND AND B. M. LEVITAN, *On the determination of a differential equation from its spectral function*, AMS Trans. Ser. 2, 1 (1955), pp. 253–304.
[10] R. E. GOODSON AND R. E. KLEIN, *A definition and some results for distributed system observability*, IEEE Trans. Automat. Control, AC-15 (1970), pp. 165–174.
[11] E. L. INCE, *Ordinary Differential Equations*, Dover, New York, 1944.
[12] S. ITO, *Fundamental solutions of parabolic differential equations and boundary value problem*, Japan J. Math., 27 (1957), pp. 55–102.
[13] ———, *Partial Differential Equations*, Baifukan Publ., Tokyo, 1967 (in Japanese).
[14] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, Berlin, 1966.

[15] T. KATO, *A second look at the essential selfadjointness of the Schrödinger operators*, in Physical Reality and Mathematical Description, Reidel, Dordrecht, Holland, 1974, pp. 193–201.

[16] S. KITAMURA AND S. NAKAGIRI, *Identifiability of spatially varying and constant parameters in distributed systems of parabolic type*, this Journal, 15 (1977), pp. 785–802.

[17] O. A. LADYŽENSKAYA, V. A. SOLONNIKOV AND N. N. URAL'CEVA, *Linear and Quasi-Linear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.

[18] L. D. LANDAU AND E. M. LIFSHITZ, *Quantum Mechanics*, Pergamon, Oxford, 1965.

[19] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, New York, 1972.

[20] R. MURAYAMA, *The Gel'fand–Levitan theory and certain inverse problems for the parabolic equation*, J. Fac. Sci. Univ. Tokyo Sect. 1A Math., 28 (1981), pp. 317–330.

[21] S. NAKAGIRI, S. KITAMURA AND H. MURAKAMI, *Mathematical treatment of the constant parameter identifiability of distributed systems of parabolic type*, Math. Sem. Notes Kobe Univ., 5 (1977), pp. 97–105.

[22] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics*, Vol. 2, Vol. 4, Academic Press, New York, 1975, 1978.

[23] A. PIERCE, *Unique identification of eigenvalues and coefficients in a parabolic problem*, this Journal, 17 (1979), pp. 494–499.

[24] M. P. POLIS AND R. E. GOODSON, *Parameter identification in distributed systems*, Proc. IEEE, 64 (1976), pp. 45–61.

[25] M. SOVA, *Cosine operator functions*, Rozprawy Matematiyczne, 49 (1966), pp. 1–47.

[26] T. SUZUKI AND R. MURAYAMA, *A unique theorem in an identification problem for coefficients of parabolic equations*, Proc. Japan Acad. Ser. A Math. Sci., 56 (1980), pp. 259–263.

[27] T. SUZUKI, *Uniqueness and nonuniqueness in an inverse problem for the parabolic equation*, J. Differential Equations, to appear.

[28] E. C. TITCHMARSH, *Introduction to the Theory of Fourier Integrals*, Oxford University Press, London, 1938.

[29] C. C. TRAVIS AND G. F. WEBB, *Cosine families and abstract nonlinear second order differential equations*, Acta Math. Acad. Sci. Hungar., 32 (1978), pp. 75–96.

[30] R. TRIGGIANI, *Extensions of rank conditions for controllability and observability to Banach spaces and unbounded operators*, this Journal, 14 (1976), pp. 313–338.

[31] ———, *On the relationship between first and second order controllable systems in Banach spaces*, this Journal, 16 (1978), pp. 847–859.

[32] C. WEICHMANN, *Linear Operators in Hilbert Spaces*, Springer-Verlag, Berlin, 1979.

[33] M. COURDESSES, *Comments on "Identifiability of spatially-varying and constant parameters in distributed systems of parabolic type"*, this Journal, 21 (1983), pp. 410–413.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

0363-0129/83/2104-0002 $01.25/0

SOLUTION OF THE BELLMAN EQUATION ASSOCIATED WITH AN INFINITE DIMENSIONAL STOCHASTIC CONTROL PROBLEM AND SYNTHESIS OF OPTIMAL CONTROL*

VIOREL BARBU† AND GIUSEPPE DA PRATO‡

Abstract. We prove the existence and uniqueness of the dynamic programming equation for control diffusion processes in Hilbert spaces.

Introduction. Consider the optimal control problem:
Minimize

$$\text{(P)} \qquad E\left(\int_0^T (g(t, x(t))) + \frac{1}{2}|u(t)|^2\right) dt + \phi_0(x(T))$$

over all u in $M_w^2\ (0, T; H)$ subject to

$$\text{(0.1)} \qquad \begin{aligned} &dx = (Ax + u)\, dt + \sqrt{\varepsilon}\, dW_t, \qquad \varepsilon > 0, \\ &x(0) = x_0. \end{aligned}$$

Here A is the infinitesimal generator of a contraction C_0-semigroup in a real separable Hilbert space H with the norm $|\cdot|$ $(\Omega, \mathcal{F}, P)$ is a probability space, W_t is a H-valued Brownian motion on $(\Omega, \mathcal{F}, P)$ and E is the expectation.

The function $g: [0, T] \times H \to \mathbb{R}$ is continuous and convex as a function of x for every $t \in [0, T]$.

This paper is concerned with a direct approach to the dynamic programming equation associated with problem (P), namely (see for instance [4], [6]):

$$\text{(0.2)} \qquad \begin{aligned} &\phi_t(t, x) + \frac{1}{2}|\phi_x(t, x)|^2 - \langle Ax, \phi_x(t, x)\rangle - \frac{\varepsilon}{2} \operatorname{Tr}(S\phi_{xx}(t, x)) = g(t, x), \\ &\phi(0, x) = \phi_0(x) \end{aligned}$$

where S is the covariance of W_1. In few words the idea (already used in [2]) consists in approximating the term $\frac{1}{2}|\phi_x|^2$ by $\alpha^{-1}(\phi - \phi_\alpha)$, where ϕ_α is the convex regularization of ϕ and after to let α tend to zero. We have previously studied in [3] this problem in the particular case where g_0 and $g(\cdot, x)$ have a sublinear growth. We remark that when g is quadratic (0.2) reduces to a Riccati equation and the corresponding control problem has been studied by several authors (see for instance [6]).

The contents of the paper are outlined below. Sections 1 and 2 are concerned with notation and preliminary results for spaces of differential functions and convex functions frequently used in the text. Section 3 studies a linearized version of problem (0.2). Section 4 gives the main result on existence and uniqueness for problem (0.2). Furthermore it is shown that for $\varepsilon \to 0$ the solution to (0.2) converges to the solution of the Hamilton–Jacobi equation

$$\text{(0.3)} \qquad \begin{aligned} &\phi_t(t, x) + \tfrac{1}{2}|\phi_x(t, x)|^2 - \langle Ax, \phi_x(t, x)\rangle = g(t, x), \\ &\phi(0, x) = \phi_0(x) \end{aligned}$$

* Received by the editors January 25, 1982, and in revised form June 30, 1982.
† University of Iasi, Romania.
‡ Scuola Normale Superiore, 56100 Pisa, Italy.

which has been studied in [2] by a different method. This result resembles the classical approach of Hamilton–Jacobi equations in finite dimensional spaces [9]. Finally in § 5 we study the synthesis for problem (P) proving the existence and uniqueness of a smooth feedback control.

1. Notation and preliminary results. Throughout in the sequel H is a real separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $|\cdot|$. For $k = 0, 1, \cdots$, denote by $C^k(H)$ the space of all k times continuously differentiable (Fréchet) functions $\phi: H \to \mathbb{R}$ which are bounded on bounded subset on H along with their derivatives up to order k. For $k = 0$ we shall simply write $C(H)$. Other notation such as $C^{(k)}([0, T] \times H)$, $k = 0, 1$, and $\mathscr{L}(H)$ is obvious. We set

$$|\phi|_{h,n} = \text{Sup } \{|\phi^{(h)}(x)(1+|x|^{2n})^{-1}|; x \in H\}, \tag{1.1}$$

$$\begin{aligned} \|\phi\|_{h,n} = \text{Sup } \{&|\phi^{(h)}(x) - \phi^{(h)}(y)||x-y|^{-1} \\ &\cdot (1+(|x| \vee |y|)^{2n})^{-1}; x \neq y \in H\} \end{aligned} \tag{1.2}$$

where $\phi^{(h)}$ stands for the derivative of order h and $|x| \vee |y| = \max(|x|, |y|)$.

LEMMA 1. *For any $\phi \in C^{k+1}(H)$ one has*

$$\|\phi\|_{k,n} = |\phi|_{k+1,n}. \tag{1.3}$$

Proof. For $|y| \leqq 1$ we have

$$\begin{aligned} &|\phi^{(k+1)}(x) \cdot y|(1+|x|^{2n})^{-1} \\ &\quad = \lim_{t \to 0} |\phi^{(k)}(x+ty) - \phi^{(k)}(x)| \cdot |t|^{-1}[1+(|x| \vee |x+ty|)^{2n}]^{-1} \leqq \|\phi\|_{k,n}|y|, \end{aligned}$$

which implies that $|\phi|_{k+1,n} \leqq \|\phi\|_{k,n}$. Conversely we have

$$\begin{aligned} &|\phi^{(k)}(x) - \phi^{(k)}(y)||x-y|^{-1}[1+(|x| \vee |y|)^{2n}]^{-1} \\ &\quad \leqq \int_0^1 |\phi^{(k+1)}((1-\lambda)x + \lambda y)\, d\lambda [1+(|x| \vee |y|)^{2n}]^{-1}. \end{aligned}$$

Since $|(1-\lambda)x + \lambda y| \leqq |x| \vee |y|$, the latter implies $\|\phi\|_{k,n} \leqq |\phi|_{k+1,n}$ as claimed.

We shall also use the following notation:

$$X = \{\phi \in C(H);\ |\phi|_{0,n_0} < +\infty\},$$

$$Y = \{\phi \in C^1(H); |\phi|_{0,n_0} + |\phi|_{1,n_1} < +\infty\},$$

$$Z = \{\phi \in C^2(H); |\phi|_{0,n_0} < +\infty, |\phi|_{1,n_1} < +\infty, |\phi|_{2,n_2} < +\infty, \|\phi\|_{2,n_3} < +\infty\}$$

where $n_0 \geqq n_1 \geqq n_2 \geqq n_3 \geqq 0$ are fixed integers. The spaces X, Y and Z are endowed with the norms

$$|\phi|_X = |\phi|_{0,n_0}, \qquad |\phi|_Y = |\phi|_{0,n_0} + |\phi|_{1,n_1}, \tag{1.4}$$

$$|\phi|_Z = |\phi|_{0,n_0} + |\phi|_{1,n_1} + |\phi|_{2,n_2} + \|\phi\|_{2,n_3}. \tag{1.5}$$

We note for the purposes of § 4 the following lemma.

LEMMA 2. *For each $M > 0$ the set*

$$\Lambda = \{\phi \in Z; |\phi|_Z \leqq M\} \tag{1.6}$$

is closed in X. Furthermore if $\{\phi_n\} \subset \Lambda$ is convergent in X to ϕ then

$$\langle y, \phi_{n,x}(x) \rangle \to \langle y, \phi_x(x) \rangle \tag{1.7}$$

uniformly on every bounded subset of H. Finally, if $S \in \mathcal{L}(H)$ is a nuclear symmetric operator then

$$\text{Tr}\,(S\phi_{n,xx}(x)) \to \text{Tr}\,(S\phi_{xx}(x)) \tag{1.8}$$

uniformly on bounded subsets of H.

Proof. Let $\{\phi_n\} \subset Z$, $|\phi_n|_Z \leqq M$ and $\phi \in X$ be such that $\phi_n \to \phi$ in X. We have to show that $\phi \in Z$ and $|\phi|_Z \leqq M$. For any $R > 0$ there exist $M_{i,R}$, $i = 0, 1, 2, 3$, such that

$$\begin{aligned} &\sup_{x \in B_R} |\phi^{(i)}(x)| \leqq M_{i,R}, \qquad i = 0, 1, 2, \\ &\sup_{\substack{x,y \in B_R \\ x \neq y}} \frac{|\phi^{(2)}(x) - \phi^{(2)}(y)|}{|x - y|} \leqq M_{3,R}, \end{aligned} \tag{1.9}$$

where $B_R = \{x \in H;\ |x| \leqq R\}$.

Let now $R > 0$ be fixed and $x, y \in B_R$, $h \in [-1, 1]$. We set

$$\psi_{n,x,y}(h) = \psi_n(h) = \phi_n(x + hy) \tag{1.10}$$

and notice that

$$\psi_n'(h) = \langle y, \phi_{nx}(x + hy)\rangle. \tag{1.11}$$

By (1.9) we have

$$\begin{aligned} &|\psi_n(h)| \leqq M_{0,2R},\ |\psi_n'(h)| \leqq M_{1,2R}|y|, \\ &|\psi_n'(h) - \psi_n'(k)| \leqq M_{2,2R}|h - k||y|^2, \qquad h, k \in [-1, 1]. \end{aligned} \tag{1.12}$$

By the Ascoli–Arzelà theorem there exists a subsequence $\{n_k\}$ of $\mathbb{N}$ such that $\{\psi_{n_k}'\}$ is uniformly convergent as $k \to \infty$. It follows that

$$\begin{aligned} &\psi_{n_k}(h) \to \phi(x + hy), \\ &\psi_{n_k}'(h) \to \frac{d}{dh}(\phi(x + hy)), \end{aligned} \qquad \text{uniformly in } [-1, 1] \tag{1.13}$$

and consequently ϕ is Gateaux differentiable (we shall denote by $D\phi(x)$ the Gateaux derivative of ϕ at x). We have

$$\psi_{n_k}'(0) = \langle y, \phi_{n_k x}(x)\rangle \to \langle y, D\phi(x)\rangle. \tag{1.14}$$

We can show now that $D\phi(x)$ is continuous in x which will imply $D\phi = \phi_x$. We have indeed

$$|\psi_{n_k,x,y}'(0) - \psi_{n_k,z,y}'(0)| = |\langle y, \phi_{n_k x}(x) - \phi_{n_k}(z)\rangle| \leqq M_{2,R}|x - z||y|, \tag{1.15}$$

from which, recalling (1.14) and letting x tend to $+\infty$ we get

$$|\langle y, D\phi(x) - D\phi(z)\rangle| \leqq M_{2,R}|x - z||y| \quad \forall y \in H, \tag{1.16}$$

$$|D\phi(x) - D\phi(z)| \leqq M_{2,R}|x - z|. \tag{1.17}$$

Consequently $\phi \in C^1(H)$, $D\phi = \phi_x$ and

$$\langle y, \phi_{n_k x}(x)\rangle \xrightarrow{k \to \infty} \langle y, \phi_x(x)\rangle \quad \forall y \in H. \tag{1.18}$$

The latter implies (1.7) by a standard argument. Now we set

$$\zeta_{n,x,y,u}(h) = \zeta_n(h) = \langle u, \phi_{nx}(x + hy)\rangle. \tag{1.19}$$

It follows that

$$\zeta_n'(h) = \phi_{nxx}(x+hy)(u, y), \tag{1.20}$$

and by (1.9)

$$|\zeta_n'(h) - \zeta_n'(k) \leqq M_{3R}|h-k||u||y|^2. \tag{1.21}$$

Using once again the Ascoli–Arzelà theorem we may conclude that there exists a subsequence $\{n_k'\}$ of $\{n_k\}$ such that

$$\begin{aligned} &\zeta_{n_k}(h) = \langle u, \phi_{n_k x}(x+hy)\rangle \to \langle u, \phi_x(x+hy)\rangle, \\ &\zeta_{n_k}'(h) \to \frac{d}{dh}\langle u, \phi_x(x+hy)\rangle. \end{aligned} \tag{1.22}$$

We set

$$\left[\frac{d}{dh}\langle u, \phi_x(x+hy)\rangle\right]_{h=0} = E\phi(x)(u, y). \tag{1.23}$$

From (1.22) it follows that

$$\phi_{nxx}(x)(u, y) \to E\phi(x)(u, y). \tag{1.24}$$

To prove that $E\phi = \phi_{xx}$ it suffices to show that E is continuous in x. We have

$$\begin{aligned} |\zeta'_{n,x,y,u}(0) - \zeta'_{n,z,y,u}(0)| &= |\phi_{nxx}(x)(u, y) - \phi_{nxx}(z)(u, y)| \\ &\leqq M_{3,R}|x-z||u||y| \end{aligned} \tag{1.25}$$

and, recalling (1.24), we get for $k \to \infty$

$$|(E\phi(x) - E\phi(z))(u, y)| \leqq M_{3,R}|x-z||u||y|. \tag{1.26}$$

It follows that $E\phi(x) = \phi_{xx}(x)$ and

$$\phi_{n_k xx}(x)(u, y) \to \phi_{xx}(x)(u, y). \tag{1.27}$$

It is also clear that

$$\phi_{nxx}(x)(u, y) \to \phi_{xx}(x)(u, y) \tag{1.28}$$

uniformly on bounded sets on H.

To prove that $|\phi|_Z \leqq M$ we proceed as follows. Let $\{e_i\}$ be an orthonormal basis in H such that

$$Se_i = \lambda_i e_i, \qquad \sum_{i=0}^{\infty} |\lambda_i| < \infty. \tag{1.29}$$

We have

$$\operatorname{Tr}(S\phi_{n,xx}(x)) = \sum_{i=0}^{\infty} \lambda_i \phi_{n,xx}(e_i, e_i) \tag{1.30}$$

which along with (1.28) and some simple calculations implies the claimed conclusion.

In the sequel we shall denote by $B([0, T]; C^h(H))$ the space of all $\phi : [0, T]\times H \to \mathbb{R}$ such that $(\partial^i\phi/\partial x^i)(z_1, z_2, \cdots, z_i)$ is continuous in $[0, T]\times H$, for $z_1, z_2, \cdots, z_i \in H$;

besides for any $R>0$ we have

$$\sup_{\substack{t\in[0,T]\\ |x|\leqq R}}\left|\frac{\partial^i\phi(t,x)}{\partial x^i}\right|<+\infty, \qquad i=0,1,\cdots,h, \quad t\in[0,T].$$

We set moreover

$$B([0,T];X)=\{\phi\in B([0,T];C(H));\ \sup_{t\in[0,T]}|\phi(t,\cdot)|_X<+\infty\},$$

$$B([0,T];Y)=\{\phi\in B([0,T];C^1(H));\ \sup_{t\in[0,T]}|\phi(t,\cdot)|_Y<+\infty\},$$

$$B([0,T];Z)=\{\phi\in B([0,T];C^2(H));\ \sup_{t\in[0,T]}|\phi(t,\cdot)|_Z<+\infty\}.$$

Let $\{\Omega,\mathcal{F},P\}$ be a complete probability space and let W be a H-valued Brownian motion. Let $\{e_k\}$ be an orthonormal basis in H and assume that W is given by

$$W(t)=\sum_{i=0}^{\infty}\sqrt{\lambda_i}\,\beta_i(t)e_i, \tag{1.31}$$

where $\lambda_i\geqq 0, i=1,2,\cdots,\sum_{i=0}^{\infty}\lambda_i<\infty$ and $\{\beta_i(t)\}$ are scalar Brownian motions mutually independent. Let S be the nuclear positive operator defined by

$$Se_i=\lambda_i e_i, \qquad i=1,\cdots. \tag{1.32}$$

We note (see [6]) that

$$\operatorname{Cov}(W_t)=tS. \tag{1.33}$$

In the sequel we shall denote by $L_w^2(0,T;H)$ (resp. $M_w^2(0,T;H)$) the space of all nonanticipative mappings $x:[0,T]\times\Omega\to H$ with respect to W, such that

$$P\left(\int_0^T|x(t)|^2\,dt<\infty\right)=1 \tag{1.34}$$

resp.

$$E\left(\int_0^T|x(s)|^2\,ds<+\infty\right). \tag{1.35}$$

For other concepts and fundamental results on Brownian motion we refer the reader to [6], [10], [11], [12].

2. Preliminaries on convex functions. In this section we recall for later use some definitions and elementary properties of some spaces of convex functions. For general concepts and results on convex analysis we refer to [1] and [5].

We shall denote by K the set of all convex functions $\phi\in C(H)$. For any $\phi\in K$ denote by ϕ_α the function

$$\phi_\alpha(x)=\inf\{(2\alpha)^{-1}|x-y|^2+\phi(y);\ y\in H\}, \qquad \alpha>0 \tag{2.1}$$

and recall that $\phi_\alpha\in K\cap C^1(H)$.

For any $\phi\in K$ denote by $\partial\phi:H\to H$ the subdifferential of ϕ, i.e.,

$$\partial\phi(x)=\{x^*\in H;\ \langle x^*,x-y\rangle\geqq\phi(x)-\phi(y),\ \forall y\in H\}.$$

If $\phi \in C^1(H)$ then $\partial\phi$ is single valued and $\partial\phi = \phi'$ (ϕ' is the derivative of ϕ). The map $\partial\phi : H \to H$ is maximal monotone, i.e.,

$$\langle x^* - y^*, x - y\rangle \geqq 0 \quad \text{for all } x^* \in \partial\phi(x),\ y^* \in \partial\phi(y)$$

and the range $R(1+\alpha\partial\phi)$ is all of H (1 is the identity operator).

In particular this implies that

$$x_\alpha = (1+\alpha\ \partial\phi)^{-1} \cdot x$$

exists for all $\alpha > 0$ and moreover $|x_\alpha - \bar{x}_\alpha| \leqq |x - \bar{x}|$ for any $x, \bar{x} \in H$. Also we have

$$\phi_\alpha(x) = \phi(x_\alpha) + (2\alpha)^{-1}|x - x_\alpha|^2 \quad \forall \alpha > 0,\ x \in H \tag{2.2}$$

and

$$F_\alpha(x) = F(x_\alpha) = \alpha^{-1}(x - x_\alpha), \tag{2.3}$$

where $F = \partial\phi$ and $F_\alpha = \phi'_\alpha$.

Assume now that $\phi \in K \cap C^2(H)$. Since $x = x_\alpha + \alpha F(x_\alpha)$ we have

$$1 = x'_\alpha + \alpha F'(x_\alpha) \cdot x'_\alpha,$$

where x'_α is the derivative of the operator $x \to x_\alpha$. Hence

$$x'_\alpha = (1+\alpha F'(x_\alpha))^{-1}. \tag{2.4}$$

In the next lemma we gather for later use some immediate properties of x_α and F_α.

LEMMA 3. *For any $\phi \in K \cap C^2(H)$ and $x, y \in H$ we have*

$$|x_\alpha| \leqq |x| + \alpha|F(0)|, \tag{2.5}$$

$$|F_\alpha(x)| \leqq |F(x)|, \tag{2.6}$$

$$|F'_\alpha(x)| \leqq |F'(x)|, \tag{2.7}$$

$$|F'_\alpha(x) - F'_\alpha(y)| \leqq |F'(x_\alpha) - F'(y_\alpha)|, \tag{2.8}$$

Proof. The proof is well known but we sketch it for the reader's convenience. Since F is monotone, we have the inequality

$$0 \leqq \langle F(x_\alpha) - F(0), x_\alpha\rangle = \alpha^{-1}\langle x - x_\alpha, x_\alpha\rangle - \langle F(0), x_\alpha\rangle$$

which implies

$$|x_\alpha|^2 \leqq |x_\alpha|(|x| + \alpha|F(0)|)$$

and (2.5) follows. To prove (2.6) we notice that

$$\begin{aligned} |F_\alpha(x)| &\leqq |\alpha^{-1}(x - (1+\alpha F)^{-1}x)| \\ &= \alpha^{-1}|(1+\alpha F)^{-1}[(1+\alpha F)x - x]| \leqq |F(x)|. \end{aligned}$$

As regards (2.7) it follows by (2.4) because F' is a positive operator. Finally, again by (2.4), we have

$$|x'_\alpha - y'_\alpha| \leqq \alpha|F'(x_\alpha) - F'(y_\alpha)|$$

while by (2.3)

$$|F'_\alpha(x) - F'_\alpha(y)| = \alpha^{-1}|x'_\alpha - y'_\alpha| \leqq |F'(x_\alpha) - F'(y_\alpha)|$$

which yields (2.8) as claimed.

For any $\varepsilon > 0$ and $\phi \in C^1(H) \cap K$ we set

$$R_{\phi,\alpha}(x) = \alpha^{-1}(\phi(x) - \phi_\alpha(x)) - \tfrac{1}{2}|\phi'(x)|^2. \tag{2.9}$$

LEMMA 4. *We have*

$$|R_{\phi,\alpha}(x)| \leqq |F(x)| \int_0^1 |F(x - \alpha t F(x_\alpha)) - F(x)|\, dt + \frac{1}{2}|F(x) - F(x_\alpha)|^2 \tag{2.10}$$

and

$$|R'_{\phi,\alpha}(x)| \leqq |F(x)| \int_0^1 |F'(x - \alpha t F(x_\alpha)) - F'(x)|\, dt + |F'(x)||F(x_\alpha) - F(x)| \tag{2.11}$$

where $F = \phi'$.

Proof. To prove (2.10) it suffices to notice the equality

$$\begin{aligned} R_{\phi,\alpha}(x) &= \frac{1}{\alpha}(\phi(x) - \phi(x_\alpha)) - \frac{1}{2}(|F(x_\alpha)|^2 + |F(x)|^2) \\ &= \int_0^1 \langle F(x - \alpha t F(x_\alpha)), F(x_\alpha)\rangle\, dt - \frac{1}{2}(|F(x_\alpha)|^2 + |F(x)|^2) \\ &= \int_0^1 \langle F(x - \alpha t F(x_\alpha)) - F(x), F(x_\alpha)\rangle\, dt - \frac{1}{2}|F(x) - F(x_\alpha)|^2. \end{aligned}$$

Finally (2.11) follows from the identity

$$\begin{aligned} R'_{\phi,\alpha}(x) &= \frac{1}{\alpha}(F(x) - F_\alpha(x)) - F'(x) \cdot F(x) \\ &= \int_0^1 F'(x - \alpha t F(x_\alpha)) \cdot F(x_\alpha)\, dt - F'(x) \cdot F(x) \\ &= \int_0^1 [F'(x - \alpha t F(x_\alpha)) - F'(x)] \cdot F(x_\alpha)\, dt + F'(x)(F(x_\alpha) - F(x)). \end{aligned}$$

LEMMA 5. *Assume that* $\phi, \bar{\phi} \in K \cap C^2(H)$. *Then for all* $x \in H$

$$\phi_\alpha(x) - \bar{\phi}_\alpha(x) \leqq \phi(\bar{x}_\alpha) - \bar{\phi}(\bar{x}_\alpha), \tag{2.12}$$

$$|x_\alpha - \bar{x}_\alpha| \leqq \alpha |F(x_\alpha) - \bar{F}(x_\alpha)|, \tag{2.13}$$

$$|F_\alpha(x) - \bar{F}_\alpha(x)| \leqq |F(x_\alpha) - \bar{F}(x_\alpha)|. \tag{2.14}$$

Proof. By (2.1) we have

$$\phi_\alpha(x) - \bar{\phi}_\alpha(x) = \inf\left\{\phi(y) \frac{1}{2\alpha}|x - y|^2;\ y \in H\right\} - \bar{\phi}(\bar{x}_\alpha) - \frac{\alpha}{2}|\bar{F}(\bar{x}_\alpha)|^2,$$

which clearly implies (2.12). Next we have

$$x_\alpha - \bar{x}_\alpha = (1 + \alpha\bar{F})^{-1}(x + \alpha(\bar{F}(x_\alpha) - F(x_\alpha))) - (1 + \alpha\bar{F})^{-1}x$$

and by (2.2)

$$|x_\alpha - \bar{x}_\alpha| \leqq \alpha |\bar{F}(x_\alpha) - F(x_\alpha)|.$$

Finally by the identity

$$F_\alpha(x)-\bar{F}_\alpha(x)=\alpha^{-1}(x_\alpha-\bar{x}_\alpha)$$

we find (2.14) as claimed.

LEMMA 6. *Assume that $\phi\in C^1(H)\cap K$. Then for every $k=0,1,2,\cdots$ there exists a continuous positive function C_k such that*

$$1+|x_\alpha|^k\leqq(1+|x|^{2k})(1+\alpha C_k(|F(0)|)) \tag{2.15}$$

for all $x\in H$ and $\alpha>0$.

Proof. It is a simple consequence of (2.5).

PROPOSITION 1. *Assume that $\phi,\bar{\phi}\in C^2(H)\cap K$. Then for all $n=0,1,\cdots$ we have*

$$|\phi_\alpha|_{0,n}\leqq(1+\alpha C_n(|F(0)|))|\phi|_{0,n}, \tag{2.16}$$

$$|\phi_\alpha|_{1,n}\leqq|\phi|_{1,n},\qquad |\phi_\alpha|_{2,n}\leqq|\phi|_{2,n}, \tag{2.17}$$

$$\|\phi_\alpha\|_{2,n}\leqq(1+\alpha C_n(|F(0)|))\|\phi\|_{2,n}, \tag{2.18}$$

$$|\phi_\alpha-\bar{\phi}_\alpha|_{0,n}\leqq(1+\alpha C_n(|F(0)|))|\phi-\bar{\phi}|_{0,n}, \tag{2.19}$$

$$|\phi_\alpha-\bar{\phi}_\alpha|_{1,n}\leqq(1+\alpha C_n(|\bar{F}(0)|))|\phi-\bar{\phi}|_{1,n}. \tag{2.20}$$

Moreover, if $|\phi|_z,|\bar{\phi}|_z\leqq\lambda$ then there exists $C(\lambda)>0$ such that

$$\begin{aligned}&|\phi_\alpha-\bar{\phi}_\alpha|_{0,n_0}+|\phi_\alpha-\bar{\phi}_\alpha|_{1,n_1}+|\phi_\alpha-\bar{\phi}_\alpha|_{2,n_1+n_3}\\&\quad\leqq(1+\alpha C(\lambda))\{|\phi-\bar{\phi}|_{0,n_0}+|\phi-\bar{\phi}|_{1,n_1}+|\phi-\bar{\phi}|_{2,n_1+n_3}\}.\end{aligned} \tag{2.21}$$

Proof. From (2.12) and (2.15) the below inequalities follow

$$\begin{aligned}\frac{\phi_\alpha(x)-\bar{\phi}_\alpha(x)}{1+|x|^{2n}}&\leqq\frac{\phi(x_\alpha)-\bar{\phi}(\bar{x}_\alpha)}{1+|\bar{x}_\alpha|^{2n}}\cdot\frac{1+|\bar{x}_\alpha|^{2n}}{1+|x|^{2n}}\\&\leqq|\phi-\bar{\phi}|_{0,n}(1+\alpha C_n(|\bar{F}(0)|)),\end{aligned}$$

which imply (2.20) and (2.16). Estimates (2.17) are immediate consequences of (2.6) and (2.7); the other inequalities are simple (although tedious) consequences of properties of ϕ_α.

PROPOSITION 2. *Assume that $\phi\in C^2(H)\cap K$ and that n_0, n_1, n_2, are nonnegative integers such that*

$$n_0\geqq2n_1(1+n_2),\qquad n_0\geqq n_1\geqq n_2. \tag{2.22}$$

Then there exists a continuous increasing mapping $\gamma:\mathbb{R}^2\to\mathbb{R}_+$ such that

$$|R_{\phi,\alpha}|_{0,n_0}\leqq\alpha\gamma(|\phi|_{1,n_1},|\phi|_{2,n_2}). \tag{2.23}$$

Proof. We have

$$\begin{aligned}|F(x-\alpha tF(x_\alpha))-F(x)|&\leqq|\phi|_{2,n_2}\alpha|F(x)|\{1+[|x|\vee|x-\alpha tF(x_\alpha)|]^{2n_2}\}\\&\leqq\alpha|\phi|_{2,n_2}|\phi|_{1,n_1}(1+|x|^{2n_1})\{1+[|x|+|F(x)|]^{2n_2}\}\\&\leqq\alpha|\phi|_{2,n_2}|\phi|_{1,n_1}(1+|x|^{2n_1})\{1+[|x|+|\phi|_{1,n_1}(1+|x|^{2n_1})]\}^{2n_2}\\&\leqq\alpha\gamma_1(|\phi|_{1,n_1},|\phi|_{2,n_2})(1+|x|^{2n_1+4n_1n_2}).\end{aligned} \tag{2.24}$$

Moreover, one has

$$
\begin{aligned}
|F(x)-F(x_\alpha)| &\leqq |\phi|_{2,n_2}|x-x_\alpha|\{1+[|x|\vee|x_\alpha|]^{2n_2}\} \\
&\leqq \alpha|\phi|_{2,n_2}|\phi|_{1,n_1}(1+|x|^{2n_1})\{1+[|x|+|F(0)|]^{2n_2}\} \\
&\leqq \alpha\gamma_2(|\phi|_{1,n_1},|\phi|_{2,n_2})(1+|x|^{2n_1+2n_2}).
\end{aligned} \tag{2.25}
$$

From (2.24) and (2.25) the conclusion follows, by virtue of (2.10).

PROPOSITION 3. *Assume that $\phi\in C^2(H)\cap K$ and $n_0\geqq n_1\geqq n_2\geqq n_3\geqq 0$. Let m_1 be a positive integer such that*

$$
m_1\geqq(2n_1+2n_1n_3)\vee(n_1+2n_2). \tag{2.26}
$$

Then there exists a continuous increasing mapping $\eta:\mathbb{R}^3\to\mathbb{R}_+$ such that

$$
|R_{\phi,\alpha}|_{1,m_1}\leqq\alpha\eta(|\phi|_{1,n_1},|\phi|_{2,n_2},\|\phi\|_{2,n_3}). \tag{2.27}
$$

Proof. We have

$$
\begin{aligned}
&|F'(x-\alpha tF(x_\alpha))-F'(x)| \\
&\quad\leqq\alpha\|\phi\|_{2,n_3}\cdot|\phi|_{1,n_1}(1+|x|^{2n_1})\{1+[|x|+|F(x)|]^{2n_3}\} \\
&\quad\leqq\alpha\|\phi\|_{2,n_3}|\phi|_{1,n_1}(1+|x|^{2n_1})\{1+[|x|+|\phi|_{1,n_1}(1+|x|^{2n_1})]^{2n_3}\} \\
&\quad\leqq\alpha\eta_1(|\phi|_{1,n_1},|\phi|_{2,n_2},\|\phi\|_{2,n_3})(1+|x|^{2n_1+4n_1n_3}).
\end{aligned} \tag{2.28}
$$

Recalling (2.11), (2.25) we get

$$
\begin{aligned}
|R'_{\phi,\alpha}(x)| &\leqq\alpha\eta_3(|\phi|_{1,n_1},|\phi|_{2,n_2},\|\phi\|_{2,n_3}) \\
&\quad\cdot\{1+|x|^{4n_1+4n_1n_3}+|x|^{2n_1+4n_2}\},
\end{aligned}
$$

as claimed.

3. The linearized problem. We shall study here the linear Cauchy problem:

$$
\begin{aligned}
&\phi_t(t,x)-\langle Ax,\phi_x(t,x)\rangle-\frac{\varepsilon}{2}\operatorname{Tr}(S\phi_{xx}(t,x))=0, \\
&\phi(0,x)=\phi_0(x),
\end{aligned} \tag{3.1}
$$

where $\phi_0\in Z$, $\varepsilon>0$ and $A:D(A)\subset H\to H$ is the infinitesimal generator of a C_0-semigroup e^{At} of contractions on H, i.e.,

$$
|e^{tA}|\leqq 1\quad\text{for all } t\geqq 0. \tag{3.2}
$$

By a solution to problem (3.1) we mean a function $\phi\in B([0,T];Z)$ which belongs to $C^1[0,T]$ for each $x\in D(A)$ and satisfies (3.1) for all $x\in D(A)$ and all $t\in[0,T]$. Consider the approximating problem

$$
\begin{aligned}
&\phi^n_t(t,x)-\langle A_nx,\phi^n_x(t,x)\rangle-\frac{\varepsilon}{2}\operatorname{Tr}(S\phi^n_{xx}(t,x))=0, \\
&\phi^n(0,x)=\phi_0(x),
\end{aligned} \tag{3.3}
$$

where $A_n=n^2(n-A)^{-1}=nA(n-A)^{-1}$ (the Yosida approximation of A).

It is well known that $\exp(tA_n)x\to\exp(tA)x$ for every x in H and $A_nx\to Ax$ for every x in $D(A)$ (see for instance [6]). Moreover, since A_n is bounded, it is easy to prove many properties in equations involving A_n (for example Itô's formula for $\phi^n(t,u)$) and afterward to pass to limit as n goes to infinity.

LEMMA 7. *For every* $\phi_0 \in Z$, *problem* (3.3) *has a unique solution* $\phi^n \in C^1([0,T]\times H)\cap B([0,T];Z)$ *given by the formula*

$$\phi^n(t,x) = E\phi_0\Big(e^{tA_n}x + \sqrt{\varepsilon}\int_0^t e^{sA_n}\,dW_{T-s}\Big). \tag{3.4}$$

Proof. For existence we first remark that $\zeta_t = W_T - W_{T-t}$ is a Brownian motion. We have:

$$\phi^n_x(t,x) = e^{tA_n^*}E\phi_{0,x}\Big(e^{tA_n}x - \sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta(s)\Big), \tag{3.5}$$

$$\phi^n_{xx}(t,x) = e^{tA_n^*}E\phi_{0,xx}\Big(e^{tA_n}x - \sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta(s)\Big)\,e^{tA_n}. \tag{3.6}$$

We notice that if $\phi \in C^2(H)$ then $\phi''(x)\in \mathscr{L}(H,\mathscr{L}(H,\mathbb{R}))$ and so we may write $(\phi''(x)\cdot y)\cdot z = \phi''(x)\cdot(y,z) = \langle \phi''(x)\cdot y, z\rangle$. In this sense we have $(\phi P)''(x) = P^*\phi''(x)P$ for all $P\in\mathscr{L}(H)$. To prove that Φ is differentiable with respect to t we notice that for each $h>0$ one has

$$\begin{aligned}
&\phi_0\Big(e^{(t+h)A_n}x - \sqrt{\varepsilon}\int_0^{t+h} e^{sA_n}\,d\zeta_s\Big) - \phi_0\Big(e^{tA_n}x - \sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta_s\Big)\\
&= \Big\langle e^{(t+h)A_n}x - e^{tA_n}x - \sqrt{\varepsilon}\int_t^{t+h} e^{sA_n}\,d\zeta_s,\ \phi_{0x}\Big(e^{tA_n}x - \sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta_s\Big)\Big\rangle\\
&\quad + \frac{1}{2}\phi_{0xx}\Big(e^{tA_n}x - \sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta_s\Big)\\
&\qquad\cdot\Big[e^{(t+h)A_n}x - e^{tA_n}x - \sqrt{\varepsilon}\int_t^{t+h} e^{sA_n}\,d\zeta_s\Big]^2\\
&\quad + \int_0^1\Big\{(1-a)\phi_{0xx}\Big((1-a)\Big[e^{tA_n}x - \sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta_s\Big]\\
&\qquad + a\Big[e^{(t+h)A_n}x - \sqrt{\varepsilon}\int_0^{t+h} e^{sA_n}\,d\zeta_s\Big]\Big)\\
&\qquad\cdot\Big[e^{(t+h)A_n}x - e^{tA_n}x - \sqrt{\varepsilon}\int_t^{t+h} e^{sA_n}\,d\zeta_s\Big]^2\Big\}\,da.
\end{aligned} \tag{3.7}$$

To calculate $\phi_t(t,x)$ we remark that

$$\lim_{h\to 0}\frac{1}{h}\,(e^{(t+h)A_n}x - e^{tA_n}x) = A_n\,e^{tA_n}x \tag{3.8}$$

and

$$E\Big\langle\int_t^{t+h} e^{sA_n}\,d\zeta_s,\ \phi_{0x}\Big(e^{tA_n}x - \sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta_s\Big)\Big\rangle = 0, \tag{3.9}$$

since ζ_s have independent increments. Moreover, one has

$$\lim_{h\to 0}\frac{1}{h}|e^{(t+h)A_n}x - e^{tA_n}x|^2 = 0, \tag{3.10}$$

$$\lim_{h\to 0} E\left\{\frac{1}{2h}\phi_{0xx}\left(e^{tA_n}x-\sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta_s\right)\left[\int_t^{t+h} e^{sA_n}\,d\zeta_s\right]^2\right\}$$

$$=\lim_{h\to 0}\left\{\frac{\varepsilon}{2}\int_t^{t+h}\operatorname{Tr}\left(S e^{sA_n^*}\phi_{0xx}\left(e^{-tA_n}x-\sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta_s\right)e^{sA_n}\,ds\right)\right\} \tag{3.11}$$

$$=\frac{\varepsilon}{2}\operatorname{Tr}\left[S e^{tA_n}\phi_{0xx}\left(e^{tA_n}x-\sqrt{\varepsilon}\int_0^t e^{sA_n}\,d\zeta_s\right)e^{tA_n}\right].$$

Observe also that the last integral in (3.7) goes to 0 for $h\to 0$ by the Lebesgue dominated convergence theorem.

It follows that

$$D_t^+\phi^n(t,x)=\langle A_n x,\phi_x^n(t,x)\rangle+\frac{\varepsilon}{2}\operatorname{Tr}\left(S\phi_{xx}^n(t,x)\right), \tag{3.12}$$

where D_t^+ means the right derivative. Since the right-hand side of (3.12) is continuous we conclude by a standard result, that (3.3) holds.

Uniqueness. Let ϕ be a solution to problem (3.3). We set $\psi(t,x)=\phi(T-t,x)$. Then ψ is a solution to the backward problem

$$\begin{aligned}&\psi_t(t,x)+\langle A_n x,\psi_x(t,x)\rangle+\frac{\varepsilon}{2}\operatorname{Tr}\left(S\psi_{xx}(t,x)\right)=0,\\ &\psi(T,x)=\phi_0(x).\end{aligned} \tag{3.13}$$

Let $u=u(s,t,x)$ be the solution to the stochastic differential equation

$$du=A_n u\,ds+\sqrt{\varepsilon}\,dW_s,\qquad u(t)=x, \tag{3.14}$$

i.e.,

$$u(s,t,x)=e^{(s-t)A_n}x+\sqrt{\varepsilon}\int_t^s e^{(s-\sigma)A_n}\,dW_\sigma=u(s). \tag{3.15}$$

By the Itô formula

$$d\psi(s,u)=\left[\psi_s(s,u)+\frac{\varepsilon}{2}\operatorname{Tr}\left(S\psi_{xx}(s,u)\right)\right]ds+\psi_x(s,u)\,du,$$

from which, by integrating in $[t,T]$ and taking the expectation we obtain

$$\psi(t,x)=E\psi(t,u(t,t,x))=E\psi(T,u(T,t,x))=E\phi_0(u(T,t,x)),$$

and therefore $\psi=\phi^n$ as claimed. The following corollary follows via a standard variation of constants formula.

COROLLARY 1. *Under the assumptions of Lemma 7, the problem*

$$\begin{aligned}&\phi_t^n(t,x)-\langle A_n x,\phi_x^n(t,x)\rangle-\frac{\varepsilon}{2}\operatorname{Tr}\left(S\phi_{xx}^n(t,x)\right)=\zeta(t,x),\\ &\phi^n(0,x)=\phi_0(x)\end{aligned} \tag{3.16}$$

has for every $\zeta\in B([0,T];Z)$ a unique solution ϕ^n given by

$$\begin{aligned}\phi^n(t,x)=&E\phi_0\left(e^{tA_n}x+\sqrt{\varepsilon}\int_0^t e^{sA_n}\,dW_{T-s}\right)\\ &+E\int_0^t\zeta\left(s,e^{(t-s)A_n}x+\sqrt{\varepsilon}\int_0^{t-s}e^{\sigma A_n}\,dW_{T-\sigma}\right)ds.\end{aligned} \tag{3.17}$$

PROPOSITION 4. *For every $\phi_0 \in Z$ problem* (3.1) *has a unique solution $\phi \in B([0, T]; Z)$ given by*

$$\phi(t, x) = E\phi_0\left(e^{tA}x + \sqrt{\varepsilon}\int_0^t e^{sA}\, dW_{T-s}\right). \tag{3.18}$$

Moreover, for each nonnegative integer m there exists $\omega_m > 0$ such that

$$|\phi(t, \cdot)|_{i,m} \leqq e^{\varepsilon\omega_m t}|\phi_0|_{i,m}, \qquad i = 0, 1, \cdots, \tag{3.19}$$

$$\|\phi(t, \cdot)\|_{i,m} \leqq e^{\varepsilon\omega_{m+1} t}\|\phi_0\|_{i,m}, \qquad i = 0, 1, \cdots. \tag{3.20}$$

Proof. Let ϕ^n given by (3.4). Inasmuch as for each $x \in H$, $e^{tA_n}x \to e^{tA}x$ uniformly on compacts we see that

$$\begin{aligned} &\phi^n(t, x) \to \phi(t, x), \\ &\operatorname{Tr}(S\phi^n_{xx}(t, x) \to \operatorname{Tr}(S\phi_{xx}(t, x), \end{aligned} \qquad \text{uniformly in } [0, T],\ T > 0.$$

Moreover, since for all $x \in D(A)\langle A_n x, \phi^n_x(t, x)\rangle \to \langle Ax, \phi_x(t, x)\rangle$ uniformly in $[0, T]$ we infer that $\phi^n_t(t, x) \to \phi_t(t, x)$ uniformly on $[0, T]$ and therefore ϕ satisfies (3.1) for all $t \in [0, T]$ and $x \in D(A)$ (we note that in this case ϕ is differentiable as a function of t for each $x \in D(A)$).

For uniqueness let $\eta \in B([0, T]; Z)$ be another solution to problem (3.1). For $x \in H$ we set $x_n = (n - A)^{-1}nx$ and notice the equation

$$\begin{aligned} \eta_t(t, x_n) - \langle A_n x_n, \eta_x(t, x_n)\rangle - \frac{\varepsilon}{2}\operatorname{Tr}(S\eta_{xx}(t, x_n)) \\ = \langle (A - A_n)x_n, \eta_x(t, x_n)\rangle. \end{aligned}$$

Then by Corollary 1,

$$\begin{aligned} \eta(t, x_n) = E\phi_0\left(e^{tA_n}x_n + \sqrt{\varepsilon}\int_0^t e^{sA_n}\, dW_{T-s}\right) \\ + E\int_0^t \left\langle (A - A_n)x_n, \eta_x\left(e^{(t-s)A_n}x_n + \sqrt{\varepsilon}\int_0^{t-s} e^{sA_n}\, dW_{T-\sigma}\right) ds\right\rangle. \end{aligned}$$

Letting n tend to $+\infty$ we get $\eta = \phi$ as claimed. To prove (3.19) we shall restrict ourselves to the case $i = 0$ the other cases being similar. Since $M_t = \int_0^t e^{sA}\, dW_{T-s}$ is a martingale it follows (see for instance [12]) that for each $m \in N$ there exists $\gamma_m > 0$ such that

$$E\left|\int_0^t e^{sA}\, dW_{T-s}\right|^{2m} \leqq \gamma_m t^m. \tag{3.21}$$

We have

$$|\phi(t, x)| \leqq |\phi_0|_{0,m}E\left(1 + \left|e^{tA}x + \sqrt{\varepsilon}\int_0^t e^{sA}\, dW_{T-s}\right|^{2m}\right).$$

It follows that there exist real bounded functions a_e, $e = 2, 3, \cdots, 2m$ such that

$$|\phi(t, x)| \leqq |\phi_0|_{0,m}E\left\{1 + \sum_{e=2}^{2m} a_e(|x|)\sqrt{\varepsilon}^e \left|\int_0^t e^{-sA}\, dW_{t-s}\right|^e\right\}(1 + |x|^{2m}).$$

By virtue of (3.21) there exists real bounded functions b_e such that

$$|\phi(t,x)| \leqq |\phi_0|_{0,m}(1+|x|^{2m})\Big(1+\sum_{e=2}^{2m} b_e(\varepsilon t)^{e/2}\Big),$$

so (3.19) is proved. The proof of (3.20) is completely similar, so it will be omitted. We shall consider now the nonhomogeneous Cauchy problem,

$$\begin{aligned} &\phi_t(t,x) - \langle Ax, \phi_x(t,x)\rangle - \frac{\varepsilon}{2}\operatorname{Tr}(S\phi_{xx}(t,x)) = g(t,x), \\ &\phi(0,x) = \phi_0(x), \end{aligned} \tag{3.22}$$

where $\phi_0 \in Z$ and $g \in B([0,T];Z)$.

By a solution to (3.22) we mean a function $\phi \in B([0,T];Z)$ which belongs to $W^{1,\infty}(0,T)$ as a function of t (for each $x \in D(A)$) and satisfies (3.22) for all $x \in D(A)$ and a.e. $t \in]0,T[$.

For later use, we notice the following existence result.

PROPOSITION 5. *For every $\phi_0 \in Z$ and $g \in B([0,T];Z)$ problem (3.22) has a unique solution $\phi \in B([0,T];Z)$ given by the formula*

$$\begin{aligned} \phi(t,x) = {} & E\phi_0\Big(e^{tA}x + \sqrt{\varepsilon}\int_0^t e^{sA}\,dW_{T-s}\Big) \\ & + E\int_0^t g\Big(s, e^{(t-s)A}x + \sqrt{\varepsilon}\int_0^{t-s} e^{\sigma A}\,dW_{T-\sigma}\Big)\,ds. \end{aligned} \tag{3.23}$$

Proof. Existence follows from Proposition 4. To prove uniqueness, arguing as in the proof of Lemma 7 it suffices to assume that A is bounded.

Let $\phi \in B([0,T];Z)$ be a solution to (3.22) where $g=0$ and $\phi_0=0$ and let $\psi(t,x) = \phi(T-t,x)$. Finally set $\psi^n(t,x) = \psi(t,P^nx)$, where

$$P^nx = \sum_{i=1}^{n} \langle x, e_i\rangle e_i. \tag{3.24}$$

Clearly $\psi^n \in C^1([0,T]\times H;\mathbb{R}) \cap C([0,T];Z)$, so we may apply the Itô formula to $\psi^n(s,u)$ (see for instance [11]) and get

$$d\psi^n(s,u) = \Big(\psi^n_s(s,u) + \frac{\varepsilon}{2}\operatorname{Tr}(S\psi^n_{xx}(s,u))\Big)\,ds + \psi^n_u(s,u)\,dW_s. \tag{3.25}$$

Integrating and taking the expectation we obtain

$$0 = E\psi^n(T,u(T)) = \psi^n(t,x) + E\int_t^T \Big[\psi^n_s(s,u(s)) + \frac{\varepsilon}{2}\operatorname{Tr}(S\psi^n_{xx}(s,u(s)))\Big]\,ds, \tag{3.26}$$

where u is the solution to (3.1) (with $A_n = A$).

As n goes to infinity we get

$$\psi(t,x) = -E\int_t^T \Big(\psi_s(s,u(s)) + \frac{\varepsilon}{2}\operatorname{Tr}(S\psi_{xx}(s,u(s)))\Big)\,ds = 0,$$

as claimed.

4. The main results. Consider the Cauchy problem

$$\begin{aligned} &\phi_t(t,x)+\frac{1}{2}|\phi_x(t,x)|^2-\langle Ax,\phi_x(t,x)\rangle-\frac{\varepsilon}{2}\operatorname{Tr}(S\phi_{xx}(t,x))=g(t,x),\\ &\phi(0,x)=\phi_0(x), \qquad \varepsilon>0, \end{aligned} \tag{4.1}$$

under the following assumptions:

(4.2) a) A is the infinitesimal generator of a C_0 semigroup of contractions;
b) $\phi_0\in Z\cap K$; $g\in B([0,T];Z)\cap\mathscr{K}$;
c) $n_0\geqq 2n_1(1+n_2)$;

where $\mathscr{K}=\{\phi\in B([0,T];C(H));\ \phi(t)\in K,\ \forall t\in[0,T]\}$.

We shall consider the approximating problem

$$\begin{aligned} &\phi^\alpha_t(t,x)+\frac{1}{\alpha}(\phi^\alpha(t,x)-\phi^\alpha_\alpha(t,x))-\langle Ax,\phi^\alpha_x(t,x)\rangle\\ &\qquad -\frac{1}{2}\varepsilon\operatorname{Tr}(S\phi^\alpha_{xx}(t,x))=g(t,x),\\ &\phi^\alpha(0,x)=\phi_0(x), \end{aligned} \tag{4.3}$$

where $\alpha\in\,]0,1]$ and ϕ^α_α is defined by (2.1), i.e.,

$$\phi^\alpha_\alpha(t,x)=\inf\left\{\phi^\alpha(t,y)+\frac{1}{2\alpha}|x-y|^2;\ y\in H\right\}.$$

A weak form of problem (4.3) is given by the following integral equation:

$$\begin{aligned} \phi^\alpha(t,x)&=\exp(-\alpha^{-1}t)E\phi_0\left(e^{tA}x+\sqrt{\varepsilon}\int_0^t e^{\sigma A}\,dW_{T-s}\right)\\ &\quad+E\int_0^t\exp(-\alpha^{-1}(t-s))(\alpha^{-1}\phi^\alpha_\alpha+g)\\ &\qquad\cdot\left(s,e^{(t-s)A}+\sqrt{\varepsilon}\int_0^{t-s}e^{\sigma A}\,dW_{T-s}\right)ds. \end{aligned} \tag{4.4}$$

PROPOSITION 6. *Under assumption* (4.2) *for every* $\alpha>0$, (4.4) *has a unique solution* $\phi^\alpha\in B([0,T];Z)$. *Moreover* ϕ^α *satisfies* (4.3) *for all* $(t,x)\in[0,T]\times D(A)$.

Finally there exist $\tilde{\omega}_i\geqq 0$, $i=0,1,2,3$ *such that*

$$\begin{aligned} |\phi^\alpha(t,\cdot)|_{i,n_i}&\leqq\exp(\varepsilon\tilde{\omega}_i t)|\phi_0|_{i,n_i}\\ &\quad+\int_0^t\exp(\varepsilon\tilde{\omega}_i(t-s))|g(s,\cdot)|_{i,n_i}\,ds, \qquad i=0,1,2, \end{aligned} \tag{4.5}$$

$$\begin{aligned} \|\phi^\alpha(t,\cdot)\|_{2,n_3}&\leqq\exp(\varepsilon\tilde{\omega}_3 t)\|\phi_0\|_{2,n_3}\\ &\quad+\int_0^t\exp(\varepsilon\tilde{\omega}_3(t-s))\|g(s,\cdot)\|_{2,n_3}\,ds. \end{aligned} \tag{4.6}$$

Proof. Set

$$\begin{aligned} \phi^0(t,x)&=e^{-t/\alpha}E\phi_0\left(e^{tA}x+\sqrt{\varepsilon}\int_0^t e^{sA}\,dW_{T-s}\right)\\ &\quad+E\int_0^t e^{-(t-s)/\alpha}g\left(s,e^{(t-s)A}x+\sqrt{\varepsilon}\int_0^{t-s}e^{\sigma A}\,dW_{T-\sigma}\right)ds, \end{aligned} \tag{4.7}$$

$$\phi^{n+1}(t,x)=\phi^0(t,x)+\frac{1}{\alpha}E\int_0^t e^{-(t-s)/\alpha}\phi_\alpha^n\left(s,e^{(t-s)A}x+\int_0^{t-s}e^{\sigma A}\,dW_{T-\sigma}\right)ds. \tag{4.8}$$

Using inequalities (2.16)–(2.20) it is not difficult to find $i=0,1,2,3$, such that

$$|\phi^n(t,\cdot)|_{i,n_i}\leqq\exp(\varepsilon\tilde{\omega}_i t)|\phi_0|_{i,n_i}+\int_0^t\exp(\varepsilon\tilde{\omega}_i(t-s))|g(s,\cdot)|_{i,n_i}\,ds,\qquad i=0,1,2, \tag{4.9}$$

$$\|\phi^n(t,\cdot)\|_{2,n_3}\leqq\exp(\varepsilon\tilde{\omega}_3 t)\|\phi_0\|_{2,n_3}+\int_0^t\exp(\varepsilon\tilde{\omega}_3(t-s))\|g(s,\cdot)\|_{2,n_3}\,ds. \tag{4.10}$$

By (4.9) and (4.10) we see that the set

$$\Gamma=\{\phi^n(t,\cdot);\,n\in N,\,t\in[0,T]\}$$

is bounded in Z.

Set now

$$|\phi|_\Sigma=|\phi|_{0,n_0}+|\phi|_{1,n_1}+|\phi|_{2,n_1+n_3}. \tag{4.11}$$

By (2.21) it follows (Γ being bounded in Z) that $\{\phi^n(t,\cdot)\}$ is a Cauchy sequence, with respect to the norm $|\cdot|_\Sigma$, uniformly in t. This implies that $\{\phi^n\}$ converges in $B([0,T];C^2(H))$ to a function ϕ^α. By (4.9) and (4.10) it follows that $\phi^\alpha\in B([0,T];Z)$ and also that (4.5) and (4.6) hold.

Finally, using Proposition 5 it is easy to check that ϕ^α is the unique solution to (4.3) for all $(t,x)\in[0,T]\times D(A)$.

THEOREM 1. *Under assumptions* (4.2) *the Cauchy problem* (4.1) *has a unique solution* $\phi\in B([0,T];X)\cap B([0,T];C^1(H))$ *such that* $\phi(\cdot,x)\in W^{1,\infty}(0,T)$ *for all* $x\in D(A)$. *Moreover, the map* $(\phi_0,g)\to\phi$ *is Lipschitz from* $XXB([0,T];Z)$ *to* $B([0,T];X)$. *Finally the set* $\{\phi(t,\cdot),t\in[0,T]\}$ *is bounded in* Z.

Proof. We shall obtain the solution ϕ to (4.1) as the limit for $\alpha\to0$ of ϕ^α. The first step is the proof of the convergence of ϕ^α. Let $\alpha,\beta>0$. By (2.10) it follows that

$$\phi_t^\beta+\frac{1}{\alpha}(\phi^\beta-\phi_\alpha^\beta)-\langle Ax,\phi_x^\beta\rangle-\frac{\varepsilon}{2}\operatorname{Tr}(S\phi_{xx}^\beta)=g+R_{\phi_\beta,\alpha}-R_{\phi_\beta,\beta},$$

$$\phi^\beta(0,x)=\phi_0(x),$$

and by Proposition 5, we get

$$\phi^\beta(t,x)=e^{-t/\alpha}E\phi_0\left(e^{tA}x+\sqrt{\varepsilon}\int_0^t e^{sA}\,dW_{T-s}\right)+E\int_0^t e^{-(t-s)/\alpha}\left(\frac{1}{\alpha}\phi_\alpha^\beta+R_{\phi_\beta,\alpha}-R_{\phi_\beta,\beta}\right)\cdot\left(s,e^{(t-s)A}x+\sqrt{\varepsilon}\int_0^{t-s}e^{\sigma A}\,dW_{T-\sigma}\right)ds. \tag{4.12}$$

By (2.20), (2.23) and (3.19) we see that

$$
\begin{aligned}
&|\phi^\alpha(t,\cdot)-\phi^\beta(t,\cdot)|_{0,n_0}\\
&\leqq\int_0^t \exp\left(-(t-s)(\alpha^{-1}+\omega_0)\right)\\
&\quad\cdot[\alpha^{-1}(1+\alpha C_{n_0}(\phi^\beta(s,\cdot)))|\phi^\alpha(s,\cdot)-\phi^\beta(s,\cdot)|_{0,n_0},\\
&\qquad\qquad +(\alpha+\beta)\gamma(|\phi^\beta|_{1,n_1},|\phi^\beta|_{2,n_2})]\,ds.
\end{aligned}
\tag{4.13}
$$

By (4.5), (4.6) and the Gronwall lemma it follows that there exists $C>0$ such that

$$
|\phi^\alpha(t,\cdot)-\phi^\beta(t,\cdot)|_{0,n_0}\leqq C(\alpha+\beta),
\tag{4.14}
$$

and all $\{\phi^\alpha(t,\cdot)\}$ belong to a closed bounded subset of Z. Hence there exists $\phi\in B([0,T];X)$ such that $\phi^\alpha\to\phi$ in $B([0,T];X)$.

We shall study the convergence of ϕ_x^α. Choose $m_1\geqq(2n_1+2n_1n_3)\vee(n_1+2n_2)$. Recalling (4.12) and (3.19) it follows by Proposition 3 that

$$
\begin{aligned}
|\phi^\alpha(t,\cdot)-\phi^\beta(t,\cdot)|_{1,m_1}&\leqq\int_0^t \exp\left(-(t-s)(\alpha^{-1}+\omega_0)\right)\\
&\quad\cdot[\alpha^{-1}(1+\alpha C(\phi^\beta(s,\cdot)))|\phi^\alpha(s,\cdot)-\phi^\beta(s,\cdot)|_{1,m_1}\\
&\quad+(\alpha+\beta)\eta(|\phi^\beta(s,\cdot)|_{1,n_1},|\phi(s,\cdot)|_{2,n_2},|\phi^\beta(s,\cdot)|_{2,n_3})]\,ds.
\end{aligned}
\tag{4.15}
$$

Using once again the Gronwall lemma we get

$$
|\phi^\alpha(t,\cdot)-\phi^\beta(t,\cdot)|_{1,m_1}\leqq C_1(\alpha+\beta),
\tag{4.16}
$$

and therefore $\phi\in B([0,T];C^1(H))$. By Lemma 2 it follows that for every $t\in[0,T]$

$$
\mathrm{Tr}\,(S\phi_{xx}^\alpha(t,x))\to\mathrm{Tr}\,(S\phi_{xx}^\alpha(t,x))
\tag{4.17}
$$

uniformly on every B_R. Recalling now that

$$
\phi_t^\alpha+\frac{1}{2}|\phi_x^\alpha|^2-\langle Ax,\phi_x^\alpha\rangle-\frac{\varepsilon}{2}\mathrm{Tr}\,(S\phi_{xx}^\alpha)=g-R_{\phi^\alpha,\alpha},
$$

and keeping in mind estimate (2.21) we may infer that for every $x\in D(A)\phi_t^\alpha(t,x)$ is bounded in $L^\infty(0,T)$ and as $\alpha\to0$

$$
\phi_t^\alpha(t,x)\to-\frac{1}{2}|\phi_x(t,x)|^2+\langle Ax,\phi_x(t,x)\rangle+\frac{\varepsilon}{2}\mathrm{Tr}\,(S\phi_{xx}(t,x))+g(t,x)
$$

uniformly on B_R, for any $t\in[0,T]$, where R is arbitrary. We have therefore proved that ϕ satisfies the conditions of Theorem 1. Let ϕ^i, $i=1,2$ be two solutions to problem (4.1) corresponding to (ϕ_0^i,g^i). We have

$$
\begin{aligned}
&\phi_t^i+\alpha^{-1}(\phi^i-\phi_\alpha^i)-\langle Ax,\phi_x^i\rangle-\frac{\varepsilon}{2}\mathrm{Tr}\,(S\phi_{xx}^i)=g^i+R_{\phi^i,\alpha},\\
&\phi^i(t,x)=\phi_0^i.
\end{aligned}
$$

This yields, recalling Proposition 5,

$$
\begin{aligned}
\phi^i(t,x)&=\exp\left(-\alpha^{-1}t\right)E\phi_0^i\left(e^{tA}x+\sqrt{\varepsilon}\int_0^t e^{sA}\,dW_{T-s}\right)\\
&\quad+E\int_0^t\exp\left(-\alpha^{-1}(t-s)\right)(\alpha^{-1}\phi_\alpha^i+g^i)\left(s,e^{(t-s)A}+\sqrt{\varepsilon}\int_0^{t-s}e^{\sigma A}\,dW_{T-\sigma}\right)ds
\end{aligned}
$$

and by (2.20), (2.23), it follows via Gronwall's lemma that

(4.18)
$$|\phi^1(t,\cdot)-\phi^2(t,\cdot)|_{0,n_0}\leqq C\Big(|\phi_0^1-\phi^2|_{0,n_0}+\int_0^t|g^1(s,\cdot)-g^2(s,\cdot)|_{0,n_0}\,ds\Big),\qquad 0\leqq t\in T$$

where C is independent of ε. In particular, we may conclude that the solution ϕ to (4.1) is unique and the proof of Theorem 1 is complete.

Remark. If Z happens to be a dense subset of X then for any $\phi_0\in X$ and $g\in B([0,T];X)$ problem (4.1) has a unique weak solution $\phi\in B([0,T];X)$.

Now we shall study the convergence of $\{\phi^\varepsilon\}$ for $\varepsilon\to 0$.

THEOREM 2. *Under the assumptions of Theorem* 1, $\phi^\varepsilon\to\phi$ *in* $B([0,T];X)\cap B([0,T];C^1(H))$, *where* $\phi(\cdot,x)\in W^{1,\infty}(0,T)$ *for all* $x\in D(A)$ *is the solution to the Hamilton–Jacobi equation*:

(4.19)
$$\begin{aligned}&\phi_t(t,x)+\tfrac{1}{2}|\phi_x(t,x)|^2-\langle Ax,\phi_x(t,x)\rangle=g(t,x)\quad\text{a.e. } t\in[0,T],\quad x\in D(A)\\&\phi(0,x)=\phi_0(x).\end{aligned}$$

Proof. First we observe that in estimates (4.5), (4.6) the constants can be taken independent of ε. For $\varepsilon,\lambda>0$ we have

$$\phi_t^\lambda+\frac{1}{2}|\phi_x^\lambda|^2=\langle Ax,\phi_x^\lambda\rangle-\frac{\varepsilon}{2}\operatorname{Tr}(S\phi_{xx}^\lambda)=g+\frac{\lambda-\varepsilon}{2}\operatorname{Tr}(S\phi_{xx}^\lambda).$$

Then by (4.5) and (4.18) we see that $\{\phi^\varepsilon\}$ is a Cauchy sequence in $B([0,T];X)$ and therefore for $\varepsilon\to 0$, $\phi^\varepsilon\to\phi$ in $B([0,T];X)$. Moreover, arguing as in the proof of inequality (4.16) we show that $\phi^\varepsilon\to\phi$ in $B([0,T];C^1(H))$. Again by estimates (4.5) and (4.6) it follows that $\{\phi^\varepsilon(t,\cdot)\}$ remain in a bounded subset of Z and for $\varepsilon\to 0$, $t\in[0,T]$ and $x\in D(A)$.

$$\phi_t^\varepsilon(t,x)\to-\tfrac{1}{2}|\phi_x(t,x)|^2+\langle Ax,\phi_x^\varepsilon(t,x)\rangle+g(t,x),$$

thereby completing the proof of Theorem 2.

We shall give now another approximation result which will be useful in the next section. Again we denote by $A_n=n^2(n-A)^{-1}-n$ the Yosida approximation of A.

PROPOSITION 7. *Assume that hypotheses of Theorem* 1 *hold for any* $n\in\mathbb{N}$ *let* ϕ^n *be the solution to the problem*

(4.20)
$$\begin{aligned}&\phi_t^n+\frac{1}{2}|\phi_x^n|^2+\langle Ax,\phi_x^n\rangle-\frac{\varepsilon}{2}\operatorname{Tr}(S\phi_{xx}^n)=g,\\&\phi^n(0,x)=\phi_0(x),\end{aligned}$$

and let ϕ *be the solution of problem* (4.1). *Then* $\phi^n\to\phi$ *in* $B([0,T];X)\cap B([0,T];C^1(H))$.

Proof. Let $\phi^{n,\alpha}$ be the solution to

(4.21)
$$\begin{aligned}&\phi_t^{n,\alpha}+\frac{1}{\alpha}(\phi^{n,\alpha}-\phi_\alpha^{n,\alpha})+\langle A_nx,\phi_x^{n,\alpha}\rangle-\frac{\varepsilon}{2}\operatorname{Tr}(S\phi_{xx}^{n,\alpha})=g,\\&\phi^{n,\alpha}(0,x)=\phi_0(x).\end{aligned}$$

Proceeding as in the proof of Theorem 1 we see that

(4.22) $\lim_{n\to\infty}\phi^{n,\alpha}=\phi^\alpha$ in $B([0,T];X)\cap B([0,T];C^1(H))$ for all $\alpha>0$.

Hence for $n \to \infty$

$$|\phi - \phi^n|_X \leqq |\phi - \phi^\alpha|_X + |\phi^\alpha - \phi^{n,\alpha}|_X + |\phi^{n,\alpha} - \phi^n|_X \to 0.$$

5. Synthesis of optimal control. Consider the Cauchy problem:

$$\begin{aligned} &\psi_t(t,x) - \tfrac{1}{2}|\psi_x(t,x)|^2 + \langle Ax, \psi_x(t,x)\rangle \\ &\quad + \frac{\varepsilon}{2}\operatorname{Tr}(S\psi_{xx}(t,x)) + V(t,x) = 0, \\ &\psi(T,x) = \phi_0(x), \end{aligned} \tag{5.1}$$

where $V(t,x) = g(T-t,x)$. Remark that (5.1) is obtained from (4.1) by setting $\psi(t,x) = \phi(T-t,x)$. Throughout this section we shall assume that assumptions (4.2) are satisfied. Then by Theorem 1, problem (5.1) has a unique solution $\psi \in B([0,T]; C^1(H))$ such that $\psi(\cdot,x) \in W^{1,\infty}(0,T)$ for every $x \in D(A)$. Notice also that by virtue of Proposition 7 the solution to the problem

$$\begin{aligned} &\psi_t^n(t,x) - \tfrac{1}{2}|\psi_x^n(t,x)|^2 + \langle A_n x, \psi_x^n(t,x)\rangle \\ &\quad + \frac{\varepsilon}{2}\operatorname{Tr}(S\psi_{xx}^n(t,x)) + V(t,x) = 0, \\ &\psi^n(T,x) = \phi_0(x) \end{aligned} \tag{5.2}$$

is convergent to ψ in the following sense:

$$\psi^n \to \psi \quad \text{in } B([0,T]; X), \tag{5.3}$$

$$\psi_x^n \to \psi_x \quad \text{uniformly in } [0,T]\times B_R. \tag{5.4}$$

We shall use these facts to prove the following lemma.

LEMMA 8. *Let $u \in M_w^2(0,T;H)$ and let ζ be the mild solution to the stochastic equation*

$$d\zeta = (A\zeta + u)\,ds + \sqrt{\varepsilon}\,dW_s, \qquad \zeta(t) = x, \quad t \leqq s \leqq T. \tag{5.5}$$

If ψ is the solution to (5.1) *then the equality*

$$\begin{aligned} &\psi(t,x) + \frac{1}{2}E\int_t^T |\psi_x(s,\zeta(s)) + u(s)|^2\,ds \\ &\quad = E\int_t^T \left(V(s,\zeta(s)) + \frac{1}{2}|u(s)|^2\right) ds + \phi_0(\zeta(T)) \end{aligned} \tag{5.6}$$

holds for all $(t,x) \in [0,T]\times H$.

Proof. Let ζ_n be the solution to

$$d\zeta_n = (A_n\zeta_n + u)\,ds + \sqrt{\varepsilon}\,dW_s, \qquad \zeta_n(t) = x, \tag{5.7}$$

and let $\psi_n : [0,T]\times H \to \mathbb{R}$ be the function defined by

$$\psi_n(t,x) = \int_0^T \psi^n(s,x)\rho_n(t-s)\,ds,$$

where $\{\rho_n\}$ is a family of C^∞-real valued functions such that

$$\operatorname{supp}(\rho_n) \subset \left]-\frac{1}{n}, \frac{1}{n}\right[, \qquad \rho_n(t) = \rho_n(-t), \quad \rho_n \geqq 0$$

and $\int_{-\infty}^{+\infty} \rho_n(t)\,dt = 1$. Clearly $\psi_n \in B([0, T]; Z)$ and $(\psi_n)_t \in B([0, T]; X)$. By (5.3) and (5.4) it follows that

$$\psi_n \to \psi \quad \text{in } B([0, T]; X). \tag{5.8}$$

By standard results on infinite dimensional stochastic equations (see for instance [7]) we know that $\zeta_n(t) \to \zeta(t)$ uniformly on $[0, T]$ with probability 1 and therefore

$$E((\psi_n)_x(t, y_n(t))) \to E\psi_x(t, y(t)) \quad \text{uniformly on } [0, T]. \tag{5.9}$$

Next by the Itô formula,

$$d\psi_n(s, \zeta_n) = (\psi_n)_s(s, \zeta_n)\,ds + \langle A_n\zeta_n\,ds + u\,ds + \sqrt{\varepsilon}\,dW_s, (\psi_n)_x(s, \zeta_n)\rangle$$

$$+\frac{\varepsilon}{2}\operatorname{Tr}(S\psi_{n,xx}(s, \zeta_n))\,ds.$$

Then, integrating on $[t, T]$ and taking the expectation gives

$$\psi_n(t, x) + \frac{1}{2}E\int_t^T |\psi_{n,x}(s, \zeta_n) + u(s)|^2\,ds$$

$$= E\int_t^T \left(V(s, \zeta_n) + \frac{1}{2}|u(s)|^2\right) ds + \phi_0(\zeta_n(T)).$$

Then, if we let n tend to $+\infty$, by (5.8) and (5.9), (5.6) follows as claimed.

The relevance of the solution ψ to (5.1) for the optimal control problem (P), is explained in Theorem 3 below.

THEOREM 3. *Assume that conditions* (4.2) *are satisfied. Then the solution* ψ *to* (5.1) *is the optimal value function of problem* (P), *i.e., for every* $t \in [0, T]$ *one has*

$$\psi(t, x) = \inf\left\{E\int_t^T \left(V(s, \zeta(s)) + \frac{1}{2}|u(s)|^2\right) ds + \phi_0(\zeta(T));\right.$$

$$\left. d\zeta = (A\zeta + u)\,ds + \sqrt{\varepsilon}\,dW_s,\ \zeta(t) = x,\ u \in M_w^2(0, T; H)\right\}. \tag{5.10}$$

Moreover, the solution ζ^+ *to the problem*

$$d\zeta = (A\zeta - \psi_x(t, \zeta))\,dt + \sqrt{\varepsilon}\,dW_t, \qquad t \in [0, T],$$

$$\zeta(0) = x \tag{5.11}$$

is an optimal trajectory to problem (P) *corresponding to the optimal control* u^+ *given by*

$$u^+(t) = -\psi_x(t, \zeta^+(t)) \quad \text{a.e. } t \in]0, T[. \tag{5.12}$$

The optimal control u^+ *is unique.*

In few words, Theorem 3 says that under assumption (4.2) $u(t) = -\psi_x(t, \zeta(t))$ is an *optimal feedback control* for the stochastic control problem (P) (see [8] for definitions and classical results on these topics).

Proof of Theorem 3. By Lemma 8 (formula (5.6)) we see that for each $(t, x) \in [0, T] \times H$, $\psi(t, x) = \Phi(t, x)$ where Φ is the optimal value functions of problem (P).

Now let (ζ^+, u^+) be a pair given by (5.10). Since $\psi_x \in B([0, T]; C(H))$ and it is monotone in x (as the derivative of a convex function) (5.9) has a unique solution ζ^+ (see [7, Thms. 4 and 7]; remark that hypothesis (24) in Theorem 7 is satisfied in

our situation because ψ_x is monotone). By (5.6) we see that for every $t \in [0, T]$

$$\psi(t, x) = E\left\{\int_t^T \left(g(s, \zeta^+(s)) + \frac{1}{2}|u^+(s)|^2\right) ds\right\} + \phi_0(\zeta(T)),$$

and therefore u^+ is optimal in problem (P).

Assume now that $(\tilde{u}, \tilde{y})$ is another optimal pair. Again by formula (5.6) it follows that

$$E\int_t^T |\psi_x(s, \tilde{y}(s)) + \tfrac{1}{2}\tilde{u}(s)|^2 \, ds = 0$$

which implies $\tilde{u} = -\psi_x(s, \tilde{\zeta}(s))$. since the solution to (5.9) is unique we infer that $\tilde{\zeta} = \zeta^+$ and $\tilde{u} = u^+$ as claimed.

REFERENCES

[1] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Sijthoff and Noordhoff, Groningen, 1978.

[2] V. BARBU AND G. DA PRATO, *Global existence for the Hamilton–Jacobi equations in Hilbert spaces*, Ann. Scuola Normale Superiore Pisa, VIII, 2 (1981), pp. 257–284.

[3] ———, *A direct method for studying the dynamic programming equation for controlled diffusion, processes in Hilbert spaces*, Numer. Funct. Anal. and Optimiz., 4(1) (1981), pp. 23–43.

[4] A. BENSOUSSAN AND J. L. LIONS, *Temps d'arrêt et contrôle impulsionnel*, Dunod, Paris, 1978.

[5] H. BRÉZIS, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[6] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1978.

[7] G. DA PRATO, M. IANNELLI AND L. TUBARO, *Semilinear stochastic differential equations*, Boll. UMI (1978), pp. 168–185.

[8] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[9] W. H. FLEMING, *The Cauchy problem for a nonlinear first order partial differential equation*, J. Differential Equations 5 (1969), pp. 515–530.

[10] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Academic Press, New York, 1975.

[11] M. METIVIER AND T. PELLAUMAIL, *Stochastic Integrals*, Academic Press, New York, 1980.

[12] E. PARDOUX, *Intégrales stochastiques Hilbertiennes*, Cahiers de Mathématiques de la Decision, Paris Dauphine, 1976.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

0363-0129/83/2104-0003 $01.25/0

APPROXIMATE CONTROLLABILITY FOR A CLASS OF SEMILINEAR ABSTRACT EQUATIONS*

HONG XING ZHOU†

Abstract. In this paper a class of control systems governed by the semilinear abstract equation: $\dot{y}+Ay=F(y)+Bv$ is considered. A sufficient condition for the approximate controllability is obtained which improves J. Henry's results on a nonlinear parabolic control systems with some more serious restrictions. It is suitable not only to the infinite-dimensional case but also to the finite-dimensional case. Two examples are given to explain the applications of the theory.

Key words. semilinear abstract system, approximate controllability theory, infinite dimension examples

1. Introduction. In this paper we will be concerned with a class of control systems governed by the semilinear abstract equation with a distributed control

$$\begin{aligned} &\dot{y}(t)+Ay(t)=F(y(t))+Bv(\cdot)(t), \qquad 0<t<T, \\ &y(0)=\eta_0. \end{aligned} \tag{1.1}$$

Here the state $y(t)$, $0\leqq t\leqq T$, takes values in the real Hilbert space X and the control $v(\cdot)$ is in another real Hilbert space V. For instance, $V=L^2(0, T; U)$ and U is a real Hilbert space. Assume the operator $-A$ generates a differentiable semigroup $S(\cdot)$ on the state space X. In (1.1) the action operator B is a linear bounded operator mapping V into $L^2(0, T; X)$. $F(\cdot)$ is some nonlinear function satisfying Hypothesis (F) in § 2.

If $F(y(\cdot))\equiv 0$, $V=L^2(0, T; U)$ and $B\in\mathscr{L}[U\to X]$, the space consisting of all linear bounded operators mapping U into X, i.e., $Bv(\cdot)(t)=Bu(t)$, then the system (1.1) becomes

$$\begin{aligned} &\dot{y}(t)+Ay(t)=Bu(t), \qquad 0<t<T, \\ &y(0)=\eta_0, \end{aligned} \tag{1.2}$$

which is called the corresponding linear system of (1.1). The controllability theory on the linear abstract control system (1.2) is well known. One of the principal results on approximate controllability is that the linear control system (1.2) is approximately controllable on $[0, T]$ if and only if $[S(t)B]^*\phi^*=0$ for $0\leqq t\leqq T$ implies $\phi^*=0$ in X^* (see [3], [1]). In [9] the existing results on controllability theory for linear partial differential equations are summarized.

As for the control systems governed by nonlinear abstract equation or nonlinear partial differential equation, there are very few papers to discuss the approximate or exact controllability problems. Using the implicit function theorem, H. O. Fattorini [4] studied the local controllability of a nonlinear wave equation with an input of the form $b(x)f(t)$. In [2] the controllability problem is considered for the N-dimensional hyperbolic equation.

In 1978, J. Henry [6] discussed approximate controllability for a nonlinear parabolic equation where the operator A is positive and $-A$ generates a holomorphic compact semigroup. He pointed out that if the range BV of the operator B in (1.1) is dense in $L^2(0, T; X)$ then under some hypotheses on the nonlinear function $F(\cdot)$ the nonlinear parabolic system (1.1) is approximately controllable. As an infinite-

* Received by the editors June 18, 1981, and in revised form April 16, 1982.

† Department of Mathematics, Shandong University, Jinan, Shandong Province, The People's Republic of China, and University of California, Los Angeles, California 90024.

dimensional parabolic equation, it may be looked at as a natural extension of a finite-dimensional ordinary differential equation. In the later case, A is a $N\times N$ matrix and B is a $N\times M$ matrix ($M\leqq N$), and approximate controllability becomes complete controllability for the linear control system. Thus Henry's hypothesis $\overline{BV}=L^2(0,T;X)$ in the finite-dimensional case is equivalent to that $M=N$ and B is a nonsingular $N\times N$ matrix. That means his result cannot be applied to such a simple second order ordinary differential equation as

$$\frac{d}{dt}\begin{pmatrix}y_1\\y_2\end{pmatrix}=\begin{pmatrix}-1 & 0\\ 0 & -4\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}+\begin{pmatrix}F_1(y_1,y_2)\\F_2(y_1,y_2)\end{pmatrix}+\begin{pmatrix}2\\1\end{pmatrix}u(t). \tag{1.3}$$

But in [7], [11] it is proved that such a system is completely controllable under some assumptions on the nonlinear functions F_1 and F_2 and the terminal time T.

In this paper, sufficient conditions—Hypothesis (B) in § 3—for the approximate controllability of the semilinear abstract system (1.1) are obtained. If the range BV of the operator B is dense in $L^2(0,T;X)$ then Hypothesis (B) is satisfied (Theorem 3.3). So this sufficient condition is more general than previous ones. It is suitable not only for a nonlinear abstract control system in Hilbert space, but also for the finite-dimensional ordinary differential equations, e.g. the nonlinear system (1.3). In § 4 two examples will be given which show that even if the range BV of the operator B is not dense in $L^2(0,T;X)$ then under some reasonable hypotheses on the nonlinear function $F(\cdot)$ and the terminal time T the semilinear parabolic system is still approximately controllable.

2. Preliminaries. Here we give some notation and introduce some lemmas concerning the properties of the solution of (1.1) corresponding to a given control $v(\cdot)$.

First, a hypothesis for the nonlinear function F is given which insures existence and uniqueness for the nonlinear equation (1.1) with a given $v(\cdot)\in V$.

Hypothesis (F). $F(\cdot)$ is a nonlinear operator mapping X into X and $F(\cdot)$ satisfies a Lipschitz condition with some positive constant K_1

$$\|F(y_1)-F(y_2)\|\leqq K_1\|y_1-y_2\| \quad \text{for } y_1, y_2\in X. \tag{2.1}$$

(Here $\|\cdot\|$ denotes the norm in X, i.e., $\|\cdot\|_X$.)

Under Hypothesis (F), $F(y(t))\in X$, $0\leqq t\leqq T$, and $F(y(\cdot))\in L^2(0,T;X)$ for any $y(\cdot)\in L^2(0,T;X)$, since

$$\|F(y(t))\|^2\leqq 2K_1^2\|y(t)\|^2+2\|F(0)\|^2, \qquad 0\leqq t\leqq T. \tag{2.2}$$

In § 1 it was pointed out that the important case for V and B is $V=L^2(0,T;U)$ and $B\in\mathscr{L}[U\to X]$. To distinguish the two kinds of case we use $B_{(0,T)}$ to denote the operator on V.

While discussing the semilinear parabolic system (1.1) on the interval $[0,T]$ we usually use its "intercept system" on the interval $[t_0,T]$ with some given initial value $\xi_0\in X$ at the initial time $t_0\in[0,T)$:

$$\begin{aligned}&\dot{y}(t)+Ay(t)=F(y(t))+B_{(t_0,T)}v(\cdot)(t), \qquad t_0<t<T,\\ &y(t_0)=\xi_0,\end{aligned} \tag{2.3}$$

where the track $y(\cdot)$ is in $L^2(t_0,T;X)$, the control $v(\cdot)$ is in V and $B_{(t_0,T)}\in\mathscr{L}[V\to L^2(t_0,T;X)]$ is the intercept of $B_{(0,T)}$ on $[t_0,T]$, i.e.,

$$B_{(t_0,T)}v(\cdot)(t)=B_{(0,T)}v(\cdot)(t) \quad \text{for } t_0\leqq t\leqq T.$$

It is proved that there exists a unique solution $y(\cdot)\in L^2(t_0, T; X)$ for the nonlinear Cauchy problem (2.3) with every given $v(\cdot)\in V$ under Hypothesis (F). Thus one may define a solution mapping, denoted by $Y(t_0, \xi_0; v)$, from $R\times X\times V$ into $L^2(t_0, T; X)$ (see [10], [5]). It is not difficult to obtain the following estimate for the solution mapping:

LEMMA 2.1. *Let $v(\cdot)\in V$ and $\xi_0\in X$. Then under Hypothesis (F) the solution mapping $Y(t_0, \xi_0; v)$ of* (2.3) *satisfies*

$$\begin{aligned}\|Y(t_0, \xi_0; v)(\cdot)\|_{L^2(t_0,T;X)} &\leqq M_1\|\xi_0\|\sqrt{T-t_0}+M_2\|F(0)\|(T-t_0)^{3/2}\\ &\quad +M_3(T-t_0)\|B_{(t_0,T)}v(\cdot)(\cdot)\|_{L^2(t_0,T;X)},\end{aligned}\tag{2.4}$$

where M_1, M_2 and M_3 are positive constants independent on t_0, ξ_0 and v. Let $v_1(\cdot)$ and $v_2(\cdot)$ be in V. Then

$$\|y_1(\cdot)-y_2(\cdot)\|_{L^2(t_0,T;X)}\leqq M_3(T-t_0)\|B_{(t_0,T)}v_1(\cdot)(\cdot)-B_{(t_0,T)}v_2(\cdot)(\cdot)\|_{L^2(t_0,T;X)}\tag{2.4$'$}$$

where $y_n(\cdot)=Y(t_0, \xi_0; v_n)(\cdot)$, $n=1, 2$.

Proof. By the semigroup method [1], [5], the solution $y(t)=Y(t_0, \xi_0; v)(t)$ of (2.3) satisfies

$$y(t)=S(t-t_0)\xi_0+\int_{t_0}^{t}S(t-s)[F(y(s))+B_{(t_0,T)}v(\cdot)(s)]\,ds,\qquad 0\leqq t\leqq T.\tag{2.5}$$

Thus

$$\begin{aligned}\|y(t)\|\leqq\|S(t-t_0)\|\|\xi_0\|&+\int_{t_0}^{t}\|S(t-s)\|\|F(y(s))-F(0)\|\,ds\\ &+\int_{t_0}^{t}\|S(t-s)\|\|F(0)+B_{(t_0,T)}v(\cdot)(s)\|\,ds.\end{aligned}$$

Denoting

$$M_A=\max_{0\leqq t\leqq T}\|S(t)\|_{\mathcal{L}[X\to X]},\tag{2.6}$$

we have

$$\begin{aligned}\|y(t)\|\leqq M_A[\|\xi_0\|+\|F(0)\|(t-t_0)+\sqrt{t-t_0}\|B_{(t_0,T)}v(\cdot)(\cdot)\|_{L^2(t_0,T;X)}]\\ +K_1M_A\int_{t_0}^{t}\|y(s)\|\,ds.\end{aligned}$$

Since $[\cdots]$ in the above inequality is monotonically increasing, using Gronwall's inequality, we have

$$\|y(t)\|\leqq M_A\,e^{K_1M_AT}[\|\xi_0\|+\|F(0)\|(t-t_0)+\sqrt{t-t_0}\|B_{(t_0,T)}v(\cdot)(\cdot)\|_{L^2(t_0,T;X)}]\tag{2.7}$$

and

$$\begin{aligned}\|Y(t_0, \xi_0; v)\|_{L^2(t_0,T;X)}&=\|y(\cdot)\|_{L^2(t_0,T;X)}\\ &\leqq\sqrt{2}M_A\,e^{K_1M_AT}\Big[\|\xi_0\|\sqrt{T-t_0}+\frac{1}{\sqrt{3}}\|F(0)\|(T-t_0)^{3/2}\\ &\quad+\frac{1}{\sqrt{2}}(T-t_0)\|B_{(t_0,T)}v(\cdot)(\cdot)\|_{L^2(t_0,T;X)}\Big].\end{aligned}\tag{2.8}$$

Formula (2.4) is required, with

$$M_1 = \sqrt{2} M_A e^{K_1 M_A T}, \quad M_2 = \frac{1}{\sqrt{3}} M_1, \quad M_3 = \frac{1}{\sqrt{2}} M_1, \tag{2.9}$$

and (2.4)′ is proved by the same way. □

Remark 2.2. If we consider concrete nonlinear control systems, then the constants M_1, M_2 and M_3 in (2.4) could be improved over that in (2.9). For example, assume the operator A is positive, i.e.,

$$(Ay, y) \geqq 0 \quad \text{for } y \in X, \tag{2.10}$$

and the nonlinear function F is negative, i.e.,

$$(F(y_1) - F(y_2), y_1 - y_2) \leqq 0 \quad \text{for } y_1, y_2 \in X. \tag{2.11}$$

Then the constants M_1, M_2 and M_3 in (2.4) take values such as

$$M_1 = 1, \quad M_2 = \frac{2}{\sqrt{3}}, \quad M_3 = \sqrt{2}. \tag{2.12}$$

That proof of (2.12) is slightly different from that of Lemma 2.1 and is omitted because the values of the constants M_1, M_2 and M_3 are not essential for the future.

Given a strongly continuous semigroup $S(t)$ for $t \geqq 0$ we define a linear bounded operator $\mathcal{S}_{t_0}$ mapping $L^2(t_0, T; X)$ into X by

$$\mathcal{S}_{t_0} p = \int_{t_0}^{T} S(T-t) p(t)\, dt \quad \text{for } p(\cdot) \in L^2(t_0, T; X). \tag{2.13}$$

Let $v(\cdot)$ be an arbitrary element in V and $y(t) = Y(t_0, \xi_0; v)(t)$. Then the pair $(y(\cdot), v(\cdot))$ satisfies (2.5) and the terminal state $y(T)$ may be rewritten as

$$y(T) = Y(t_0, \xi_0; v)(T) = S(T-t_0)\xi_0 + \mathcal{S}_{t_0} F(y(\cdot)) + \mathcal{S}_{t_0} B_{(t_0,T)} v(\cdot)(\cdot). \tag{2.14}$$

So the reachable set for the intercept system (2.3) at the terminal time T is dependent on t_0 and ξ_0. We denote it by

$$K(t_0, \xi_0) = \{\xi_T | \xi_T = Y(t_0, \xi_0; v)(T) \text{ for some } v \in V\}. \tag{2.15}$$

If the reachable set $K(t_0, \xi_0)$ is dense in X for any given $\xi_0 \in X$ then we say that the intercept system (2.3) is approximately controllable on $[t_0, T]$. Here, for convenience, we give an equivalent definition of approximate controllability of the intercept system (2.3) on $[t_0, T]$:

DEFINITION. Assume ξ_0 is arbitrarily given in X. The intercept system (2.3) is called *approximately controllable* on $[t_0, T]$ if for any given $\varepsilon > 0$ and $\xi_T \in X$ there exists some control $v_\varepsilon(\cdot) \in V$ such that

$$\| \xi_T - S(T-t_0)\xi_0 - \mathcal{S}_{t_0} F(y_\varepsilon) - \mathcal{S}_{t_0} B_{(t_0,T)} v_\varepsilon \| < \varepsilon, \tag{2.16}$$

where $y_\varepsilon(t) = Y(t_0, \xi_0; v_\varepsilon)(t)$, $t_0 \leqq t \leqq T$.

Remark. The definition of approximate controllability for the intercept system (2.3) is slightly different from the ordinary one in which $\xi_0 = 0$. But there is not any essential distinction.

Obviously, the linear intercept system

$$\dot{y}(t) + Ay(t) = u(t), \qquad t_0 < t < T,$$
$$y(t_0) = \xi_0$$

is approximately controllable and the reachable set at the terminal T includes the domain $D(A)$ of the operator A.

3. Approximate controllability. Before discussing approximate controllability for the semilinear abstract system (1.1) on $[0, T]$ we deal first with the approximate controllability problem for the corresponding intercept system (2.3) on $[t_0, T]$. We prove here that some intercept system is approximately controllable under some assumptions on the nonlinear function F, the action operator $B_{(t_0,T)}$ and the length $T-t_0$ of the time interval.

Hypothesis (B). For every arbitrarily given $\varepsilon>0$ and $p(\cdot)\in L^2(t_0, T; X)$ there exists some $v(\cdot)\in V$ such that

$$\|\mathcal{S}_{t_0}p-\mathcal{S}_{t_0}B_{(t_0,T)}v\|<\varepsilon; \tag{3.1}$$

$$\|B_{(t_0,T)}v(\cdot)(\cdot)\|_{L^2(t_0,T;X)}\leqq q_1\|p(\cdot)\|_{L^2(t_0,T;X)}; \tag{3.2}$$

where q_1 is a positive constant independent of $p(\cdot)$;

$$\text{The constant } q_1 \text{ satisfies } M_3(T-t_0)K_1q_1<1. \tag{3.3}$$

Remark. Hypothesis (3.1) is equivalent to the approximate controllability of the corresponding linear system.

THEOREM 3.1. *Let the operator $B_{(t_0,T)}$, the nonlinear function F and the length $(T-t_0)$ of the time interval $[t_0, T]$ for the intercept system* (2.3) *on $[t_0, T]$ satisfy Hypothesis* (B). *Then this intercept system* (2.3) *is approximately controllable on $[t_0, T]$.*

Proof. Since the domain $D(A)$ of the operator A is dense it is sufficient to prove

$$D(A)\subset\overline{K(t_0,\xi_0)}; \tag{3.4}$$

i.e., for any given $\varepsilon>0$ and $\xi_T\in D(A)$ there exists an $v_\varepsilon(\cdot)\in V$ such that

$$\|\xi_T-S(T-t_0)\xi_0-\mathcal{S}_{t_0}F(y_\varepsilon)-\mathcal{S}_{t_0}B_{(t_0,T)}v_\varepsilon\|<\varepsilon, \tag{3.5}$$

where $y_\varepsilon(\cdot)=Y(t_0,\xi_0;v_\varepsilon)(\cdot)$ satisfies

$$\begin{aligned} y_\varepsilon(t)&=S(t-t_0)\xi_0+\int_{t_0}^{t}S(t-s)F(y_\varepsilon(s))\,ds\\ &\quad+\int_{t_0}^{t}S(t-s)B_{(t_0,T)}v_\varepsilon(\cdot)(s)\,ds, \qquad t_0\leqq t\leqq T. \end{aligned} \tag{3.6}$$

As $\xi_T\in D(A)$ and $S(T-t_0)\xi_0\in D(A)$ there exists some $p(\cdot)\in C^1([t_0, T]; X)$ such that

$$\mathcal{S}_{t_0}p=\xi_T-S(T-t_0)\xi_0,$$

e.g., $p(t)=1/(T-t_0)[\xi_1+(t-t_0)A\xi_1]$, where $\xi_1=\xi_T-S(T-t_0)\xi_0$.

We construct a sequence recursively as follows:

Assume $v_1(\cdot)\in V$ is arbitrarily given. By Hypothesis (3.1) there exists some $v_2(\cdot)\in V$ such that

$$\|\xi_T-S(T-t_0)\xi_0-\mathcal{S}_{t_0}F(y_1)-\mathcal{S}_{t_0}B_{(t_0,T)}v_2\|<\frac{\varepsilon}{2^2},$$

where

$$y_1(t)=Y(t_0,\xi_0;v_1)(t), \qquad t_0\leqq t\leqq T.$$

For $v_2(\cdot)$ thus obtained, we determine $w_2(\cdot)\in V$ by Hypotheses (3.1) and (3.2) such that

$$\|\mathscr{S}_{t_0}[F(y_2)-F(y_1)]-\mathscr{S}_{t_0}B_{(t_0,T)}w_2\|<\frac{\varepsilon}{2^3}$$

and

$$\begin{aligned}\|B_{(t_0,T)}w_2(\cdot)(\cdot)\|_{L^2(t_0,T;X)}&\leqq q_1\|F(y_2)(\cdot)-F(y_1)(\cdot)\|_{L^2(t_0,T;X)}\\ &\leqq q_1K_1\|y_2(\cdot)-y_1(\cdot)\|_{L^2(t_0,T;X)}\\ &\leqq q_1K_1M_3(T-t_0)\|B_{(t_0,T)}v_2(\cdot)(\cdot)-B_{(t_0,T)}v_1(\cdot)(\cdot)\|_{L^2(t_0,T;X)},\end{aligned}\tag{3.7}$$

where

$$y_n(t)=Y(t_0,\xi_0;v_n)(t),\qquad n=1,2,\quad t_0\leqq t\leqq T.$$

Thus we may define

$$v_3(\cdot)=v_2(\cdot)-w_2(\cdot)\quad\text{in } V,$$

which has the following property:

$$\begin{aligned}&\|\xi_T-S(T-t_0)\xi_0-\mathscr{S}_{t_0}F(y_2)-\mathscr{S}_{t_0}B_{(t_0,T)}v_3\|\\ &\quad=\|\xi_T-S(T-t_0)\xi_0-\mathscr{S}_{t_0}F(y_1)\\ &\qquad-\mathscr{S}_{t_0}B_{(t_0,T)}v_2+\mathscr{S}_{t_0}B_{(t_0,T)}w_2-\mathscr{S}_{t_0}[F(y_2)-F(y_1)]\|\\ &\quad<\left(\frac{1}{2^2}+\frac{1}{2^3}\right)\varepsilon.\end{aligned}\tag{3.8}$$

By induction, it is proved that there exists a sequence of $v_n(\cdot)$ in V such that

$$\|\xi_T-S(T-t_0)\xi_0-\mathscr{S}_{t_0}F(y_n)-\mathscr{S}_{t_0}B_{(t_0,T)}v_{n+1}\|<\left(\frac{1}{2^2}+\cdots+\frac{1}{2^{n+1}}\right)\varepsilon,\tag{3.9}$$

where

$$y_n(t)=Y(t_0,\xi_0;v_n)(t),\qquad t_0\leqq t\leqq t,\quad n=1,2,\cdots\tag{3.10}$$

and

$$\begin{aligned}&\|B_{(t_0,T)}v_{n+1}(\cdot)(\cdot)-B_{(t_0,T)}v_n(\cdot)(\cdot)\|_{L^2(t_0,T;X)}\\ &\qquad\leqq q_1K_1M_3(T-t_0)\|B_{(t_0,T)}v_n(\cdot)(\cdot)-B_{(t_0,T)}v_{n-1}(\cdot)(\cdot)\|_{L^2(t_0,T;X)}.\end{aligned}\tag{3.11}$$

By Hypothesis (3.3) the sequence $\{B_{(t_0,T)}v_n;\ n=1,2,\cdots\}$ is a Cauchy sequence in the Banach space $L^2(t_0,T;X)$ and there exists some $f(\cdot)$ in $L^2(t_0,T;X)$ such that

$$\lim_{n\to\infty}B_{(t_0,T)}v_n(\cdot)(\cdot)=f(\cdot)\quad\text{in } L^2(t_0,T;X).$$

Therefore, for any given $\varepsilon>0$ there exists some integer N_ε such that

$$\|\mathscr{S}_{t_0}B_{(t_0,T)}v_{N_\varepsilon+1}-\mathscr{S}_{t_0}B_{(t_0,T)}v_{N_\varepsilon}\|<\frac{\varepsilon}{2}\tag{3.12}$$

and

$$
\begin{aligned}
&\|\xi_T - S(T-t_0)\xi_0 - \mathscr{S}_{t_0}F(y_{N_\varepsilon}) - \mathscr{S}_{t_0}B_{(t_0,T)}v_{N_\varepsilon}\| \\
&\quad \leqq \|\xi_T - S(T-t_0)\xi_0 - \mathscr{S}_{t_0}F(y_{N_\varepsilon}) - \mathscr{S}_{t_0}B_{(t_0,T)}v_{N_\varepsilon+1}\| \\
&\qquad + \|\mathscr{S}_{t_0}B_{(t_0,T)}v_{N_\varepsilon+1} - \mathscr{S}_{t_0}B_{(t_0,T)}v_{N_\varepsilon}\| \\
&\quad < \left(\frac{1}{2^2} + \cdots + \frac{1}{2^{N_\varepsilon+1}}\right)\varepsilon + \frac{\varepsilon}{2} \leqq \varepsilon,
\end{aligned} \tag{3.13}
$$

in which $y_{N_\varepsilon}(\cdot) = Y(t_0, \xi_0; v_{N_\varepsilon})(\cdot)$. These are the required inequality (3.5) and (3.6). Thus the intercept nonlinear system (2.3) is approximately controllable under Hypotheses (B) and (F). □

Remark. The limiting element f on the sequence $\{B_{(t_0,T)}v_n\}$ is only used to obtain the inequality (3.12). In generality, the control sequence $\{v_n\}$ is not convergent and there is not any "limiting control" $\bar{v} \in V$ such that $f = B_{(t_0,T)}\bar{v}$. □

From Theorem 3.1 the following corollary on the approximate controllability for the nonlinear system (1.1) on $[0, T]$ is obtained immediately.

COROLLARY 3.2. *Consider the semilinear parabolic system* (1.1) *on* $[0, T]$. *If there exists some* $t_0 \in [0, T)$ *such that Hypothesis* (B) *is satisfied for its intercept system* (2.3) *on* $[t_0, T]$, *then under Hypothesis* (F) *the original system* (1.1) *is approximately controllable for every given* $\eta_0 \in X$ *on the time interval* $[0, T]$. □

The conclusions in Theorem 3.1 and Corollary 3.2 are suitable for more general cases. For some concrete nonlinear system Hypothesis (B) may be verified. We discuss two cases here:

(1) the range $B_{(0,T)}V$ of the operator B is dense in $L^2(0, T; X)$,

(2) the finite-dimensional case.

THEOREM 3.3. *Suppose the range* $B_{(0,T)}V$ *of the operator* $B_{(0,T)}$ *is dense in* $L^2(0, T; X)$. *Then under Hypothesis* (F) *the nonlinear parabolic system* (1.1) *is approximately controllable for every given* $\eta_0 \in X$ *on* $[0, T]$.

Proof. Denote the intercept of $B_{(0,T)}$ on $[t_0, T]$ by $B_{(t_0,T)}$ for every given $t_0 \in [0, T)$. Then it is easy to see that the range $B_{(t_0,T)}V$ of the operator $B_{(t_0,T)}$ is dense in $L^2(t_0, T; X)$. In fact, for any given $p(\cdot) \in L^2(t_0, T; X)$ there exists a sequence of $v_n(\cdot) \in V$ such that

$$
\lim_{n\to\infty} \|B_{(0,T)}v_n(\cdot)(\cdot) - p^*(\cdot)\|_{L^2(0,T;X)} = 0,
$$

where $p^*(\cdot) \in L^2(0, T; X)$ is the extension of $p(\cdot)$ with $p^*(t) = 0$ for $0 \leqq t < t_0$. As

$$
\|B_{(t_0,T)}v_n(\cdot)(\cdot) - p(\cdot)\|_{L^2(t_0,T;X)} \leqq \|B_{(0,T)}v_n(\cdot)(\cdot) - p^*(\cdot)\|_{L^2(0,T;X)}.
$$

That means the range $B_{(t_0,T)}V$ is dense in $L^2(t_0, T; X)$ for every given $t_0 \in [0, T)$.

Now choose an arbitrary t_0 such that

$$
0 < T - t_0 < \frac{1}{M_3K_1}, \tag{3.14}
$$

for example $t_0 = \max\{0, T - (\frac{1}{2}M_3K_1)\}$. Thus for any given $\varepsilon > 0$ and $p(\cdot) \in L^2(t_0, T; X)$ there exists some $v(\cdot) \in V$ such that if

$$
\|B_{(t_0,T)}v(\cdot)(\cdot) - p(\cdot)\|_{L^2(t_0,T;X)} < \delta\|p(\cdot)\|_{L^2(t_0,T;X)},
$$

$$
\|\mathscr{S}_{t_0}p - \mathscr{S}_{t_0}B_{(t_0,T)}v\| < \varepsilon,
$$

where $\delta > 0$ is any given constant. Thus

$$\|B_{(t_0,T)}v(\cdot)(\cdot)\|_{L^2(t_0,T;X)} \leqq \|p(\cdot)\|_{L^2(t_0,T;X)}(\delta+1). \tag{3.15}$$

By the definition (3.14) of t_0 Hypothesis (B) is satisfied for this t_0 and the operator $B_{(t_0,T)}$ with dense range. Hence the approximate controllability of the system (1.1) is obtained on $[0, T]$ where $T > 0$ is any given terminal time. □

In the rest of this section the finite-dimensional case will be considered. Suppose $X = \mathbb{R}^N$, $U = \mathbb{R}^M (M < N)$. The time-dependent nonlinear lumped-parameter control system is described as follows:

$$\begin{aligned} &\frac{dx(t)}{dt} + A(t)x(t) = F(x(t)) + B(t)u(t), \qquad 0 \leqq t \leqq T, \\ &x(0) = \eta, \end{aligned} \tag{3.16}$$

where $A(t)$ and $B(t)$ are $N \times N$ and $N \times M$ continuous matrices on $[0, T]$ respectively. The corresponding linear system of (3.16) is

$$\frac{dx(t)}{dt} + A(t)x(t) = B(t)u(t). \tag{3.17}$$

Denote the fundamental solution matrix of (3.17) by $S(t, s)$, $0 \leqq s \leqq t \leqq T$. Then we have a preliminary lemma on controllability (see [8]):

LEMMA 3.4. *Let the linear system* (3.17) *be completely controllable on* $[t_0, T]$ *and the matrix* $G(t_0, T)$ *be defined by*

$$G(t_0, T) = \int_{t_0}^{T} S(T, t)B(t)B(t)^*S(T, t)^* \, dt. \tag{3.18}$$

Then for every $h \in \mathbb{R}^N$ the control $u(t) = B(t)^*S(T, t)^*[G(t_0, T)]^{-1}h$ has the min-norm property:

$$\|u(\cdot)\|_{L^2(t_0,T;\mathbb{R}^M)} = \inf\left\{\|v(\cdot)\|_{L^2(t_0,T;\mathbb{R}^M)}: \int_{t_0}^{T} S(T, t)B(t)v(t) \, dt = h\right\}. \tag{3.19}$$

THEOREM 3.5. *Let* $t_0 \in [0, T)$. *Assume the linear system* (3.17) *is completely controllable on* $[t_0, T]$, *nonlinear function* F *satisfies Hypothesis* (F) *in* § 2, *and the condition*

$$M_3K_1M_A^2\|B(\cdot)\|^2(T - t_0)^{3/2}\|G(t_0, T)^{-1}\| < 1 \tag{3.20}$$

is satisfied for the nonlinear system (3.16), *where*

$$M_A = \max_{0 \leqq s \leqq t \leqq T} \|S(t, s)\|_{\mathscr{L}[\mathbb{R}^N \to \mathbb{R}^N]}.$$

Then the nonlinear system (3.16) *is completely controllable on* $[0, T]$ *for any given initial state* $\eta_0 \in X$.

Proof. Let $f(\cdot) \in L^2(t_0, T; \mathbb{R}^N)$ be an arbitrarily given function. Then by the complete controllability of the linear system (3.17) on the time interval $[t_0, T]$ there exists an $u(\cdot) \in L^2(t_0, T; \mathbb{R}^M)$ such that

$$\int_{t_0}^{T} S(T, t)B(t)u(t) \, dt = \int_{t_0}^{T} S(T, t)f(t) \, dt$$

and

$$u(t) = B(t)^*S(T, t)^*G(t_0, T)^{-1} \int_{t_0}^{T} S(T, t)f(t) \, dt.$$

Thus

$$\|u(\cdot)\|_{L^2(t_0,T;\mathbb{R}^M)} \leqq q_1^* \|f(\cdot)\|_{L^2(t_0,T;\mathbb{R}^N)}, \tag{3.21}$$

where

$$q_1^* = M_A^2 \|B(\cdot)\|(T-t_0)^{1/2}\|G(t_0, T)^{-1}\|.$$

Similar to Theorem 3.1, for any given ξ_T and ξ_0 in $\mathbb{R}^N$ there exists a sequence $\{v_n(\cdot); n = 1, 2, \cdots\}$ in $L^2(t_0, T; \mathbb{R}^M)$ such that

$$\begin{aligned} 0 = \xi_T - S(T, s)\xi_0 - \int_{t_0}^{T} S(T, s)F(x_n(s))\, ds \\ - \int_{t_0}^{T} S(T, s)B(s)v_{n+1}(s)\, ds, \qquad n = 1, 2, \cdots \end{aligned} \tag{3.22}$$

and

$$\begin{aligned} &\|v_{n+1}(\cdot) - v_n(\cdot)\|_{L^2(t_0,T;\mathbb{R}^M)} \\ &\quad \leqq q_1^* K_1 M_3 (T-t_0)\|B(\cdot)\|\|v_n(\cdot) - v_{n-1}(\cdot)\|_{L^2(t_0,T;\mathbb{R}^M)}, \qquad n = 2, 3, \cdots \end{aligned} \tag{3.23}$$

where $x_n(t) = X(t_0, \xi_0; v_n)(t)$ is the solution of the intercept system on $[t_0, T]$ of (3.16), i.e.,

$$x_n(t) = S(t, t_0)\xi_0 + \int_{t_0}^{t} S(t, s)F(x_n(s))\, ds + \int_{t_0}^{t} S(t, s)B(s)v_n(s)\, ds. \tag{3.24}$$

Here we notice that since the linear system (3.17) is finite-dimensional the inequality (3.9) in Theorem 3.1 becomes an equality (3.22) here. By Hypothesis (3.20) the sequence $\{v_n(\cdot); n = 1, 2, \cdots\}$ is a Cauchy sequence in $L^2(t_0, T; \mathbb{R}^M)$ and there exists some control $u(\cdot)$ in $L^2(t_0, T; \mathbb{R}^M)$ such that

$$\lim_{n\to\infty} v_n(\cdot) = u(\cdot) \quad \text{in } L^2(t_0, T; \mathbb{R}^M). \tag{3.25}$$

Since the semilinear system (3.16) is a finite-dimensional ordinary differential equation, the solution mapping $X(t_0, \xi_0; v)(\cdot)$ is continuous from $L^2(t_0, T; \mathbb{R}^M)$ into $C([t_0, T]; \mathbb{R}^N)$. Thus from (3.24) we have

$$x(t) = S(t, t_0)\xi_0 + \int_{t_0}^{t} S(t, s)F(x(s))\, ds + \int_{t_0}^{t} S(t, s)B(s)u(s)\, ds, \tag{3.26}$$

where

$$x(t) = X(t_0, \xi_0; u)(t), \qquad t_0 \leqq t \leqq T,$$

and from (3.22) we have

$$x(T) = \xi_T. \tag{3.27}$$

Since the initial value ξ_0 at $t = t_0$ is arbitrary, the complete controllability for the system (3.16) on $[0, T]$ is shown immediately.

Remark. There are some papers which discuss the complete controllability problem for the nonlinear lumped-parameter system (3.16). Here the condition (3.20) for the complete controllability is similar to one in [9].

4. Examples. In this section some examples of approximate controllability of the systems governed by the semilinear heat equations will be given. For these systems

approximate controllability does not follow from Henry's theorem. But Theorem 3.1 can be applied to show that these systems are approximately controllable.

Example 1. Let $X = L^2(0, \pi)$ and $e_n(x) = \sqrt{2/\pi} \sin nx$, $n = 1, 2, \cdots$. Then $\{e_n, n = 1, 2, \cdots\}$ is an orthonormal base for X. Define an infinite-dimensional space U by

$$U = \left\{ u \,\middle|\, u = \sum_{n=2}^{\infty} u_n e_n \text{ with } \sum_{n=2}^{\infty} u_n^2 < \infty \right\}. \tag{4.1}$$

The norm in U is defined by $\|u\|_U = (\sum_{n=2}^{\infty} u_n^2)^{1/2}$. Define a mapping $B \in \mathscr{L}[U \to X]$ as follows:

$$Bu = 2u_2 e_1 + \sum_{n=2}^{\infty} u_n e_n \quad \text{for } u = \sum_{n=2}^{\infty} u_n e_n \in U. \tag{4.2}$$

Obviously, $\|B\|_{\mathscr{L}[U \to X]} \leqq \sqrt{5}$.

Consider a system governed by the semilinear heat equation

$$\begin{aligned} &\frac{\partial y(t, x)}{\partial t} = \frac{\partial^2 y(t, x)}{\partial x^2} + F(x, y(t, x)) + Bu(t, x), \qquad 0 < t < T, \quad 0 < x < \pi, \\ &y(t, 0) = y(t, \pi) = 0, \qquad 0 \leqq t \leqq T, \\ &y(0, x) = \eta_0(x), \qquad 0 \leqq x \leqq \pi, \end{aligned} \tag{4.3}$$

where the operator B is defined by (4.2) and $u(\cdot, \cdot) \in L^2(0, T; X) = L^2((0, T) \times (0, \pi))$. Here the nonlinear function F is considered as an operator satisfying Hypothesis (F). Assume the initial value satisfies the compatibility condition, i.e., $\eta_0(0) = \eta_0(\pi) = 0$. In this specific case the range $B_{(t,T)}V$ of the operator B is not dense in $L^2(t_0, T; X)$.

The linear system corresponding to the nonlinear heat system is

$$\begin{aligned} &\frac{\partial y(t, x)}{\partial t} = \frac{\partial^2 y(t, x)}{\partial x^2} + Bu(t, x), \qquad 0 < t < T, \quad 0 < x < \pi, \\ &y(t, 0) = y(t, \pi) = 0, \qquad 0 \leqq t \leqq T, \\ &y(0, x) = 0, \qquad 0 \leqq x \leqq \pi, \end{aligned} \tag{4.4}$$

(without loss of generality, we may assume $\eta_0(x) \equiv 0$). Let

$$y(t, x) = \sum_{n=1}^{\infty} y_n(t) e_n(x), \qquad u(t, x) = \sum_{n=2}^{\infty} u_n(t) e_n(x).$$

Then this heat equation (4.4) is equivalent to the following infinite set of ordinary differential equations:

$$\begin{aligned} &\frac{d}{dt} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -4 \end{pmatrix} \begin{pmatrix} y_1(t) \\ y_2(t) \end{pmatrix} + \begin{pmatrix} 2 \\ 1 \end{pmatrix} u_2(t), \\ &y_1(0) = y_2(0) = 0, \end{aligned} \tag{4.5}$$

and

$$\begin{aligned} &\frac{d}{dt} y_n(t) = -n^2 y_n(t) + u_n(t), \\ &y_n(0) = 0, \end{aligned} \qquad n = 3, 4, \cdots. \tag{4.6}$$

Obviously, as the second order system (4.5) is completely controllable in $\mathbb{R}^2$ for any given $T > 0$, the heat system (4.4) is approximately controllable.

Let $f(\cdot)$ be an arbitrary element in $L^2(t_0, T; X)$ and $h \in X$ be defined by

$$h = \int_{t_0}^{T} S(T-t)f(t)\,dt. \tag{4.7}$$

Assume that

$$f(t) = \sum_{n=1}^{\infty} f_n(t)e_n \quad \text{and} \quad h = \sum_{n=1}^{\infty} h_n e_n.$$

According to Lemma 3.4

$$\int_{t_0}^{T} |\hat{f}_n(t)|^2\,dt \leqq \int_{t_0}^{T} |f_n(t)|^2\,dt, \qquad n = 1, 2, \cdots,$$

where

$$\hat{f}_n(t) = \frac{2n^2}{1-e^{-2n^2(T-t_0)}} h_n e^{-n^2(T-t)}, \qquad t_0 \leqq t \leqq T, \quad n = 1, 2, \cdots. \tag{4.8}$$

Define $\hat{f}(t) = \sum_{n=1}^{\infty} \hat{f}_n(t)e_n$. Then

$$\|\hat{f}(\cdot)\|_{L^2(t_0,T;X)} \leqq \|f(\cdot)\|_{L^2(t_0,T;X)}. \tag{4.9}$$

We now claim that there exists some constant q_1 such that

$$\|Bu(\cdot)\|_{L^2(t_0,T;X)} \leqq q_1 \|f(\cdot)\|_{L^2(t_0,T;X)}, \tag{4.10}$$

where the control $u(\cdot)$ satisfies

$$\int_{t_0}^{T} S(T-t)Bu(t)\,dt = h = \int_{t_0}^{T} S(T-t)f(t)\,dt. \tag{4.11}$$

In fact, consider such a control $u(t) = \sum_{n=2}^{\infty} u_n(t)e_n$, where $u_2(t), u_3(t), \cdots$ are defined by

$$\begin{aligned} u_2(t) &= (2 \quad 1)\begin{pmatrix} e^{-(T-t)} & 0 \\ 0 & e^{-4(T-t)} \end{pmatrix} G_b(t_0, T)^{-1}\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}, \qquad t_0 \leqq t \leqq T, \\ u_n(t) &= \frac{2n^2}{1-e^{-2n^2(T-t_0)}} h_n e^{-n^2(T-t)}, \qquad t_0 \leqq t \leqq T, \quad n = 3, 4, \cdots. \end{aligned} \tag{4.12}$$

The matrix $G_b(t_0, T)$ may be obtained as follows

$$\begin{aligned} G_b(t_0, T) &= \int_{t_0}^{T} \begin{pmatrix} 4e^{-2(T-t)} & 2e^{-5(T-t)} \\ 2e^{-5(T-t)} & e^{-8(T-t)} \end{pmatrix} dt \\ &= \begin{pmatrix} 2[1-e^{-2(T-t_0)}] & \frac{2}{5}[1-e^{-5(T-t_0)}] \\ \frac{2}{5}[1-e^{-5(T-t_0)}] & \frac{1}{8}[1-e^{-8(T-t_0)}] \end{pmatrix}. \end{aligned}$$

To illustrate the applications of Theorem 3.1, concrete initial and terminal times are given, for example, $t_0 = 0.9$ and $T = 1$. In this special case one has

$$G_b(0.9, 1) = \begin{pmatrix} 0.3625 & 0.1574 \\ 0.1574 & 0.0688 \end{pmatrix}, \qquad G_b(0.9, 1)^{-1} = \begin{pmatrix} 416.5 & -952.8 \\ -952.8 & 2194 \end{pmatrix}.$$

Thus

$$u_2(t) = 2(416.5h_1 - 952.8h_2)\, e^{-(1-t)} + (-952.8h_1 + 2194h_2)\, e^{-4(1-t)}$$

and

$$\begin{aligned} \int_{t_0}^{T} |u_2(t)|^2\, dt &= 416h_1^2 + 2449h_2^2 - 2123h_1h_2 \\ &\leqq (21.5|h_1| + 49.5|h_2|)^2. \end{aligned} \tag{4.13}$$

Taking the norms of $\hat{f}_1(\cdot)$ and $\hat{f}_2(\cdot)$ in $L^2(0.9, 1)$ gives

$$|h_1| = 0.301|\hat{f}_1(\cdot)|_{L^2(0.9,1)} \quad \text{and} \quad |h_2| = 0.262|\hat{f}_2(\cdot)|_{L^2(0.9,1)}.$$

From (4.13) we obtain immediately that

$$|u_2(\cdot)|_{L^2(0.9,1)} \leqq 19.5\|\hat{f}(\cdot)\|_{L^2(0.9,1;X)}$$

and

$$\begin{aligned} \|Bu(\cdot)\|_{L^2(0.9,1;X)} &= \left(5\int_{0.9}^{1} |u_2(t)|^2\, dt + \sum_{n=3}^{\infty} \int_{0.9}^{1} |u_n(t)|^2\, dt\right)^{1/2} \\ &\leqq (5|u_2(\cdot)|^2_{L^2(0.9,1)} + \|\hat{f}(\cdot)\|^2_{L^2(0.9,1;X)})^{1/2} \\ &\leqq 43.6\|\hat{f}(\cdot)\|_{L^2(0.9,1;X)} \\ &\leqq 43.6\|f(\cdot)\|_{L^2(0.9,1;X)}. \end{aligned} \tag{4.14}$$

This is (4.10) with $q_1 = 43.6$, $t_0 = 0.9$ and $T = 1$. Thus the condition (3.3) in Hypothesis (B) reduces to

$$6.2K_1 < 1; \tag{4.15}$$

i.e., if the Lipschitz constant K_1 for the nonlinear function F satisfies the requirement (4.15) then the nonlinear heat system (4.3) is approximately controllable for every given initial state $\eta_0(\cdot) \in L^2(0, \pi)$ on the time interval $[0, 1]$. □

Example 2. Let $X = L^2(0, \pi)$ and $\{e_n,\ n = 1, 2, \cdots\}$ be an orthonormal base as the one in Example 1. Define

$$V = L^2(0, T; X). \tag{4.16}$$

For every $u(\cdot) \in V$ of the form $u(t) = \sum_{n=1}^{\infty} u_n(t)e_n$ define

$$B_{(0,T)}u(\cdot)(t) = \sum_{n=1}^{\infty} \hat{u}_n(t)e_n, \tag{4.17}$$

where

$$\hat{u}_n(t) = \begin{cases} 0 & 0 \leqq t < T\left(1 - \dfrac{1}{n^2}\right), \\ u_n(t), & T\left(1 - \dfrac{1}{n^2}\right) \leqq t \leqq T, \end{cases} \qquad n = 1, 2, \cdots. \tag{4.18}$$

Since

$$\|B_{(0,T)}u(\cdot)(\cdot)\|_{L^2(0,T;X)} \leqq \|u(\cdot)\|_{L^2(0,T;X)}$$

the operator $B_{(0,T)}$ is bounded in $\mathscr{L}[V \to L^2(0, T; X)]$. It is not difficult to see that the range $B_{(t_0,T)}V$ of the intercept $B_{(t_0,T)}$ of $B_{(0,T)}$ is not dense in $L^2(t_0, T; X)$ for

any $t_0 \in [0, T)$. In fact, if $t_0 \in [0, T)$ is given, there exists some positive integer N such that

$$T\left(1-\frac{1}{N^2}\right) \leqq t_0 < T\left(1-\frac{1}{(N+1)^2}\right).$$

Hence for any given function $f(\cdot) \in L^2(t_0, T; X)$ and $u(\cdot) \in V$ one has

$$\begin{aligned}
&\int_{t_0}^{T} \|f(t) - B_{(t_0,T)}u(\cdot)(t)\|^2 \, dt \\
&\quad = \sum_{n=1}^{\infty} \int_{t_0}^{T} |f_n(t) - \hat{u}_n(t)|^2 \, dt \\
&\quad = \sum_{n=1}^{N} \int_{t_0}^{T} |f_n(t) - u_n(t)|^2 \, dt + \sum_{n=N+1}^{\infty} \int_{T(1-(1/n^2))}^{T} |f_n(t) - u_n(t)|^2 \, dt \\
&\qquad\qquad + \sum_{n=N+1}^{\infty} \int_{t_0}^{T(1-(1/n^2))} |f_n(t)|^2 \, dt.
\end{aligned}$$

Consider such a function $g(t) = \sum_{n=1}^{\infty} g_n(t) e_n$, where $g_n(\cdot)$ is defined by

$$g_n(t) = 0 \quad \text{for } t_0 \leqq t \leqq T, \quad n = 1, 2, \cdots, N,$$

$$g_n(t) = \begin{cases} \dfrac{N+1}{n\sqrt{T - t_0 - \dfrac{T}{(N+1)^2}}} & \text{for } t_0 \leqq t \leqq T\left(1-\dfrac{1}{n^2}\right), \\ & \qquad n = N+1, N+2, \cdots. \\ 0 & \text{for } T\left(1-\dfrac{1}{n^2}\right) < t \leqq T, \end{cases}$$

Since

$$\int_{t_0}^{T} g_n^2(t) \, dt = \frac{(N+1)^2}{n^2} \frac{T - t_0 - T/n^2}{T - t_0 - T/(N+1)^2}, \qquad n = N+1, N+2, \cdots,$$

thus

$$\begin{aligned}
\frac{(N+1)^2(T-t_0)}{T - t_0 - T/(N+1)^2} \sum_{n=N+1}^{\infty} \frac{1}{n^2} &\geqq \frac{(N+1)^2}{T - t_0 - T/(N+1)^2} \sum_{n=N+1}^{\infty} \frac{T - t_0 - T/n^2}{n^2} \\
&\geqq (N+1)^2 \sum_{n=N+1}^{\infty} \frac{1}{n^2} > 1.
\end{aligned}$$

This means that

$$g(\cdot) \in L^2(t_0, T; X) \quad \text{and} \quad \int_{t_0}^{T} \|g(t) - B_{(t_0,T)}u(\cdot)(t)\|^2 \, dt \geqq 1$$

for every $u \in V$. Hence $\overline{B_{(t_0,T)}V} \neq L^2(t_0, T; X)$ for any $t_0 \in [0, T)$.

Consider a nonlinear heat control system (1.1) with corresponding linear system

$$\begin{aligned}
&\frac{\partial y(t,x)}{\partial t} = \frac{\partial^2 y(t,x)}{\partial x^2} + B_{(0,T)}u(\cdot,\cdot)(t,x), \qquad 0 < t < T, \quad 0 < x < \pi, \\
&y(t,0) = y(t,\pi) = 0, \qquad 0 \leqq t \leqq T, \qquad (4.19) \\
&y(0,x) = 0, \qquad 0 \leqq x \leqq \pi.
\end{aligned}$$

We claim that for every given $f(\cdot)\in L^2(0, T; X)$ there exists an $u(\cdot)\in V$ such that

$$\int_0^T S(T-t)B_{(0,T)}u(\cdot)(t)\,dt = \int_0^T S(T-t)f(t)\,dt \tag{4.20}$$

and

$$\|B_{(0,T)}u(\cdot)(\cdot)\|_{L^2(0,T;X)} \leqq (1-e^{-2T})^{-1/2}\|f(\cdot)\|_{L^2(0,T;X)}.$$

As in the example above consider the infinite set of lumped-parameter control systems equivalent to (4.19):

$$\begin{aligned} &\frac{dy_n(t)}{dt} = -n^2 y_n(t) + \hat{u}_n(t), \qquad 0<t<T, \\ &y_n(0) = 0. \end{aligned} \tag{4.21}$$

By the definition of $B_{(0,T)}$ (or $\hat{u}_n(\cdot)$) (4.21) is equivalent to

$$\begin{aligned} &\frac{dy_n(t)}{dt} = -n^2 y_n(t) + \tilde{u}_n(t), \qquad T\left(1-\frac{1}{n^2}\right)<t<T, \\ &y_n\left(T-\frac{T}{n^2}\right) = 0. \end{aligned} \tag{4.22}$$

If $f(\cdot)\in L^2(0, T; X)$ and h is defined by (4.7) with $t_0 = 0$, then the control $\tilde{u}_n(t)$, $T(1-(1/n^2)) \leqq t \leqq T$, in (4.22) may be selected as

$$\tilde{u}_n(t) = \frac{2n^2}{1-e^{-2T}}\, h_n\, e^{-n^2(T-t)}, \qquad T\left(1-\frac{1}{n^2}\right) \leqq t \leqq T, \tag{4.23}$$

which satisfies

$$h_n = \int_{T(1-(1/n^2))}^T e^{-n^2(T-t)}\tilde{u}_n(t)\,dt.$$

Define

$$u(t) = \sum_{n=1}^{\infty} u_n(t)e_n, \qquad u_n(t) = \begin{cases} 0, & 0 \leqq t < T\left(1-\dfrac{1}{n^2}\right), \\ \tilde{u}_n(t), & T\left(1-\dfrac{1}{n_2}\right) \leqq t \leqq T. \end{cases}$$

Then (4.20) is satisfied. The estimate (4.21) is verified by the definitions of $\tilde{u}_n(\cdot)$ and $u(\cdot)$ and (4.8) with $t_0 = 0$:

$$\begin{aligned} \|B_{(0,T)}u(\cdot)(\cdot)\|^2_{L^2(0,T;X)} &= \sum_{n=1}^{\infty} \int_{T(1-(1/n^2))}^T |\tilde{u}_n(t)|^2\,dt \\ &= \frac{1}{1-e^{-2T}} \sum_{n=1}^{\infty} 2n^2 h_n^2 \\ &= \frac{1}{1-e^{-2T}} \sum_{n=1}^{\infty} (1-e^{-2n^2T}) \int_0^T |\hat{f}_n(t)|^2\,dt \\ &\leqq \frac{1}{1-e^{-2T}} \|f(\cdot)\|^2_{L^2(0,T;X)}. \end{aligned}$$

Now return to the nonlinear heat control system (1.1). If the Lipschitz constant K_1 for the nonlinear function F satisfies

$$K_1 T\left(\frac{2}{1-e^{-2T}}\right)^{1/2} < 1, \tag{4.24}$$

for instance, $K_1 < 0.66$ with $T = 1$, then the nonlinear system is approximately controllable on $[0, T]$ under Hypothesis (F). □

Acknowledgments. The author wishes to express his thanks to Professor H. O. Fattorini for reading the manuscript and many stimulating discussions and to Professor T. I. Seidman for his comments on Remark 2.2.

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.
[2] G. CHEN, W. H. MILLS AND G. CROSTA, *Exact controllability theorems and numerical simulations for some nonlinear differential equations*, Pennsylvania State University, University Park, 1980, this Journal, 19 (1981), pp. 765–790.
[3] H. O. FATTORINI, *On complete controllability of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.
[4] ———, *Local controllability of a nonlinear wave equation*, Math. Systems Theory, 9 (1975), pp. 30–45.
[5] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, New York, 1981.
[6] J. HENRY, *Etude de la controlabilité de certains équations paraboliques non-linéaires*, Thèse d'état, Paris, June 1978.
[7] K. MIRZA AND B. F. WOMACK, *On the controllability of a class of non-linear systems*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 531–535.
[8] D. L. RUSSELL, *Mathematics of Finite-Dimensional Linear Control Systems, Theory and Design*, Marcel Dekker, New York and Basel, 1979.
[9] ———, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639–739.
[10] T. I. SEIDMAN AND H. X. ZHOU, *Existence and uniqueness of optimal controls for a quasilinear parabolic equation*, this Journal, 20 (1982), pp. 747–762.
[11] E. L. TONKOV, *Controllability of a nonlinear system in a linear approximation*, Prikl Matem. i Mech., 38 (1974), pp. 599–606. (In Russian.)

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

0363-0129/83/2104-0004 $01.25/0

SINGULAR PERTURBATION IN MAYER'S PROBLEM FOR LINEAR SYSTEMS*

A. L. DONTCHEV† AND V. M. VELIOV†

Abstract. The well-posedness of Mayer's problem for a linear system with a small parameter in the derivatives and constrained controls is studied. The convergence of the optimal control when the parameter tends to zero is analyzed. The state constrained case is discussed.

Key words. singular perturbation, optimal control, Mayer's problem, well-posedness

1. Introduction.

Consider the following singularly perturbed control system:

$$\begin{aligned} &\dot{x} = A_1(t)x + A_2(t)y + B_1(t)u(t), && x(0) = x^0, \\ &\lambda\dot{y} = A_3(t)x + A_4(t)y + B_2(t)u(t), && y(0) = y^0 \end{aligned} \tag{1}$$

for $t \in [0, T]$, where the final time $T \in (0, +\infty)$ is fixed, the "slow" state $x(t) \in R^m$ and the "fast" state $y(t) \in R^n$. The set of feasible controls is

$$U = \{u(\cdot) \in L^1(R^r, (0, T)); u(t) \in V \subset R^r \text{ a.e. } t \in (0, T)\}. \tag{2}$$

The singular perturbation is provided by the positive scalar parameter λ, $\lambda \in (0, T^2)$. We assume that:

(A1) The matrices $A_i(t), i = 1, \cdots, 4, B_j(t), j = 1, 2$, are continuous; the eigenvalues of the matrix $A_4(t)$ have strictly negative real parts for $t \in [0, T]$.

For $\lambda = 0$ we obtain the reduced system

$$\dot{x} = A_0(t)x + B_0(t)u(t), \qquad x(0) = x^0, \tag{3a}$$

$$y(t) = -A_4^{-1}(t)(A_3(t)x(t) + B_2(t)u(t)) \tag{3b}$$

for $t \in [0, T]$, where $A_0 = A_1 - A_2A_4^{-1}A_3$, $B_0 = B_1 - A_2A_4^{-1}B_2$.

Let K_λ be the attainable set for the system (1) at the time T, that is $K_\lambda = \{z \in R^{m+n}; \exists u(\cdot) \in U, z = (x(T), y(T))$, where $(x(\cdot), y(\cdot))$ is the solution of (1) corresponding to $u(\cdot)\}$. This paper is concerned with the well-posedness of the order reduction procedure for the following optimal control problem:

$$\inf\{g(x, y), (x, y) \in K_\lambda\} = \hat{g}_\lambda. \tag{4}$$

Formally substituting (3) into (4) we get the problem

$$\inf\{g(x, -A_4^{-1}(T)(A_3(T)x + B_2(T)v)), x \in P_0, v \in V\} = \tilde{g}, \tag{5}$$

where P_0 is the attainable set for the system (3a). Such a definition of the limit problem, however, does not provide well-posedness of (4): see the following example.

Example 1. For $t \in [0, 1]$ consider the system

$$\dot{x} = y_1 - y_2, \qquad x(0) = -0.5, \tag{6a}$$

$$\begin{aligned} &\lambda\dot{y}_1 = -y_1 + u(t), && y_1(0) = 0, \\ &\lambda\dot{y}_2 = -2y_2 + u(t), && y_2(0) = 0. \end{aligned} \tag{6b}$$

* Received by the editors July 17, 1981, and in revised form March 15, 1982.
† Institute of Mathematics, Bulgaria Academy of Sciences, 1090 Sofia, P.O. Box 373, Bulgaria.

Let $u(t)\in[-1, 1]$ and $g(x, y_1, y_2)=x^2+y_1^2+(y_2+0.25)^2$. The problem (5) for $\lambda=0$ becomes

$$\tilde{g}=\min\{x^2(1)+v^2+(0.5v+0.25)^2, v\in[-1, 1]\}$$

for the reduced system

$$\dot{x}=0.5u(t), \quad x(0)=-0.5, \quad u(t)\in[-1, 1]. \tag{7}$$

We obtain in fact two independent problems: a Mayer's problem for (7) and a one-dimensional minimization. Clearly, the optimal control $\tilde{u}(t)\equiv 1$ and $\tilde{g}=0.05$. Applying the control

$$u(t)=\begin{cases} 1 & \text{for } t\in[0, 1+\lambda \ln 0.5], \\ -1 & \text{for } t\in(1+\lambda \ln 0.5, 1] \end{cases}$$

to the perturbed system (6a, b) we get

$$x(1)=\lambda(-0.5+\ln 0.5+e^{-1/\lambda}-0.25\, e^{-2/\lambda}),$$
$$y_1(1)=e^{-1/\lambda}, \qquad y_2(1)=-0.25-0.5\, e^{-2/\lambda},$$

and

$$\hat{g}_\lambda \leqq g(x(1), y_1(1), y_2(1))\to 0, \qquad \lambda\to 0.$$

Hence, for small λ

$$\hat{g}_\lambda<\tilde{g}-0.01.$$

Thus the question arises of how to define an optimal control problem for the reduced system (3) which is a "limit" of the perturbed problem (4). This paper gives an answer to this question.

Convergence of the solutions of singularly perturbed Mayer's problems has been investigated by Dmitriev [4] for a system which is linear with respect to the state, and by Binding [2] for a general nonlinear system. In both papers the performance index does not depend explicitly on the fast states. In this case the effect illustrated in Example 1 disappears. Considering a more general performance index we limit our investigations here to linear systems.

The discontinuity effect in the terminal part of the performance index was first observed by Glizer and Dmitriev [7] for a linear-quadratic problem without constraints. The investigations in this direction are developed in [5] for strictly convex control problems. The analysis in [5] is based on the observation that, if the reduced and the fast systems are controllable (with unconstrained controls) then the attainable set for small λ is the entire space. This approach, however, cannot be applied to the constrained case.

In § 2 we give a definition of the limit problem, corresponding to the considered one, which guarantees well-posedness without any controllability assumptions. Section 3 deals with the optimal performance convergence. A possible generalization of our analysis to problems with functionals including an integral part is discussed. In § 4 we concern ourselves with the convergence of the optimal controls. Section 5 is devoted to state constrained problems.

2. Statement of the limit problem. In the sequel we assume that:

(A2) The function $g(\cdot, \cdot)$ is continuous; the set V is compact and convex.

Since for each $\lambda>0$ the attainable set K_λ is compact, there exist an optimal control $\hat{u}_\lambda(\cdot)$ and a corresponding final state $(\hat{x}_\lambda, \hat{y}_\lambda)$ for problem (4). For every $x\in R^m$

we define the set

$$R(x) = -A_4^{-1}(T)A_3(T)x + R, \tag{8}$$

where

$$R = \int_0^{+\infty} \exp(A_4(T)s)B_2(T)V\,ds.$$

Here the integral of the set-valued function is taken in the sense of Aumann [1]. From the stability of the matrix $A_4(T)$ (see Assumption (A1)) it follows that this integral exists. Moreover, R is a convex and compact set in R^n.

Let

$$K_0 = \{(x, y) \in R^{m+n};\ x \in P_0,\ y \in R(x)\}. \tag{9}$$

We define the following limit problem:

$$\inf\{g(x, y), (x, y) \in K_0\} = \hat{g}_0. \tag{10}$$

Since the set K_0 is compact, there exists an optimal control $\hat{u}_0(\cdot)$, which when applied to (3a) drives the initial state x^0 to $\hat{x}_0$, where $(\hat{x}_0, \hat{y}_0)$ is the solution of (10). Note that the limit problem (10) does not depend on the initial condition $y(0) = y^0$.

We can rewrite the limit problem as follows

$$\inf\{g_0(x), x \in P_0\}, \tag{11}$$

where the (continuous) function $g_0(\cdot)$ is defined by

$$g_0(x) = \inf\{g^0(x, y), g^0(x, y) = g(x, y - A_4^{-1}(T)A_3(T)x), y \in R\}. \tag{12}$$

Such a definition suggests that the limit problem could be solved as a two-stage optimization problem. The goal function $g_0(x)$ of the "outer" problem (11) is to be evaluated by means of the "inner" problem (12). Clearly, in so far as the set R can be effectively approximated, the "inner" problem is a parametric mathematical programming problem. Generally, problem (12) can be considered as a parametric Mayer's problem with performance index $g^0(x, y)$ on infinite time horizon for the system

$$\dot{y} = A_4(T)y + B_2(T)u(t), \qquad y(0) = 0,$$

the attainable set of which at $t = +\infty$ is exactly R. The "outer" problem (11) for the slow state x remains Mayer's problem over $[0, T]$. We shall not go into computational details further noting only that if $g(x, y)$ is separable one can select two important cases: 1) the function $g(x, \cdot)$ is linear; 2) $A_3(T) = 0$. If one of these conditions holds, the problems (11) and (12) can be solved independently. Moreover, then the "outer" problem (11) coincides with the corresponding "outer" problem in (5). The solution of (12) gives only a shift constant for $g_0(x)$, which provides the well-posedness of the order reduction.

The difference between problems (5) and (10) follows from the fact that

$$-A_4^{-1}(T)B_2(T)V \subsetneq R.$$

For our Example 1 this relation is illustrated in Fig. 1. The set $R(x) = R$ has been obtained analytically in the following way:

The boundary ∂R_λ of the attainable set R_λ for the system (6b) at the time $t = 1/\lambda$ can be achieved by means of controls having one switching. Using the switching point

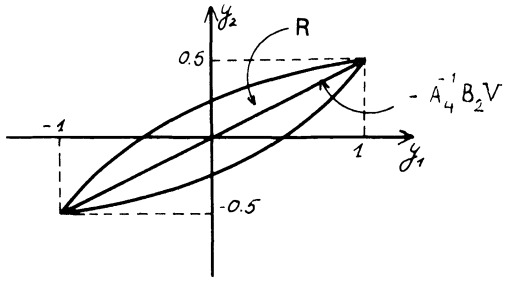


FIG. 1

t_λ as a parameter, one can get that each point $(y_1, y_2) \in \partial R_\lambda$ satisfies

$$y_1 = \pm\left(2\exp\left(\frac{t_\lambda - 1}{\lambda}\right)\right) + O(\lambda), \qquad y_2 = \pm\left(\exp\left(\frac{2(t_\lambda - 1)}{\lambda}\right) - 0.5\right) + O(\lambda).$$

Eliminating t_λ and tending to zero with λ, we obtain that

$$R = \{y = (y_1, y_2) \in R^2;\ -1 \leqq y_1 \leqq 1,\ 0.25(y_1+1)^2 - 0.5 \leqq y_2 \leqq -0.25(y_1-1)^2 + 0.5\}.$$

Since $A_3 = 0$ and $g(x, y)$ is separable, the limit problem consists of two independent problems: the "inner" one is

$$\inf\{y_1^2 + (y_2 + 0.25)^2,\ (y_1, y_2) \in R\} = 0,$$

and the "outer"

$$\min x^2(1) \quad \text{subject to (7).}$$

3. Convergence of the optimal performance. From assumption (A1) it follows that there exist numbers σ_0, $\sigma > 0$, such that if $\phi(t, \tau, \lambda)$ is the fundamental matrix solution of the equation $\lambda\dot{y} = A_4(t)y$, normalized to the identity at $t = \tau$, then

$$|\phi(t, \tau, \lambda)| \leqq \sigma_0 \exp\left(-\sigma\frac{t-\tau}{\lambda}\right) \tag{13}$$

for each t, τ, $t \geqq \tau$ and $\lambda > 0$.

LEMMA 1. *Let $\rho_H(K_\lambda, K_0)$ be the Hausdorff distance between the sets K_λ and K_0. Then*

$$\lim_{\lambda \to 0} \rho_H(K_\lambda, K_0) = 0. \tag{14}$$

Proof. Let $(x_0, y_0) \in K_0$. There exists a control $u_0(\cdot) \in U$ such that the corresponding solution $x_0(\cdot)$ of (3a) satisfies $x_0(T) = x_0$ and there exists an integrable function $v_0(\cdot)$ defined on $[0, +\infty)$ $v_0(t) \in V$ for $t \in [0, +\infty)$ such that

$$y_0 = -A_4^{-1}(T)A_3(T)x_0 + \int_0^{+\infty} \exp(A_4(T)s)B_2(T)v_0(s)\,ds.$$

Define the control

$$u_\lambda(t) = \begin{cases} u_0(t) & \text{for } t \in [0, T - \sqrt{\lambda})], \\ v_0\left(\dfrac{T-t}{\lambda}\right) & \text{for } t \in (T - \sqrt{\lambda}, T]. \end{cases}$$

Clearly, $u_\lambda(\cdot)\in U$. Moreover, $\lim_{\lambda\to 0} u_\lambda(t)=u_0(t)$ for almost all $t\in(0,T)$. Let $(x_\lambda(\cdot), y_\lambda(\cdot))$ be the solution of the perturbed system (1) which results from $u_\lambda(\cdot)$. By Lemma A in the Appendix it follows that $\lim_{\lambda\to 0} x_\lambda(T)=x_0$. Let $\bar{y}_\lambda(\cdot)$ be the solution of the equation

$$\lambda\dot{y}=A_4(T)y+A_3(T)x_\lambda(T)+B_2(T)u_\lambda(t),\qquad t\in[0,T],\quad y(0)=y^0, \tag{15}$$

and let $\Delta y_\lambda(\cdot)=y_\lambda(\cdot)-\bar{y}_\lambda(\cdot)$, $\Delta x_\lambda(t)=x_\lambda(t)-x_\lambda(T)$, $\Delta A_i(t)=A_i(t)-A_i(T)$, $i=3,4$, $\Delta B_2(t)=B_2(t)-B_2(T)$.

We have

$$\begin{aligned}\lambda\,\Delta\dot{y}_\lambda &= A_4(T)\Delta y_\lambda+\Delta A_4(t)y_\lambda(t)+\Delta A_3(t)x_\lambda(t)\\ &\quad +A_3(T)\Delta x_\lambda(t)+\Delta B_2(t)u_\lambda(t),\qquad \Delta y_\lambda(0)=0.\end{aligned} \tag{16}$$

Using the compactness of V, (13) and the Gronwall lemma one can easily prove that $y_\lambda(t)$, $\bar{y}_\lambda(t)$ and $x_\lambda(t)$ are bounded uniformly in t and λ, hence $x_\lambda(\cdot)$ is Lipschitz continuous with respect to t uniformly in λ. From (13) and (16) we get

$$\begin{aligned}|\Delta y_\lambda(T)|\leqq\frac{\sigma_0}{\lambda}\int_0^{T-\sqrt{\lambda}}&\exp\left(-\sigma\frac{T-t}{\lambda}\right)|\Delta A_4(t)y_\lambda(t)+\Delta A_3(t)x_\lambda(t)\\ &+A_3(T)\Delta x_\lambda(t)+\Delta B_2(t)u_\lambda(t)|\,dt\\ +\max_{T-\sqrt{\lambda}\leqq t\leqq T}&(|\Delta A_4(t)y_\lambda(t)|+|\Delta A_3(t)x_\lambda(T)|\\ &+|A_3(T)\Delta x_\lambda(t)|+|\Delta B_2(t)u_\lambda(t)|)\frac{\sigma_0}{\lambda}\int_{T-\sqrt{\lambda}}^{T}\exp\left(-\sigma\frac{T-t}{\lambda}\right)dt.\end{aligned}$$

Hence

$$\lim_{\lambda\to 0}|\Delta y_\lambda(T)|=0. \tag{17}$$

We have

$$\lim_{\lambda\to 0}\frac{1}{\lambda}\int_0^T\exp\left(A_4(T)\frac{T-t}{\lambda}\right)A_3(T)x_\lambda(T)\,dt=-A_4^{-1}(T)A_3(T)x_0 \tag{18}$$

and

$$\begin{aligned}&\frac{1}{\lambda}\int_0^T\exp\left(A_4(T)\frac{T-t}{\lambda}\right)B_2(T)u_\lambda(t)\,dt\\ &\quad=\frac{1}{\lambda}\int_0^{T-\sqrt{\lambda}}\exp\left(A_4(T)\frac{T-t}{\lambda}\right)B_2(T)u_0(t)\,dt\\ &\qquad+\int_0^{T/\sqrt{\lambda}}\exp\left(A_4(T)s\right)B_2(T)v_0(s)\,ds.\end{aligned} \tag{19}$$

Applying (17), (18) and (19) to the Cauchy formula for (15) we conclude that

$$\lim_{\lambda\to 0}|y_\lambda(T)-y_0|=0.$$

Hence, there exists a sequence $(x_\lambda, y_\lambda)\in K_\lambda$ such that

$$\lim_{\lambda\to 0}(x_\lambda, y_\lambda)=(x_0, y_0). \tag{20}$$

Now, let us assume that there exist sequences $\{\lambda_k\}$, $\lim_{k\to+\infty}\lambda_k = 0$ and $\{(x_k, y_k)\}$, $(x_k, y_k)\in K_{\lambda_k}$, $\lim_{k\to+\infty}(x_k, y_k) = (x_0, y_0)\notin K_0$. Let the control $u_k(\cdot)$ correspond to (x_k, y_k) according to (1). The sequence $\{u_k(\cdot)\}$ has a weak limit point $u_0(\cdot)$ in $L^2(R^r, (0, T))$ and let the trajectory $x_0(\cdot)$ result from (3a) for $u_0(\cdot)$. Then, by Lemma A(i), $x_0 = x_0(T)\in P_0$. As before we denote $\Delta y_k(\cdot) = y_k(\cdot) - \bar{y}_k(\cdot)$, where $\bar{y}_k(\cdot)$ is the solution of (15) for $u_k(\cdot)$ and λ_k. By repeating the arguments in (17), we get $\lim_{k\to+\infty}\Delta y_k(T) = 0$, hence $\lim_{k\to+\infty}\bar{y}_k(T) = y_0$. Moreover, using (13) we have

$$\bar{y}_k(T) = -A_4^{-1}(T)A_3(T)x_0 + \int_0^{T/\lambda_k}\exp(A_4(T)s)B_2(T)u_k(T-\lambda_k s)\,ds + \varphi_k^1$$

$$\in R(x_0) + \varphi_k^2,$$

where

$$\lim_{k\to+\infty}\varphi_k^1 = 0, \qquad \lim_{k\to+\infty}\varphi_k^2 = 0.$$

Hence $y_0\in R(x_0)$ and $(x_0, y_0)\in K_0$ which is a contradiction. This, combined with (20), completes the proof.

From (14) and the continuity of $g(\cdot,\cdot)$ one can obtain:

THEOREM 1. *The following relation holds*:

$$\lim_{\lambda\to 0}\hat{g}_\lambda = \hat{g}_0.$$

The proof is standard and therefore it is omitted.

Remark 1. Consider the problem for minimizing the following more general functional:

$$J_\lambda(u(\cdot)) = g(x(T), y(T)) + \int_0^T f(x(t), y(t), u(t))\,dt$$

subject to (1) and (2), where the function g is continuous and the function f satisfies the Carathéodory condition, i.e., it is continuous with respect to (x, y, u) and measurable with respect to t. We assume additionally that the integral part of $J_\lambda(\cdot)$ is lower semicontinuous in the uniform topology for x and in the L^2 weak topology for (y, u); for sufficient conditions see [8, p. 380]. The performance index for the limit problem will have the form

$$J_0(u(\cdot)) = g_0(x(T)) + \int_0^T f(x(t), -A_4^{-1}(t)(A_3(t)x(t) + B_2(t)u(t)), u(t))\,dt.$$

Choosing an L^2 weakly convergent subsequence of $\hat{u}_\lambda(\cdot)$ and using Lemma A(i), (iii) one can get

$$J_0(\hat{u}_0(\cdot)) \leqq \liminf_{\lambda\to 0} J_\lambda(\hat{u}_\lambda(\cdot)). \tag{21}$$

By Lusin's theorem, for each $\varepsilon > 0$, one can find a continuous control $u^\varepsilon(\cdot)\in U$ such that

$$J_0(u^\varepsilon(\cdot)) \leqq J_0(\hat{u}_0(\cdot)) + \varepsilon. \tag{22}$$

Let $x^\varepsilon = x^\varepsilon(T)$ correspond to $u^\varepsilon(\cdot)$ according to (3a) and let y^ε satisfy $g_0(x^\varepsilon) = g(x^\varepsilon, y^\varepsilon)$. Then there exists a control $v^\varepsilon(t)\in V$ for $t\in[0, +\infty)$ such that

$$y^\varepsilon = -A_4^{-1}(T)A_3(T)x^\varepsilon + \int_0^{+\infty}\exp(A_4(T)s)B_2(T)v^\varepsilon(s)\,ds.$$

Introduce the control

$$u_\lambda(t)=\begin{cases} u^\varepsilon(t) & \text{for } t\in[0,\ T-\sqrt{\lambda}],\\ v^\varepsilon\left(\dfrac{T-t}{\lambda}\right) & \text{for } t\in(T-\sqrt{\lambda},\ T].\end{cases}$$

Since $u^\varepsilon(\cdot)$ is continuous, by Lemma A the corresponding trajectory $(x_\lambda^\varepsilon(\cdot), y_\lambda^\varepsilon(\cdot))$ from (1) is pointwise convergent on $(0, T)$ to the trajectory $(x^\varepsilon(\cdot), y^\varepsilon(\cdot))$ corresponding to $u^\varepsilon(\cdot)$ according to (3). Thus

$$\limsup_{\lambda\to 0} J_\lambda(\hat{u}_\lambda(\cdot))\leqq\limsup_{\lambda\to 0} J_\lambda(u_\lambda(\cdot))=J_0(u^\varepsilon(\cdot)).$$

Since ε is arbitrarily small, combining this relation with (21) and (22) we finally obtain

$$\lim_{\lambda\to 0} J_\lambda(\hat{u}_\lambda(\cdot))=J_0(\hat{u}_0(\cdot)).$$

Remark 2. The result obtained in Lemma 1 may be interpreted in the following way: Let $Z_\lambda(t)$ be the family of solutions of the differential inclusion

$$\lambda\dot{y}\in A_4(t)y+B_2(t)V,\qquad y(0)=y^0$$

for $t\in[0, T]$ and $\lambda\in(0, T^2)$. Define the set

$$Z_0(t)=\int_0^{+\infty}\exp(A_4(t)s)B_2(t)V\,ds.$$

Then, for each $t\in(0, T]$

$$\lim_{\lambda\to 0}\rho_H(Z_\lambda(t), Z_0(t))=0.$$

4. Convergence of the optimal control. Throughout this section we assume that (A1) and the following conditions hold:

(A3) The set V is a compact and convex polyhedron in R^r. The components of the matrices $A_1(t)$ and $A_3(t)$ are in $C^{m-2}[0, T]$ and the components of $A_2(t)$, $A_4^{-1}(t)$, $B_1(t)$ and $B_2(t)$ are in $C^{m-1}[0, T]$. For the matrices $C_j(t)$, defined by the relations

$$C_1(t)=B_0(t),\quad C_j(t)=-A_0(t)C_{j-1}(t)+\dot{C}_{j-1}(t),\qquad j=2,\cdots,m,$$

the general position hypothesis holds, that is, if the vector l is parallel to an edge of V, then the vectors $C_1(t)l,\cdots,C_m(t)l$ are linearly independent, see [9, p. 201].

(A4) The function $g(\cdot,\cdot)$ is locally Lipschitz continuous. For each solution $(\hat{x}_0, \hat{y}_0)$ of the limit problem (10) if $(p, q)\in\partial_C g(\hat{x}_0, \hat{y}_0)$ then $p-(A_4^{-1}(T)A_3(T))^*q\neq 0$, where $\partial_C g(\cdot)$ is the subgradient defined by F. Clarke, see [3].

We denote transposition by an asterisk.

As it will be further shown, the condition (A4) is a sufficient condition for $g_0(x)$ to achieve its minimum at the boundary ∂P_0 of the set P_0. (The general position hypothesis implies that $\operatorname{Int} P_0\neq\varnothing$.) Moreover, for sufficiently small λ the function $g(x, y)$ achieves its minimum at the boundary of K_λ.

From Theorem 1 it follows that every L^2 weak limit point of the optimal controls $\hat{u}_\lambda(\cdot)$ is an optimal control for the limit problem. We strengthen this result in the following theorem.

THEOREM 2. *For every $\varepsilon>0$ and for every sequence $\{\lambda_k\}$, $\lim_{k\to+\infty}\lambda_k=0$, for which the sequence of optimal controls $\{\hat{u}_{\lambda_k}(\cdot)\}$ is L^2 weakly convergent to $\hat{u}_0(\cdot)$, there exists*

$N>0$ so that if $k>N$ one can choose a finite number of intervals $\Delta_1, \cdots, \Delta_p$ such that meas $\bigcup_{i=1}^{p} \Delta_i < \varepsilon$ *and* $\hat{u}_{\lambda_k}(t) = \hat{u}_0(t)$ *for almost all* $t \in [0, T] \setminus \bigcup_{i=1}^{p} \Delta_i$.

Proof. Let $(\hat{x}_k(\cdot), \hat{y}_k(\cdot))$ be the optimal trajectory corresponding to $\hat{u}_{\lambda_k}(\cdot)$ and λ_k. We denote by $NM(z_0)$ the normal cone to a convex set $M \subset R^r$ at the point $z_0 \in M$, i.e.

$$NM(z_0) = \{l \in R^r,\ l^*(z - z_0) \leqq 0 \text{ for all } z \in M\}.$$

From [3, Thm. 1] it follows that for each $k = 1, 2, \cdots$ there exists a vector $(p_k, q_k) \in -\partial_C g(\hat{x}_k(T), \hat{y}_k(T)) \cap NK_{\lambda_k}(\hat{x}_k(T), \hat{y}_k(T))$. Moreover, from [3, Lemma 1] we conclude that if there exists $\lim_{k\to+\infty}(x_k(T), y_k(T))$ then from the sequence $\{(p_k, q_k)\}$ one can choose a convergent subsequence.

Let us assume that the statement of the theorem is false for some $\varepsilon_0 > 0$ and for some sequence $\{\lambda_k\}$ for which $\lim_{k\to+\infty} \hat{u}_{\lambda_k}(\cdot) = \hat{u}_0(\cdot)$ in the weak topology of $L^2(R^r, (0, T))$. Without loss of generality we suppose that

$$\lim_{k\to+\infty}(\hat{x}_k(T), \hat{y}_k(T)) = (\hat{x}_0, \hat{y}_0), \qquad \lim_{k\to+\infty}(p_x, q_k) = (p_0, q_0).$$

From Theorem 1 it follows that $(\hat{x}_0, \hat{y}_0)$ solves the limit problem (10) and by Lemma 1 we get that $(p_0, q_0) \in NK_0(\hat{x}_0, \hat{y}_0)$.

We shall prove that

$$\hat{p} = p_0 - (A_4^{-1}(T)A_3(T))^* q_0 \in NP_0(\hat{x}_0). \tag{23}$$

Let $\bar{x} \in P_0$ and $\bar{u}(\cdot) \in U$ be the corresponding control according to (3a). Since $\hat{y}_0 \in R(\hat{x}_0)$ there exists $\hat{v}_0(t) \in V$ for $t \in [0, +\infty)$ such that

$$\hat{y}_0 = -A_4^{-1}(T)A_3(T)\hat{x}_0 + \int_0^{+\infty} \exp(A_4(T)s)B_2(T)\hat{v}_0(s)\,ds.$$

Define the control

$$\bar{u}_k(t) = \begin{cases} \bar{u}(t) & \text{for } t \in [0, T - \sqrt{\lambda_k}], \\ \hat{v}_0\left(\dfrac{T-t}{\lambda_k}\right) & \text{for } t \in (T - \sqrt{\lambda_k}, T]. \end{cases}$$

If $(\bar{x}_k(\cdot), \bar{y}_k(\cdot))$ corresponds to $\bar{u}_k(\cdot)$ according to (1), for $\lambda = \lambda_k$, then from Lemma A(ii) it follows that $\lim_{k\to+\infty} \bar{x}_k(T) = \bar{x}$. Moreover, from the proof of Lemma 1 we get

$$\begin{aligned} \lim_{k\to+\infty} \bar{y}_k(T) &= -A_4^{-1}(T)A_3(T)\bar{x} + \int_0^{+\infty} \exp(A_4(T)s)B_2(T)v_0(s)\,ds \\ &= -A_4^{-1}(T)A_3(T)(\bar{x} - \hat{x}_0) + \hat{y}_0. \end{aligned}$$

Then

$$\begin{aligned} \hat{p}^*(\bar{x} - \hat{x}_0) &= p_0^*(\bar{x} - \hat{x}_0) + q_0^*(-A_4^{-1}(T)A_3(T)(\bar{x} - \hat{x}_0)) \\ &= \lim_{k\to+\infty}(p_k^*(\bar{x}_k(T) - \hat{x}_k(T)) + q_k^*(\bar{y}_k(T) - \hat{y}_k(T))) \leqq 0, \end{aligned}$$

since $(p_k, q_k) \in NK_{\lambda_k}(\hat{x}_k(T), \hat{y}_k(T))$. This proves (23).

Let us denote by $\psi_0(\cdot)$ the solution of the adjoint equation

$$\dot{\psi} = -A_0^*(t)\psi, \qquad \psi(T) = \hat{p}. \tag{24}$$

From [3, Prop. 7] we have $(p_0, q_0) \in -\partial_C g(\hat{x}_0, \hat{y}_0)$, hence by (A4) $\hat{p} \neq 0$. In [9, p. 202] it is proved that if the general position hypothesis holds and $\psi_0(\cdot) \neq 0$, then there

exists a finite number of points $t_2, \cdots, t_{p-1}$ such that $\psi_0^*(t)B_0(t)l \neq 0$ for every $t \in [\varepsilon_0/4, T-\varepsilon_0/4]\backslash\{t_2, \cdots, t_{p-1}\}$ and for every vector l which is parallel to an edge of V. Let the intervals $\Delta_1, \cdots, \Delta_p$ be centered at $0, t_2, \cdots, t_{p-1}, T$ and meas $\Delta_j < \varepsilon_0/2p$, $j = 1, \cdots, p$. Since $\hat{p} \in NP_0(\hat{x}_0)$ the optimal control $\hat{u}_0(\cdot)$ is uniquely defined on $[0, T]\backslash\cup_{j=1}^p \Delta_j$ by the maximum principle

$$\psi_0^*(t)B_0(t)\hat{u}_0(t) = \max\{\psi_0^*(t)B_0(t)v, v \in V\}. \tag{25}$$

Relations (24) and (25) can be rewritten as

$$\dot{\psi} = -A_1^*(t)\psi - A_3^*(t)\eta, \qquad \psi(T) = \hat{p},$$

$$0 = -A_2^*(t)\psi - A_4^*(t)\eta,$$

$$(\psi^*(t)B_1(t) + \eta^*(t)B_2(t))\hat{u}_0(t) = \max\{(\psi^*(t)B_1(t) + \eta^*(t)B_2(t))v, v \in V\}.$$

Since $(p_k, q_k) \in NK_{\lambda_k}(\hat{x}_k(T), \hat{y}_k(T))$, for the perturbed problem we have

$$(\psi_k^*(t)B_1(t) + \eta_k^*(t)B_2(t))\hat{u}_{\lambda_k}(t) = \max\{(\psi_k^*(t)B_1(t) + \eta_k^*(t)B_2(t))v, v \in V\},$$

where $(\psi_k(\cdot), \eta_k(\cdot))$ solves

$$\dot{\psi} = -A_1^*(t)\psi - A_3^*(t)\eta, \qquad \psi_k(T) = p_k,$$

$$\lambda_k\dot{\eta}_k = -A_2^*(t)\psi - A_4^*(t)\eta, \qquad \eta_k(T) = \frac{q_k}{\lambda_k}.$$

From Lemma A(i), (iv) it follows that

$$\lim_{k\to+\infty} \max_{0\leq t\leq T-\varepsilon_0/2} (|\psi_k(t) - \psi(t)| + |\eta_k(t) - \eta(t)|) = 0.$$

Hence, for sufficiently large k and for every vector l, which is parallel to an edge of V we have

$$\xi_k^*(t)l = (\psi_k^*(t)B_1(t) + \eta_k^*(t)B_2(t))l \neq 0$$

and

$$\max\{\xi_k^*(t)v, v \in V\} = \xi_k^*(t)\hat{u}_0(t)$$

for each $t \in [0, T]\backslash\cup_{j=1}^p \Delta_j$. The obtained contradiction completes the proof.

Remark 3. If the limit problem (10) has a unique solution $\hat{u}_0(\cdot)$ then Theorem 2 can be formulated in the following way: for each $\varepsilon > 0$ there exists $\Lambda > 0$ such that if $\lambda \in (0, \Lambda)$ and $\hat{u}_\lambda(\cdot)$ is an optimal control for (4) then there exists a finite number of intervals $\Delta_1, \cdots, \Delta_p$ such that meas $\cup_{j=1}^p \Delta_j < \varepsilon$ and $\hat{u}_\lambda(t) = \hat{u}_0(t)$ for almost all $t \in [0, T]\backslash\cup_{j=1}^p \Delta_j$.

In order to obtain uniqueness of $\hat{u}_0(\cdot)$ it is sufficient to assume that $g(\cdot, \cdot)$ is convex. We prove this statement. Let $(\hat{x}_1, \hat{y}_1) \in K_0$ and $(\hat{x}_2, \hat{y}_2) \in K_0$ be two different solutions of (10). Then $(\hat{x}_0, \hat{y}_0) = ((\hat{x}_1, \hat{y}_1) + (\hat{x}_2, \hat{y}_2))/2$ is a solution. Moreover, since the set P_0 is strictly convex, we have $\hat{x}_0 \in \operatorname{Int} P_0$. From [3, Thm. 1] we conclude that there exists $(\tilde{p}, \tilde{q}) \in -\partial_C g(\hat{x}_0, \hat{y}_0) \cap NK_0(\hat{x}_0, \hat{y}_0)$. Let $x \in P_0$ be arbitrarily chosen and let $y = -A_4^{-1}(T)A_3(T)(x - \hat{x}_0) + \hat{y}_0$. Clearly $y \in R(x)$. Then

$$0 \geq \tilde{p}^*(x - \hat{x}_0) + \tilde{q}^*(y - \hat{y}_0) = (\tilde{p} - (A_4^{-1}(T)A_3(T))^*\tilde{q})^*(x - \hat{x}_0),$$

which combined with $\hat{x}_0 \in \operatorname{Int} P_0$ implies that $\tilde{p} - (A_4^{-1}(T)A_3(T))^*\tilde{q} = 0$. This contradicts assumption (A4). Thus, the general position hypothesis implies that the optimal control $\hat{u}_0(\cdot)$ is unique, see [9, p. 139].

By repeating the arguments in [6, Thm. 3.2] one can prove:

THEOREM 3. *Suppose that the optimal control $\hat{u}_0(\cdot)$ is unique. Then for every $\varepsilon > 0$ there exists $\Lambda > 0$ such that if $\lambda \in (0, \Lambda)$ and $(\hat{x}_\lambda(\cdot), \hat{y}_\lambda(\cdot))$ is an optimal trajectory for the perturbed problem (4), then there exists a finite number of intervals $\Delta_1, \cdots, \Delta_p$ such that* meas $\bigcup_{j=1}^{p} \Delta_j < \varepsilon$ *and*

$$\max_{t\in[0,T]} |\hat{x}_\lambda(t) - \hat{x}_0(t)| + \sup_{t\in[0,T]\setminus\bigcup_{j=1}^{p}\Delta_j} |\hat{y}_\lambda(t) - \hat{y}_0(t)| < \varepsilon,$$

where $\hat{x}_0(\cdot)$ is the optimal trajectory for problem (10) and $\hat{y}_0(\cdot)$ satisfies (3b) for $u = \hat{u}_0(t)$ and $x = \hat{x}_0(t)$.

Remark 4. Following [5] one can develop conditions for semiuniform convergence of the optimal controls for problems with an integral performance index, see Remark 1.

5. State constraints. In this section we show that the analysis of § 3 can be extended to problem (4) with additional state constraints of the form

$$x(t) \in X \quad \text{for all } t \in [0, T], \tag{26}$$

assuming that (A1), (A2) and the following condition hold:

(A5) The set $X \subset R^m$ is closed and convex and has nonempty interior. There exists a control $\tilde{u}(\cdot) \in U$ such that if $\tilde{x}(\cdot)$ is the corresponding solution of (3a) then $\tilde{x}(t) \in \operatorname{Int} X$ for all $t \in [0, T]$.

We denote as U_x the set of admissible controls for the reduced system, that is: $u(\cdot) \in U_x$ if $u(\cdot) \in U$ and the corresponding state of (3a) satisfies (26). As before, let P_0 be the attainable set for the system (3a) (with controls in U_x) and let K_λ be the attainable set for the full-order system (1). The sets $R(x)$ and K_0 are defined as in (8) and (9). The limit problem has the form of (10).

In this case the statements of Lemma 1 and Theorem 1 hold true as well. We need only the following modification of the proof of Lemma 1.

Let $(x_0, y_0) \in K_0$, $u_0(\cdot) \in U_x$, $v_0(t) \in V$ for $t \in [0, +\infty)$ and $u_\lambda(\cdot)$ be chosen as in the proof of Lemma 1. By Lemma A(ii) we get that $\lim_{\lambda\to 0} \|x_\lambda(\cdot) - x_0(\cdot)\|_C = 0$, where $x_\lambda(\cdot)$ solves (1) for $u = u_\lambda(\cdot)$. There exists a function $\varepsilon(\lambda) \in (0, 1)$ such that $\lim_{\lambda\to 0} \varepsilon(\lambda) = 0$ and $\lim_{\lambda\to 0} (\|x_\lambda(\cdot) - x_0(\cdot)\|_C + \sqrt{\lambda})/\varepsilon(\lambda) = 0$. Define the control

$$\bar{u}_\lambda(t) = \begin{cases} (1-\varepsilon(\lambda))u_0(t) + \varepsilon(\lambda)\tilde{u}(t) & \text{for } t \in [0, T-\sqrt{\lambda}], \\ v_0\left(\dfrac{T-t}{\lambda}\right) & \text{for } t \in (T-\sqrt{\lambda}, T]. \end{cases}$$

Clearly, $\bar{u}_\lambda(\cdot) \in U$. Let $(\bar{x}_\lambda(\cdot), \bar{y}_\lambda(\cdot))$ correspond to $\bar{u}_\lambda(\cdot)$ according to (1). Since $x_0(t) \in X$ one can easily deduce that

$$\bar{x}_\lambda(t) = (1-\varepsilon(\lambda))x_\lambda(t) + \varepsilon(\lambda)\tilde{x}_\lambda(t) \in X \quad \text{for each } t \in [0, T-\sqrt{\lambda}].$$

We show that there exists a constant $\alpha > 0$ such that

$$\operatorname{dist}(\bar{x}_\lambda(T-\sqrt{\lambda}), \partial X) \geqq \alpha\varepsilon(\lambda).$$

Denote $t_\lambda = T - \sqrt{\lambda}$ and let

$$|\bar{x}_\lambda(t_\lambda) - z_\lambda| = \operatorname{dist}(\bar{x}_\lambda(T-\sqrt{\lambda}), \partial X), \qquad l_\lambda = z_\lambda - (1-\varepsilon(\lambda))x_0(t_\lambda) - \varepsilon(\lambda)\tilde{x}(t_\lambda).$$

Since $z_\lambda \in \partial X$ and $\tilde{x}(t_\lambda) \in \operatorname{Int} X$, $l_\lambda \neq 0$. From (A5) it follows that there exists $\varepsilon_0 > 0$ such that $\tilde{x}(t_\lambda) + \varepsilon_0 l_\lambda / |l_\lambda| \in X$. Then

$$(1-\varepsilon(\lambda))x_0(t_\lambda) + \varepsilon(\lambda)\tilde{x}(t_\lambda) + \frac{\varepsilon_0 \varepsilon(\lambda) l_\lambda}{|l_\lambda|} \in X.$$

By the definition of l_λ we get $\varepsilon_0 \varepsilon(\lambda) \leqq |l_\lambda|$. For λ sufficiently small we have

$$\begin{aligned}
|\bar{x}_\lambda(t_\lambda) - z_\lambda| &= |(1-\varepsilon(\lambda))x_\lambda(t_\lambda) + \varepsilon(\lambda)\tilde{x}_\lambda(t_\lambda) - z_\lambda| \\
&\geqq |l_\lambda| - (1-\varepsilon(\lambda))|x_\lambda(t_\lambda) - x_0(t_\lambda)| - \varepsilon(\lambda)|\tilde{x}_\lambda(t_\lambda) - \tilde{x}(t_\lambda)| \\
&\geqq \varepsilon_0 \varepsilon(\lambda) - (1-\varepsilon(\lambda))\|x_\lambda(\cdot) - x_0(\cdot)\|_C - \varepsilon(\lambda)\|\tilde{x}_\lambda(\cdot) - \tilde{x}(\cdot)\|_C \\
&\geqq \varepsilon_0 \varepsilon(\lambda)/2.
\end{aligned}$$

Thus, since $|\bar{x}_\lambda(T - \sqrt{\lambda}) - \bar{x}_\lambda(t)| = O(\sqrt{\lambda})$ for all $t \in [T - \sqrt{\lambda}, T]$ and $\lim_{\lambda \to 0} \sqrt{\lambda}/\varepsilon(\lambda) = 0$, for λ sufficiently small we obtain that $\bar{x}_\lambda(t) \in X$ for all $t \in [0, T]$. This means that $\bar{u}_\lambda(\cdot) \in U_x$. As in Lemma 1 one can prove that $\lim_{\lambda \to 0} \bar{x}_\lambda(T) = x_0$ and $\lim_{\lambda \to 0} \bar{y}_\lambda(T) = y_0$, this is completely analogous to the proof of Lemma 1.

The presence of state constraints for the fast variables complicates the situation considerably. The following example shows that even in the case where the function g is not dependent on the fast states, the substitution $\lambda = 0$ in (1) does not define a limit problem.

Example 2. Minimize $x^2(1)$ subject to

$$\begin{aligned}
&\dot{x} = y_1, && x(0) = -1, \\
&\lambda \dot{y}_1 = -y_1 + y_2, && y_1(0) = 0, \\
&\lambda \dot{y}_2 = -y_2 + u(t), && y_2(0) = e, \\
&t \in [0, 1], \quad u(t) \in [0, 1], \quad y_1(t) \in [-1, 1].
\end{aligned}$$

The "reduced" problem consists of minimizing $x^2(1)$ for

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

and has a solution $\hat{u}(t) \equiv 1$, which gives $\hat{x}(1) = 0$.

For the perturbed problem we have

$$y_2(t) = e^{-t/\lambda} e + \frac{1}{\lambda} \int_0^t e^{-(t-s)/\lambda} u(s)\, ds \geqq e^{1-t/\lambda},$$

$$y_1(t) = \frac{1}{\lambda} \int_0^t e^{-(t-s)/\lambda} y_2(s)\, ds \geqq \frac{t}{\lambda} e^{1-t/\lambda}.$$

Hence the only feasible control is $u(t) \equiv 0$, which gives value 1 for the performance index.

In the case considered the fast state constraint "transfers" the singularity to the slow system and changes essentially the attainable set of the full-order system. It can be shown that for some special sequences, for example $\{\lambda_k\}$ such that $\lim_{k \to +\infty} ((\lambda_{k+1} - \lambda_k)/\lambda_k) = 0$, the sequence of the attainable sets is a fundamental sequence according to the Hausdorff metric, hence it has a limit set. The description of this limit set, however, remains an open question.

Appendix. Denote by $(x_k(\cdot), y_k(\cdot))$ the solution of the equation

$$\begin{aligned} \dot{x} &= A_1(t)x + A_2(t)y + \varphi_0(t) + \Delta\varphi_k(t), \qquad x(0) = v_k, \\ \lambda_k \dot{y} &= A_3(t)x + A_4(t)y + \psi_0(t) + \Delta\psi_k(t), \qquad y(0) = w_k, \end{aligned} \tag{27}$$

where $k = 1, 2, \cdots, \lim_{k\to+\infty} \lambda_k = 0$, $\varphi_0(\cdot) \in L^2(R^m, (0, T))$, $\psi_0(\cdot) \in L^2(R^n, (0, T))$, and $v_k \in R^m$, $w_k \in R^n$; $\{\Delta\varphi_k(\cdot)\}$, $\{\Delta\psi_k(\cdot)\}$ are given sequences of functions. The solution of

$$\begin{aligned} \dot{x} &= A_1(t)x + A_2(t)y(t) + \varphi_0(t), \qquad x(0) = v_0, \\ 0 &= A_3(t)x(t) + A_4(t)y(t) + \psi_0(t) \end{aligned} \tag{28}$$

will be denoted by $(x_0(\cdot), y_0(\cdot))$. We assume that the matrices $A_i(t)$, $i = 1, \cdots, 4$, satisfy (A1).

LEMMA A. (i) *Let* $\lim_{k\to+\infty} v_k = v_0$, $w_k = \omega_k/\lambda_k$, $\lim_{k\to+\infty} \omega_k = \omega_0$, *and the sequences* $\{\Delta\varphi_k(\cdot)\}$ *and* $\{\Delta\psi_k(\cdot)\}$ *are weakly convergent to zero functions in* $L^2(R^m, (0, T))$ *and* $L^2(R^n, (0, T))$ *respectively. Let* $x_0(\cdot)$ *be determined by* (28) *with the initial condition* $x(0) = v_0 - A_2(0)_4^{-1}(0)\omega_0$. *Then the sequence* $\{x_k(\cdot)\}$ *is uniformly bounded on* $[0, T]$ *and for every* $\theta \in (0, T)$

$$\lim_{k\to+\infty} \max_{\theta\leq t\leq T} |x_k(t) - x_0(t)| = 0. \tag{29}$$

(ii) *If, additionally,* $\omega_0 = 0$, *then*

$$\lim_{k\to+\infty} \|x_k(\cdot) - x_0(\cdot)\|_C = 0. \tag{30}$$

(iii) *Let* $\lim_{k\to+\infty} \sqrt{\lambda_k}\, w_k = 0$ *and all the above conditions hold. Then the sequence* $\{y_k(\cdot)\}$ *is* L^2 *weakly convergent to* $y_0(\cdot)$.

(iv) *If all the above conditions are satisfied and, additionally,* $\psi_0(\cdot) \in C(R^n, [0, T])$, *the sequence* $\{w_k\}$ *is bounded, and for every* $\theta_1 \in (0, T)$

$$\lim_{k\to+\infty} \max_{0\leq t\leq\theta_1} |\Delta\psi_k(t)| = 0,$$

then for every $\theta \in (0, T/2)$

$$\lim_{k\to+\infty} \max_{\theta\leq t\leq T-\theta} |y_k(t) - y_0(t)| = 0. \tag{31}$$

Proof. Denoting $\Delta x_k(\cdot) = x_k(\cdot) - x_0(\cdot)$, $\Delta y_k(\cdot) = y_k(\cdot) - y_0(\cdot)$ we have

$$\begin{aligned} \Delta x_k(t) &= v_k - v_0 + A_2(0)A_4^{-1}(0)\omega_0 \\ &\quad + \int_0^t (A_1(\tau)\Delta x_k(\tau) + A_2(\tau)\Delta y_k(\tau) + \Delta\varphi_k(\tau))\, d\tau, \end{aligned} \tag{32}$$

$$\begin{aligned} \Delta y_k(t) &= \phi(t, 0, \lambda_k)w_k \\ &\quad + \frac{1}{\lambda_k}\int_0^t \phi(t, \tau, \lambda_k)(A_3(\tau)\Delta x_k(\tau) + \Delta\psi_k(\tau) - A_4(\tau)y_0(\tau))\, d\tau - y_0(t). \end{aligned} \tag{33}$$

In the sequel we use the following standard result: if $p(\cdot) \in L^1(R^1, (0, T))$, $q(\cdot) \in L^2(R^1, (0, T))$ and

$$r(t) = \int_0^t p(t - \tau)q(\tau)\, d\tau$$

then

$$\|r(\cdot)\|_{L^2} \leqq \|p(\cdot)\|_{L^1}\|q(\cdot)\|_{L^2}. \tag{34}$$

Let δ be an arbitrary positive number. Choose a function $y^\delta(\cdot) \in C^1(R^n, [0, T])$ such that $\|y^\delta(\cdot) - y_0(\cdot)\|_{L^2} < \delta$. In view of (13), for every $t \in [0, T]$

$$\begin{aligned} &\left| \int_0^t \frac{\partial}{\partial \tau} \phi(t, \tau, \lambda_k) y^\delta(\tau)\, d\tau - y_0(t) \right| \\ &\leqq c\left(|y^\delta(t) - y^0(t)| + |y^\delta(0)| \exp\left(-\sigma \frac{t}{\lambda_k}\right) + \lambda_k \|\dot{y}^\delta(\cdot)\|_c \right), \end{aligned} \tag{35}$$

where c is a generic constant. Let

$$\bar{y}_k(t) = \frac{1}{\lambda_k} \int_0^t \phi(t, \tau, \lambda_k) A_4(\tau) y_0(\tau)\, d\tau. \tag{36}$$

Then, since δ can be arbitrary small, by (34) and (35), integrating by parts we obtain

$$\lim_{k \to +\infty} \|\bar{y}_k(\cdot) - y_0(\cdot)\|_{L^2} = 0. \tag{37}$$

For an arbitrary but fixed $\varepsilon > 0$ one can choose matrices $A_2^\varepsilon(t)$ and $A^\varepsilon(t)$, whose components are C^1, such that $\|A_2(\cdot) - A_2^\varepsilon(\cdot)\|_c < \varepsilon$ and $\|A_4^{-1}(\cdot) - A^\varepsilon(\cdot)\|_c < \varepsilon$. From

$$\begin{aligned} &\frac{1}{\lambda_k} \int_0^t A_2(\tau) \phi(\tau, 0, \lambda_k)\, d\tau \\ &\quad = \int_0^t A_2(\tau) A_4^{-1}(\tau) \frac{\partial}{\partial \tau} \phi(\tau, 0, \lambda_k)\, d\tau \\ &\quad = \frac{1}{\lambda_k} \int_0^t (A_2(\tau) A_4^{-1}(\tau) - A_2^\varepsilon(\tau) A^\varepsilon(\tau)) A_4(\tau) \phi(\tau, 0, \lambda_k)\, d\tau \\ &\qquad + \int_0^t A_2^\varepsilon(\tau) A^\varepsilon(\tau) \frac{\partial}{\partial \tau} \phi(\tau, 0, \lambda_k)\, d\tau, \end{aligned}$$

integrating by parts and taking advantage of (13) we get

$$\begin{aligned} &\left| \frac{1}{\lambda_k} \int_0^t A_2(\tau) \phi(\tau, 0, \lambda_k) \omega_k\, d\tau + A_2(0) A_4^{-1}(0) \omega_0 \right| \\ &\qquad \leqq c\left(\varepsilon + |\omega_k| \exp\left(-\sigma \frac{t}{\lambda_k}\right) + c_1(\varepsilon) \lambda_k + |\omega_k - \omega_0| \right), \end{aligned} \tag{38}$$

where $c_1(\varepsilon) \geqq \|d/dt(A_2^\varepsilon A^\varepsilon)(\cdot)\|_c$. Denote $\xi_k(t) = A_3(t) \Delta x_k(t) + \Delta \psi_k(t)$ and

$$\eta_k(t) = \frac{1}{\lambda_k} \int_0^t \phi(t, \tau, \lambda_k) \xi_k(\tau)\, d\tau.$$

Applying (34) we have

$$\|\eta_k(\cdot)\|_{L^2} \leqq \frac{\sigma_0}{\sigma} \|\xi_k(\cdot)\|_{L^2}. \tag{39}$$

Using (13), (34), (39), the Hoelder inequality, and integrating by parts we obtain

$$
\begin{aligned}
&\left|\int_0^t A_2(\tau)\eta_k(\tau)\,d\tau\right| \\
&\quad\leqq \int_0^t |A_2(\tau)-A_2^\varepsilon(\tau)|\,|\eta_k(\tau)|\,d\tau + \left|\int_0^t A_2^\varepsilon(\tau)A^\varepsilon(\tau)A_4(\tau)\eta_k(\tau)\,d\tau\right| \\
&\qquad + \int_0^t |A_2^\varepsilon(\tau)(A_4^{-1}(\tau)-A^\varepsilon(\tau))A_4(\tau)\eta_k(\tau)|\,d\tau \\
&\quad\leqq c\varepsilon\|\xi_k(\cdot)\|_{L^2} + \left|\int_0^t A_2^\varepsilon(\tau)A^\varepsilon(\tau)\frac{\partial}{\partial\tau}\int_0^\tau \phi(\tau,s,\lambda_k)\xi_k(s)\,ds\,d\tau\right| \\
&\qquad + \left|\int_0^t A_2^\varepsilon(\tau)A^\varepsilon(\tau)\xi_k(\tau)\,d\tau\right| \qquad (40)\\
&\quad\leqq c\varepsilon\|\xi_k(\cdot)\|_{L^2} + \lambda_k\left(|A_2^\varepsilon(t)A^\varepsilon(t)\eta_k(t)| + \left|\int_0^t \frac{d}{d\tau}(A_2^\varepsilon(\tau)A^\varepsilon(\tau))\eta_k(\tau)\,d\tau\right|\right) \\
&\qquad + \left|\int_0^t A_2^\varepsilon(\tau)A^\varepsilon(\tau)\xi_k(\tau)\,d\tau\right| \\
&\quad\leqq c\left((\varepsilon+\sqrt{\lambda_k}+c_1(\varepsilon)\lambda_k)\|\xi_k(\cdot)\|_{L^2} + \int_0^t |\Delta x_k(\tau)|\,d\tau\right) \\
&\qquad + \left|\int_0^t A_2^\varepsilon(\tau)A^\varepsilon(\tau)\Delta\psi_k(\tau)\,d\tau\right|.
\end{aligned}
$$

Taking into account (32), (33), (38), (39) and (40), we get

$$
\begin{aligned}
|\Delta x_k(t)| \leqq |v_k - v_0| + c\Big(&\varepsilon + c_1(\varepsilon)\lambda_k + |\omega_k|\exp\left(-\sigma\frac{t}{\lambda_k}\right) + |\omega_k-\omega_0| \\
&+ (\varepsilon+\sqrt{\lambda_k}+c_1(\varepsilon)\lambda_k)(\|\Delta x_k(\cdot)\|_{L^2} + \|\Delta\psi_k(\cdot)\|_{L^2}) \\
&+ \|\bar{y}_k(\cdot)-y_0(\cdot)\|_{L^2} + \int_0^t |\Delta x_k(\tau)|\,d\tau\Big) \qquad (41)\\
&+ \left|\int_0^t \Delta\varphi_k(\tau)\,d\tau\right| + \left|\int_0^t A_2^\varepsilon(\tau)A^\varepsilon(\tau)\Delta\psi_k(\tau)\,d\tau\right|.
\end{aligned}
$$

Let us recall that if the sequence $\{z_k(\cdot)\}$ is L^2 weakly convergent to zero, then

$$\lim_{k\to+\infty}\max_{0\leqq t\leqq T}\left|\int_0^t z_k(\tau)\,d\tau\right| = 0,$$

and the sequence of the norms is bounded. Applying the Gronwall lemma to (41) we conclude that

$$\|\Delta x_k(\cdot)\|_{L^2} \leqq T\|\Delta x_k(\cdot)\|_c \leqq c(\varepsilon+\sqrt{\lambda_k}+c_1(\varepsilon)\lambda_k)\|\Delta x_k(\cdot)\|_{L^2} + c.$$

Choosing $\varepsilon < 1/c$ and tending to zero with λ_k we get

$$\lim_{k\to+\infty}\sup\|\Delta x_k(\cdot)\|_{L^2} < +\infty. \qquad (42)$$

Using this result in (41) we obtain finally

$$|\Delta x_k(t)| \leqq c\left(\varepsilon + \int_0^t |\Delta x_k(\tau)|\, d\tau + |w_k| \exp\left(-\sigma \frac{t}{\lambda_k}\right) + \delta_k\right), \tag{43}$$

where $\lim_{k\to+\infty} \delta_k = 0$ uniformly in $[0, T]$. From the Gronwall lemma we obtain that for every $\theta \in (0, T)$

$$\lim_{k\to+\infty} \max_{\theta \leqq t \leqq T} |\Delta x_k(t)| \leqq c\varepsilon.$$

Since ε can be arbitrary small and $\Delta x_k(\cdot)$ does not depend on ε, this relation implies (29).

The relation (30) follows immediately from (43).

If $\lim_{k\to+\infty} \sqrt{\lambda_k}\, w_k = 0$, then from (13), (29), (33), (34) and (38)

$$\lim_{k\to+\infty} \sup \|\Delta y_k(\cdot)\|_{L^2} < +\infty.$$

Hence, if we show that for every $t \in [0, T]$

$$\lim_{k\to+\infty} \int_0^t \Delta y_k(\tau)\, d\tau = 0, \tag{44}$$

then the statement of (iii) will be proved. Using a sequence of inequalities similar to (40) we have

$$\left|\frac{1}{\lambda_k} \int_0^t \int_0^\tau \phi(\tau, s, \lambda_k) \Delta\psi_k(s)\, ds\, d\tau\right| \leqq c(\varepsilon + \sqrt{\lambda_k} + c_1(\varepsilon)\lambda_k) \|\Delta\psi_k(\cdot)\|_{L^2} + \left|\int_0^t A^\varepsilon(\tau) \Delta\psi_k(\tau)\, d\tau\right|.$$

In view of this relation, (30) and (37) we obtain (44).

Next, let the conditions in (iv) hold. Since $y_0(\cdot) \in C(R^n, [0, T])$, one can choose $y^\delta(\cdot) \in C^1(R^n, [0, T])$ such that $\|y^\delta(\cdot) - y_0(\cdot)\|_c < \delta$. Then

$$\begin{aligned} |\bar{y}_k(t) - y_0(t)| &\leqq \left|\frac{1}{\lambda_k} \int_0^t \phi(t, \tau, \lambda_k) A_4(\tau)(y_0(\tau) - y^\delta(\tau))\, d\tau\right| \\ &\quad + \left|\int_0^t \frac{\partial}{\partial\tau} \phi(t, \tau, \lambda_k) y^\delta(\tau)\, d\tau - y_0(t)\right| \\ &\leqq c\left(\delta + |y^\delta(0)| \exp\left(-\sigma \frac{t}{\lambda_k}\right) + \lambda_k \|\dot{y}^\delta(\cdot)\|_c\right). \end{aligned} \tag{45}$$

By substituting (45) into (33) we obtain (31). The proof is complete.

REFERENCES

[1] R. J. AUMANN, *Integrals of set valued functions*, J. Math. Anal. Appl., 12 (1965), pp. 1–12.

[2] P. BINDING, *Singularly perturbed optimal control systems. I: Convergence*, this Journal, 14 (1976), pp. 591–612.

[3] F. CLARKE, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 155–174.

[4] M. G. DMITRIEV, *The continuity of the solution of the Mayer problem with respect to singular perturbations*, J. Vych. Mat. i Mat. Fiz., 123 (1972), pp. 788–791. (In Russian.)

[5] A. L. DONTCHEV, *On the order reduction of optimal control systems*, Proc. Conference on Model Validity and Credibility, Oct. 1980, IIASA.

[6] T. R. GIČEV AND A. L. DONTCHEV, *Convergence of the solutions of the singularly perturbed time-optimal problem*, Prikl. Matem. i Mech., 43 (1979), pp. 466–474. (In Russian.)

[7] V. J. GLIZER AND M. G. DMITRIEV, *On the continuity of the solution of the analytic regulator design problem with singular perturbation*, Prikl. Matem. i Mech., 41B (1977), pp. 573–576. (In Russian.)

[8] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, Nauka, Moscow, 1974. (In Russian.)

[9] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE AND E. F. MISHCHENKO, *Mathematical Theory of Optimal Processes*, Nauka, Moscow, 1969. (In Russian.)

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

0363-0129/83/2104-0005 $01.25/0

THE CONSTRUCTION OF THE SOLUTION OF AN OPTIMAL CONTROL PROBLEM DESCRIBED BY A VOLTERRA INTEGRAL EQUATION*

GUSTAF GRIPENBERG†

Abstract. A constructive method is developed for finding the nonnegative function u so that $\int_0^\infty u(s)\,ds$ is as small as possible when $\liminf_{t\to\infty} y(t)>0$, where y satisfies the equation $y(t)=u(t)+\int_0^t a(t-s)g(y(s))\,ds$, $t\geqq 0$.

Key words. optimal control, integral equation

1. Introduction. The purpose of this paper is to study the following problem: Find a nonnegative function u on $\mathbb{R}^+=[0,\infty)$ so that $\int_0^\infty u(t)\,dt$ is minimized under the condition that $\liminf_{t\to\infty} y(t)\geqq\inf\{\omega\in\mathbb{R}^+ | g(\omega)\int_0^\infty a(t)\,dt>\omega\}$ when

$$y(t)=u(t)+\int_0^t a(t-s)g(y(s))\,ds,\qquad t\in\mathbb{R}^+. \tag{1.1}$$

We assume that a is integrable, nonnegative and nonincreasing, $g(\omega)=0$ on $[0,\omega_0]$ and that g is nonnegative and concave on $[\omega_0,\infty)$.

This problem arises for example from the following kind of "investment" or "growth" model: Let $y(t)$ be the flow of available "resources" (of some kind) and let $g(y(t))$ be the "investments". Due to diminishing returns it is to be expected that g is concave and that if there are not enough "resources" available, then no "investments" are made. The available "resources" are determined by previous "investments" and "exterior inputs" $u(t)$ as in (1.1). The problem is to minimize the total inputs $\int_0^\infty u(t)\,dt$ so that self-sustained growth is achieved, i.e., the returns on the "investments" suffice for "consumption" and "re-investments".

Another equation that can be put in form (1.1) is the following:

$$v(t)=g_0\Big(u(t)+\int_0^t a(t-s)g_1(v(s)\,ds)\Big),\qquad t\in\mathbb{R}^+.$$

(To get (1.1) take $y(t)=u(t)+\int_0^t a(t-s)g_1(v(s))\,ds$ and $g=g_0\circ g_1$.) This equation arises from a model where $v(t)$ is the "output" of the "economy", $g_1(v(t))$ represents the "investments", $\int_0^t a(t-s)g_1(v(s))\,ds$ is the "capital stock" due to previous "investments" and $u(t)$ is "capital stock" derived from external sources and one wants to minimize the use of this external "capital stock".

These models, as they are formulated here, are clearly quite simple, and hence somewhat unrealistic, ones, but the main point of this paper is to show how one can, at least in principle, calculate the optimal solution. The proof that such a solution exists will be quite a small part of the argument.

For more general results on the optimization of functional equations, see e.g. [1], [3], [4] and [8]. Here we will use only one relatively straightforward idea from these more general results where the main emphasis is on how to find necessary conditions for optimality. Since we will here try to find the solution to the rather specific problem at hand, most of the analysis is closely tied to the assumptions made concerning (1.1).

* Received by the editors August 19, 1981, and in final revised form July 25, 1982.
† Institute of Mathematics, Helsinki University of Technology, SF-02150 Espoo 15, Finland.

2. Statement of results. We will establish the following result concerning the problem formulated in the introduction.

THEOREM. *Assume that*

(2.1) $a:\mathbb{R}^+\to\mathbb{R}^+$ *is nonincreasing and* $\int_0^\infty a(s)\,ds=1$,

(2.2) $g:\mathbb{R}^+\to\mathbb{R}^+$ *is continuous,* $g(\omega)=0, \omega\in[0,\omega_0], \omega_0>0$ *and* g *is twice continuously differentiable nondecreasing and concave on* $[\omega_0,\infty)$,

(2.3) $g(\omega_j)=\omega_j$, $j=1,2$ *where* $\omega_0<\omega_1<\omega_2$ *and* $g''(\omega)<0$ *on* $[\omega_0,\omega_2]$.

Then there exists a unique, a.e., function $u^*\in U$ *such that* $\int_0^\infty u^*(t)\,dt=\inf\{\int_0^\infty u(t)\,dt|u\in U\}$ *where*

(2.4) $U=\{u:\mathbb{R}^+\to\mathbb{R}^+|u$ *is measurable,* $\liminf_{t\to\infty} y(t)\geqq\omega_1$ *if* y *is the solution of* (1.1) *and* $\inf\{t|y(t)-u(t)>0\}=0\}$.

The function u^* *is continuous and if* y^* *is the corresponding optimal solution of* (1.1), *then*

(2.5) $y^*(t)>\omega_0$, $u^*(t)=0$ *if* $g'(y^*(t))<1$, *a.e.,* $t\in\mathbb{R}^+$ *and* $\lim_{t\to\infty} y^*(t)=\omega_1$.

Moreover, the functions u^* *and* y^* *can be found as the uniform limits on compact subsets of* $\mathbb{R}^+$ *of functions that are constructed using iteration procedures involving evaluations of integrals and functions.*

Note that the theorem above does not make the claim that the procedure for finding u^* and y^* would be computationally very efficient. It may very well be the case that much simpler approximation procedures will give the desired result.

The fact that the kernel a is nonincreasing is needed because the function g is not concave on the whole of $\mathbb{R}^+$, but this assumption also has the desirable consequence that if $\lim_{t\to\infty} u(t)=0$, then $\lim_{t\to\infty} y(t)$ exists provided y is bounded, cf. [5]. It will also be shown in the proof that if $u(t)\geqq u^*(t)$ with strict inequality on some set of positive measure, then $\liminf_{t\to\infty} y(t)=\omega_2$.

It is easy to see that if the last condition in the definition of the set U of admissible controls is dropped, then the optimal solution is no longer unique.

The main idea of the proof is, of course, to replace the optimization problem above by other problems that approximate the original one in the right sense and that are such that they can be solved. One difficulty is of course the infinite horizon of the problem and the solution is to take a finite interval $[0,T]$ and letting $T\to\infty$. But when we work on $[0,T]$ the condition $\liminf_{t\to\infty} y(t)\geqq\omega_1$ is meaningless and instead we minimize $\mu\int_0^T u(t)\,dt-(y(T)-u(T))$, where μ is a certain carefully chosen constant. But in order for this approach to work we must replace the function g by another function $g(\delta,\cdot)$ that equals g on $[\omega_1,\omega_2]$, and is concave on $\mathbb{R}$ and satisfies $\sup_{\omega\in\mathbb{R}} g(\delta,\omega)\to\omega_2$ as $\delta\to 0$.

3. Proof of the theorem. First we observe that if we define x by $x(t)=y(t)-u(t)$, then (1.1) becomes

$$x(t)=\int_0^t a(t-s)g(x(s)+u(s))\,ds,\qquad t\in\mathbb{R}^+. \tag{3.1}$$

If $\lim_{t\to\infty} x(t)=x_0$ and $u(t)\geqq 0$, $\int_0^\infty u(t)\,dt<\infty$, then $\liminf_{t\to\infty} y(t)=x_0$. On the other hand, under the same assumptions on u, it follows from (2.2) and (2.3) that $g(x(\cdot)+u(\cdot))-g(x(\cdot))\in L^1(\mathbb{R}^+)$ for any nonnegative function x and therefore it is possible to show, see, e.g., [6], that (3.1) has a unique, nonnegative and continuous solution

x. But it also follows from results in [5] that $\lim_{t\to\infty} x(t)$ exists and is equal to $\liminf_{t\to\infty} y(t)$. Thus we see that we can just as well consider (3.1) and the problem of minimizing $\int_0^\infty u(s)\,ds$ under the condition that the solution x of (3.1) (which we denote: $x = G(u)$), satisfies $\lim_{t\to\infty} x(t) \geqq \omega_1$.

Next we construct some new functions that will be used instead of the function g in (3.1). There are several other functions that would serve the same purposes. Let $\delta \geqq 0$ and define

$$
g(\delta,\omega)=\begin{cases} g''(\omega_0+)(1+\omega_0-\omega)\log(1+\omega_0-\omega) \\ \qquad +(g'(\omega_0+)+g''(\omega_0+))(\omega-\omega_0), & \omega\leqq\omega_0, \\ g(\omega), & \omega_0<\omega\leqq\omega_2, \\ g(\omega_2)+(g'(\omega_2)\delta^{-2}+g''(\omega_2)(3\delta)^{-1})((\omega-\omega_2-\delta)^3+\delta^3) \\ \qquad +(g'(\omega_2)2^{-1}\delta^{-3}+g''(\omega_2)(2\delta)^{-2})((\omega-\omega_2-\delta)^4-\delta^4), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \omega_2<\omega\leqq\omega_2+\delta, \\ g(\omega_2)+g'(\omega_2)2^{-1}\delta+g''(\omega_2)(12)^{-1}\delta^2, & \omega>\omega_2+\delta. \end{cases} \tag{3.2}
$$

It is easy to check that if $\delta_0>0$ is such that $g'(\omega_2)>-\delta_0 g''(\omega_2)$, then $g(\delta,\cdot)$ is nondecreasing and concave on $\mathbb{R}^+$ for $\delta\in[0,\delta_0)$ and if $\delta>0$, then $g(\delta,\cdot)$ is twice continuously differentiable. Observe that by (2.2) and (2.3) $g'(\omega_2)>0$ because otherwise $g'(\omega)<0$ for some $\omega>\omega_2$ as $g''(\omega_2)<0$. Note also that $g(\delta,\omega)\leqq\omega$ when $\omega\geqq\omega_2$, since $g(\delta,\cdot)$ is concave and $g(\delta,\omega_j)=\omega_j$, $j=1,2$ and hence $g(\delta,\omega)\leqq\omega_2+\delta$, $\delta\in[0,\delta_0)$.

It follows from standard results that if u is nonnegative and measurable on $\mathbb{R}^+$ and $\delta\in[0,\delta_0)$, then the equation

$$
x(t)=\int_0^t a(t-s)g(\delta,x(s)+u(s))\,ds, \qquad t\in\mathbb{R}^+ \tag{3.3}
$$

has a unique solution on $\mathbb{R}^+$ and there exists a continuous nondecreasing function $c_0:\mathbb{R}^+\to\mathbb{R}^+$, independent of u, such that

$$
-c_0(t)\leqq x(t)\leqq\omega_2+\delta, \qquad t\in\mathbb{R}^+. \tag{3.4}
$$

To see this, one uses the facts that $g(\delta,\omega)\int_0^\infty a(t)\,dt\leqq\omega_2+\delta$ and that $x(t)\geqq z(t)$ where z satisfies $z'(t)=a(0)g(\delta,z(t))$, $z(0)=0$. We denote the solution of (3.3) by $x=G(\delta,u)$.

For arbitrary $T>0$, $\mu>0$ and $\delta\in[0,\delta_0)$ we consider the following:

Problem (T,δ). *Find a nonnegative measurable function* u *on* $[0,T]$ *such that* $J(T,\delta,u)\overset{\text{def}}{=}\mu\int_0^T u(t)\,dt-x(T)$ *is minimized when* $x=G(\delta,u)$.

(Later we will choose μ so that (3.4) is satisfied and therefore we will not write out the dependence on μ below.)

First we prove that this problem has a solution. This proof is not "constructive", but we need it below.

LEMMA 1. *For each* $T>0$ *and* $\delta\in[0,\delta_0)$ *Problem* (T,δ) *has a unique, a.e., solution* $u_*(T,\delta,\cdot)$.

Proof. Fix $T>0$ and $\delta\in[0,\delta_0)$. Since $J(T,\delta,u)$ is bounded from below, see (3.4), it follows that there exists a sequence $\{u_j\}_{j=1}^\infty$ of nonnegative functions such that $J(T,\delta,u_j)\to\inf_{u\geqq0}J(T,\delta,u)$, as $j\to\infty$. We let $x_j=G(\delta,u_j)$. Since $g(\delta,\cdot)$ is a constant on $(\omega_2+\delta,\infty)$ it follows from (3.4) that we may assume that

$$
u_n(t)\leqq\omega_2+\delta+c_0(t), \qquad t\in[0,T]. \tag{3.5}
$$

Since the functions $g(\delta, x_j(\cdot)+u_j(\cdot))$ are uniformly bounded by (3.2) and (3.4), it follows from the integrability of a and (3.3) that the functions x_n are uniformly bounded and equicontinuous. Hence we conclude, using also the fact that the integrals $\int_0^T u_j(t)^2\,dt$ are uniformly bounded, see (3.5), that there exists a continuous function y and a bounded, nonnegative, measurable function v such that for some subsequences, also denoted by $\{x_j\}$ and $\{u_j\}$,

$$\begin{aligned} &x_j \to y \quad \text{uniformly on } [0, T], \\ &u_j \to v \quad \text{weakly in } L^2(0, T) \text{ as } j \to \infty, \end{aligned} \tag{3.6}$$

and

$$\mu \int_0^T v(t)\,dt - y(T) = \inf_{u \geqq 0} J(T, \delta, u). \tag{3.7}$$

It follows from (3.6) and [7, Thm. 3.13] that there exist numbers α_{ij} such that

$$\alpha_{ij} \geqq 0, \quad \sum_{i=j}^{I(j)} \alpha_{ij} = 1, \quad \sum_{i=j}^{I(j)} \alpha_{ij} u_i \to v \quad \text{in } L^2(0, T) \quad \text{as } j \to \infty. \tag{3.8}$$

Using the concavity of $g(\delta, \cdot)$ and the fact that $a(t) \geqq 0$ we get from (3.3) and (3.8)

$$\sum_{i=j}^{I(j)} \alpha_{ij} x_i(t) \leqq \int_0^t a(t-s) g\left(\delta, \sum_{i=j}^{I(j)} \alpha_{ij}(x_i(s)+u_i(s))\right) ds$$

and if we let $j \to \infty$, then we obtain from (3.6) and (3.8), since $g(\delta, \cdot)$ is locally Lipschitz-continuous, that

$$y(t) \leqq \int_0^t a(t-s) g(\delta, y(s)+v(s))\,ds.$$

As $g(\delta, \cdot)$ is nondecreasing and $a(t) \geqq 0$, this inequality implies that $x = G(\delta, v)$ satisfies $x(t) \geqq y(t)$, $t \in [0, T]$. Hence (3.7) shows that $J(T, \delta, v) = \inf_{u \geqq 0} J(T, \delta, u)$.

Next we establish the uniqueness of this optimal solution. Suppose that we have two functions u_1 and u_2 that give the same minimal value to $J(T, \delta, \cdot)$ and differ on a set of positive measure. Therefore the solutions $x_j = G(\delta, u_j)$ $j = 1, 2$ cannot be identical and $x_1 + u_1$ and $x_2 + u_2$ must also differ on a set of positive measure. Since we obtain the minimum, we must have $x_j(t) + u_j(t) \leqq \omega_2 + \delta$, hence the function $g(\delta, \cdot)$ is strictly increasing and strictly concave on the interval under consideration. Let $u(t) = (u_1(t) + u_2(t))/2$, and $\bar{x}(t) = (x_1(t) + x_2(t))/2$, $t \in [0, T]$. It follows that $(g(\delta, x_1(t) + u_1(t)) + g(\delta, x_2(t) + u_2(t)))/2 \leqq g(\delta, \bar{x}(t) + u(t))$, i.e., $\bar{x}(t) \leqq \int_0^t a(t-s) g(\delta, \bar{x}(s) + u(s))\,ds$ with strict inequality on a set of positive measure. Let $x = G(\delta, u)$. By the previous result and the fact that $g(\delta, \cdot)$ is nondecreasing it follows that $x(t) \geqq \bar{x}(t)$ with strict inequality on a nonempty open set. If $x(T) > \bar{x}(T)$, then we immediately get a contradiction. Otherwise we let $t_0 = \sup\{t \in [0, T] \mid x(t) > \bar{x}(t)\}$, so that $x(t_0) = \bar{x}(t_0)$. Since $g(\delta, \cdot)$ is strictly increasing on $(-\infty, \omega_2 + \delta)$ and a is nonincreasing and $\neq 0$, there exists a number $\tau > 0$ such that $\bar{x}(t) + u(t) = \omega_2 + \delta$ on $(t_0 - \tau, t_0)$. But then $x(t) + u(t) > \omega_2 + \delta$ on a set of positive measure but this is by (3.4) impossible since u is also optimal. This contradiction shows that the optimal solution $u_*(T, \delta, \cdot)$ is unique and the proof of Lemma 1 is completed.

We fix for the moment $T > 0$ and $\delta \in (0, \delta_0)$ and we write $x_*(\cdot) = x_*(T, \delta, \cdot) = G(\delta, u_*(T, \delta, \cdot))$, $u_*(\cdot) = u_*(T, \delta, \cdot)$. We will derive an equation for u_* and x_* that we will then solve. We write $g'(\delta, \omega) = d/d\omega\, g(\delta, \omega)$.

LEMMA 2. *If $\delta \in (0, \delta_0)$ and $T>0$, then there exists a unique solution $q_*(\cdot) = q_*(T, \delta, \cdot)$ and $x_*(T, \delta, \cdot)$ of the equations*

$$q(t) = a(T-t) + \int_t^T a(s-t) \min \{\mu, g'(\delta, x(s))q(s)\}\, ds, \qquad t \in [0, T] \tag{3.9}$$

and

$$x(t) = \int_0^t a(t-s) g(\delta, \max \{x(s), Q(\delta, q(s))\})\, ds, \qquad t \in [0, T], \tag{3.10}$$

where $Q(\delta, \cdot)$ is defined by

$$\omega g'(\delta, Q(\delta, \omega)) = \mu, \quad \omega > 0, \quad Q(\delta, \omega) = -\infty, \quad \omega \leqq 0. \tag{3.11}$$

Moreover,

$$u_*(T, \delta, t) = \max \{0, Q(\delta, q_*(T, \delta, t)) - x_*(T, \delta, t)\}, \qquad t \in [0, T]. \tag{3.12}$$

Proof. Let $t_0 \in (0, T)$ be a Lebesgue point for the functions $u_*(\cdot)$ and $g(\delta, x_*(\cdot) + u_*(\cdot))$. Let $v \geqq 0$ and $\varepsilon > 0$ be arbitrary numbers and define

$$u(\varepsilon, t) = \begin{cases} u_*(t), & t \in [0, T] \backslash [t_0, t_0 + \varepsilon], \\ v, & t \in [t_0, t_0 + \varepsilon]. \end{cases} \tag{3.13}$$

Since u_* is optimal we have

$$\liminf_{\varepsilon \to 0+} \frac{J(T, \delta, u(\varepsilon, \cdot)) - J(T, \delta, u_*)}{\varepsilon} \geqq 0. \tag{3.14}$$

Because t_0 is a Lebesgue point of u_* we have by (3.13)

$$\lim_{\varepsilon \to 0+} \int_0^T \frac{(u(\varepsilon, t) - u_*(t))\, dt}{\varepsilon} = v - u_*(t_0). \tag{3.15}$$

Let $x(\varepsilon, \cdot) = G(\delta, u(\varepsilon, \cdot))$. It is straightforward to prove that $x(\varepsilon, \cdot) \to x_*$ uniformly on $[0, T]$ as $\varepsilon \to 0+$ and that $\lim_{\varepsilon \to 0+} (x(\varepsilon, t) - x_*(t))/\varepsilon = y(t)$ exists and satisfies the equation

$$\begin{aligned} y(t) &= \int_0^t a(t-s) g'(\delta, x_*(s) + u_*(s)) y(s)\, ds \\ &\quad + a(t-t_0)(g(\delta, x_*(t_0) + v) - g(\delta, x_*(t_0) + u_*(t_0))). \end{aligned} \tag{3.16}$$

Here we used the facts that t_0 is a Lebesgue point and that we may modify a on a denumerable set so that a is left-continuous and $a(t) = 0$, $t \leqq 0$. Let $h(t) = g'(\delta, x_*(t) + u_*(t))$ and define $R(t, s)$ to be the solution of the equation, see [2],

$$\begin{aligned} R(t, s) &= a(t-s)h(s) + \int_s^t a(t-u)h(u)R(u, s)\, du \\ &= a(t-s)h(s) + \int_s^t R(t, u)a(u-s)h(s)\, du, \qquad s \leqq t. \end{aligned}$$

This equation shows that $R(t, s) = 0$ if $h(s) = 0$ and hence $r(t, s) = R(t, s)/h(s)$ satisfies

$$\begin{aligned} r(t, s) &= a(t-s) + \int_s^t a(t-u)h(u)r(u, s)\, du \\ &= a(t-s) + \int_s^t r(t, u)h(u)a(u-s)\, du, \qquad s \leqq t. \end{aligned} \tag{3.17}$$

Using the first equality of (3.17) in (3.16) we obtain

$$y(T)=q_*(t_0)(g(\delta, x_*(t_0)+v)-g(\delta, x_*(t_0)+u_*(t_0))) \tag{3.18}$$

where $q_*(t)=r(T, t)$. From the second equality in (3.17) and our definition of h we get

$$q_*(t)=a(T-t)+\int_t^T a(s-t)g'(\delta, x_*(s)+u_*(s))q_*(s)\,ds. \tag{3.19}$$

Now we are able to conclude from (3.14), (3.15) and (3.18), our definitions for J and y and our choices of v and t_0, that

$$\begin{aligned}&\min_{v\in\mathbb{R}^+}\{\mu v-q_*(t)g(\delta, x_*(t)+v)\}\\ &\qquad=\mu u_*(t)-q_*(t)g(\delta, x_*(t)+u_*(t)) \quad \text{for a.e. } t\in[0, T].\end{aligned} \tag{3.20}$$

From (3.19), (3.20) and the fact that $g(\delta, \cdot)$ is concave, we conclude that q_* and x_* satisfy (3.9) and (3.10) and that (3.12) holds.

It remains for us to prove that the solution q_*, x_* is unique. Suppose that this is not the case but that there exists another solution q_0 and x_0. If we define u_0 by $u_0(t)=\max\{0, Q(\delta, q_0(t))-x_0(t)\}$, then $x_0=G(\delta, u_0)$ and u_0 and u_* cannot be identical a.e., since otherwise we could deduce from standard uniqueness results for Volterra equations, see e.g. [6], first that $x_0\equiv x_*$ and then that $q_0\equiv q_*$. But then Lemma 1 implies that

$$J(T, \delta, u_0)>J(T, \delta, u_*). \tag{3.21}$$

Let $u_\alpha=(1-\alpha)u_0+\alpha u_*$ and $x_\alpha=G(\delta, u_\alpha)$. By the same argument that was used in the proof of Lemma 1, we see that $x_\alpha(T)\geqq(1-\alpha)x_0(T)+\alpha x_*(T)$. Hence

$$\limsup_{\alpha\to 0+}(J(T, \delta, u_\alpha)-J(T, \delta, u_0))/\alpha\leqq J(T, \delta, u_*)-J(T, \delta, u_0). \tag{3.22}$$

On the other hand it is easy to see that $\lim_{\alpha\to 0+}(x_\alpha(t)-x_0(t))/\alpha=v(t)$ exists and satisfies

$$v(t)=\int_0^t a(t-s)g'(\delta, x_0(s)+u_0(s))(v(s)+u_*(s)-u_0(s))\,ds.$$

Arguing in the same way as above, we conclude that

$$v(T)=\int_0^T q_0(s)g'(\delta, x_0(s)+u_0(s))(u_*(s)-u_0(s))\,ds.$$

But this means that

$$\begin{aligned}&\lim_{\alpha\to 0+}(J(T, \delta, u_\alpha)-J(T, \delta, u_0))/\alpha\\ &\qquad=\int_0^T(\mu-q_0(s)g'(\delta, x_0(s)+u_0(s)))(u_*(s)-u_0(s))\,ds\geqq 0\end{aligned}$$

by the definition of u_0 and the fact that $u_*\geqq 0$. Combining this inequality with (3.21) and (3.22) we obtain a contradiction and the proof of Lemma 2 is completed.

Next we proceed to show how this solution q_*, x_* of (3.9) and (3.10) can be found. Since the functions max and min are not continuously differentiable and we want to apply the implicit function theorem, we need some approximations. Let

$\alpha \in (0, \mu)$ and define

$$p(\alpha, \omega) = \begin{cases} \dfrac{\omega^2}{\alpha} + \dfrac{\alpha}{4}, & |\omega| \leqq \dfrac{\alpha}{2}, \\ |\omega|, & |\omega| > \dfrac{\alpha}{2}. \end{cases} \tag{3.23}$$

We also define

$$m(\alpha, \gamma, \omega) = \frac{\gamma g'(\delta, \omega) + \mu - p(\alpha, \gamma g'(\delta, \omega) - \mu)}{2} \tag{3.24}$$

and

$$n(\alpha, \gamma, \omega) = g(\delta, \omega_2 + \delta) - 2^{-1} \int_\omega^\infty g'(\delta, s)(1 - p_\omega(\alpha, \gamma g'(\delta, s) - \mu))\, ds. \tag{3.25}$$

It is straightforward to check that the functions p, m and n are continuously differentiable in all their arguments and that for every $\omega_* \in \mathbb{R}$ the first partial derivatives are uniformly bounded when $\alpha \in (0, \mu)$, $\gamma \in \mathbb{R}$ and $\omega \in [\omega_*, \infty)$. To see this we use the facts that $p_\omega(\alpha, \omega)$ is a constant for $|\omega| > \alpha/2$ and that $g'(\delta, \omega)^2/g''(\delta, \omega)$ is uniformly bounded on every set of the form $[\omega_*, \infty)$, cf. (3.2).

Let $X = (q, x)$ be an element in $V = L^\infty(0, T; \mathbb{R}) \times C([0, T]; \mathbb{R})$. For $\lambda \in [0, 1]$, $\alpha \in (0, \mu)$ and $X \in V$ we define the mapping $F(\lambda, \alpha, X)$ by

$$\begin{aligned} F(\lambda, \alpha, X)(t) = \Big(& q(t) - \lambda \Big(a(T-t) + \int_t^T a(s-t) m(\alpha, q(s), x(s))\, ds \Big), \\ & x(t) - \lambda \int_0^t a(t-s) n(\alpha, q(s), x(s))\, ds \Big), \qquad t \in [0, T]. \end{aligned} \tag{3.26}$$

It follows from the differentiability properties of m and n that F is continuously Fréchet-differentiable: $V \to V$, but we need a stronger result.

LEMMA 3. *If $\lambda \in [0, 1]$, $\alpha \in (0, \mu)$ and $X = (q, x)$ is such that $0 \leqq q(t) \leqq c_1$, $c_2 \leqq x(t) \leqq c_3$, $t \in [0, T]$, for some constants c_j, $j = 1, 2, 3$, then $F_X(\lambda, \alpha, X)$ is invertible and the norm of the inverse is bounded by a constant independent of λ, α and X. Moreover, $F_X(\lambda, \alpha, X)^{-1} w$, $w \in V$, can be found as the uniform limit of certain iteration procedures.*

Proof. Fix λ, α and X such that the assumptions of Lemma 3 are satisfied. We put

$$h_1(t) = m_\gamma(\alpha, q(t), x(t)), \quad h_2(t) = m_\omega(\alpha, q(t), x(t)), \quad h_3(t) = n_\gamma(\alpha, q(t), x(t)) \tag{3.27}$$

and we note that by (3.24), (3.25) and (3.27) we also have the important result that

$$n_\omega(\alpha, q(t), x(t)) = h_1(t). \tag{3.28}$$

We observe that there exists a constant c_4 (depending only on c_1, c_2, c_3, g and δ), such that

$$0 \leqq h_1(t) \leqq c_4, \quad -c_4 \leqq h_2(t) \leqq 0, \quad 0 \leqq h_3(t) \leqq c_4, \quad t \in [0, T]. \tag{3.29}$$

The equation $F_X(\lambda, \alpha, X)(v_1, v_2) = (w_1, w_2)$ can be written in the form

$$v_1(t) - \lambda \int_t^T a(s-t) h_1(s) v_1(s)\, ds - \lambda \int_t^T a(s-t) h_2(s) v_2(s)\, ds = w_1(t), \tag{3.30}$$

$$v_2(t) - \lambda \int_0^t a(t-s) h_3(s) v_1(s)\, ds - \lambda \int_0^t a(t-s) h_1(s) v_2(s)\, ds = w_2(t). \tag{3.31}$$

Define the function $r_\lambda(t, s)$ to be the solution of the equation

$$\begin{aligned} r_\lambda(t, s) &= \lambda a(t-s) + \lambda \int_s^t r_\lambda(t, u) h_1(u) a(u-s)\, du \\ &= \lambda a(t-s) + \lambda \int_s^t a(t-u) h_1(u) r_\lambda(u, s)\, du, \qquad 0 \leqq s \leqq t \leqq T. \end{aligned} \tag{3.32}$$

It is straightforward to check that this equation has a unique solution that can be found by iteration and that the iteration procedure converges uniformly, cf. [2] and the proof of Lemma 2 above. Thus one also sees from (2.1) and (3.29) that there exists a constant c_5 such that,

$$0 \leqq r_\lambda(t, s) \leqq c_5, \qquad 0 \leqq s \leqq t \leqq T. \tag{3.33}$$

Using (3.32) we can rewrite the (3.30) and (3.31) as

$$v_1(t) = \int_t^T r_\lambda(s, t) h_2(s) v_2(s)\, ds + w_1(t) + \int_t^T r_\lambda(s, t) w_1(s)\, ds, \tag{3.34}$$

$$v_2(t) = \int_0^t r_\lambda(t, s) h_3(s) v_1(s)\, ds + w_2(t) + \int_0^t r_\lambda(t, s) w_2(s)\, ds. \tag{3.35}$$

If we let $y(t) = h_3(t)^{1/2} v_1(t)$, then we obtain from (3.34) and (3.35) the equation

$$y(t) + \int_0^T H(t, s) y(s)\, ds = f(t), \tag{3.36}$$

where

$$H(t, s) = -\int_{\max\{s,t\}}^T r_\lambda(\tau, t) r_\lambda(\tau, s) h_2(\tau) h_3^{1/2}(s) h_3^{1/2}(t)\, d\tau$$

and

$$\begin{aligned} f(t) = h_3(t)^{1/2} \Big(& w_1(t) + \int_t^T r_\lambda(s, t) w_1(s)\, ds \\ & + \int_t^T r_\lambda(s, t) h_2(s) \Big(w_2(s) + \int_0^s r_\lambda(s, \tau) w_2(\tau)\, d\tau \Big)\, ds \Big). \end{aligned}$$

The important thing about the function H is that by (3.29) it defines a monotone operator in $L^2(0, T)$, i.e.,

$$\int_0^T z(t) \int_0^T H(t, s) z(s)\, ds \geqq 0, \qquad z \in L^2(0, T).$$

Since it is also a bounded operator, cf. (3.29), (3.33), we can solve equation (3.36) iteratively by

$$y_{j+1}(t) = y_j(t) - \nu \Big(y_j(t) + \int_0^T H(t, s) y_j(s)\, ds - f(t) \Big),$$

where ν is a sufficiently small positive number, and $y_j \to y$ in $L^2(0, T)$. From (3.36) we also obtain the bound

$$\int_0^T |y(t)|^2\, dt \leqq \int_0^T |f(t)|^2\, dt.$$

Since we now know the function y, we can get v_2 from (3.35) and then v_1 from (3.34). Since y appears inside the integral we get the desired bounds on v_1 and v_2 in terms of the sup-norms of w_1 and w_2 and we also obtain the uniform convergence of the approximate solutions. This completes the proof of Lemma 3.

When $\alpha \in (0, \mu)$ and $\lambda = 0$ we have $F(\alpha, 0, 0) = 0$ and if for some other $\lambda \in (0, 1]$ we have found $X = (q, x)$ such that $F(\alpha, \lambda, X) = 0$, then we conclude from (3.26) that $0 \leqq q(t) \leqq a(0+) + \mu$, since (2.1) holds, and $0 \leqq m(\alpha, \gamma, \omega) \leqq \mu$. On the other hand we also have $g(\delta, \omega_2 + \delta) \geqq n(\alpha, \gamma, \omega) \geqq g(\delta, \omega)$ and hence we see that $x(t)$ satisfies (3.4). But this means that we can apply Lemma 3 and the implicit function theorem to construct a function $X(\lambda, \alpha)$ such that

$$F(\lambda, \alpha, X(\lambda, \alpha)) = 0, \qquad \alpha \in (0, \mu), \quad \lambda \in [0, 1]. \tag{3.37}$$

We note that the construction of this function $X(\lambda, \alpha)$ relies on the iteration procedure (i.e., the Newton method), used to establish the implicit function theorem.

Since $X(1, \alpha)$ satisfies bounds of the form given in the hypothesis of Lemma 3, we conclude that $F_\alpha(\alpha, \lambda, X(1, \alpha))$ is uniformly bounded. Invoking the implicit function theorem and (3.37) we see that $X(1, \alpha)$ is continuously differentiable with respect to α and the derivative is uniformly bounded. But this means that $X(1, \alpha)$ converges as $\alpha \to 0+$ and in view of the definitions (3.24)–(3.26) and the uniqueness result given in Lemma 2, we see that the limit must be (q_*, x_*). Thus we have been able to construct the solution of Problem (T, δ). It remains to show that this solution converges to the solution of the original problem as $\delta \to 0$ and $T \to \infty$.

Next we establish a result concerning the asymptotic behavior of the optimal solution x_*, and this proof will involve our choice of μ.

LEMMA 4. *If $\delta \in (0, \delta_0)$ is sufficiently small, $T > 0$ is sufficiently large and we define $u_*(T, \delta, t) = 0$, $t \geqq T$, then $x_*(T, \delta, t) \geqq \omega_1$, $t \geqq T$.*

Proof. Choose a number $\omega_3 \in (\omega_1, \omega_2)$, so that $g(\delta, \omega_3) > \omega_3$ and let $\rho = (g(\delta, \omega_3) - \omega_3)/2$. Let τ_0 be a number such that

$$\omega_4 \in (\omega_1, \omega_2) \quad \text{if } \omega_4 = \omega_2 + \rho - g(\delta, \omega_3) \int_0^{\tau_0} a(\tau)\, d\tau + \omega_3, \tag{3.38}$$

and

$$g(\delta, \omega_3) \int_0^{\tau_0} a(s)\, ds + g(\delta, \beta) \int_{\tau_0}^{\infty} a(s)\, ds > \omega_3, \qquad a(\tau_0) > 0, \tag{3.39}$$

where β is a negative number such that

$\omega \leqq \beta$ implies that

$$\tau_1 = \inf \left\{ \tau \,|\, \omega_2 + \delta_0 - \omega - 2^{-1} |g(\delta, \omega)| \int_0^{\tau} a(t)\, dt < 0 \right\} < \infty,$$

$$a(0+)\eta g'(\delta, \omega)\, e^{a(0+)g'(\delta,\omega)\tau_1} < -\omega + \omega_1 - \eta, \tag{3.40}$$

$$|g(\delta, \omega)| \int_0^{\tau_1} a(t)\, dt > 2a(0+)\eta g'(\delta, \omega)\, e^{a(0+)g'(\delta,\omega)\tau_1}.$$

Here

$$\eta = \int_0^{\infty} \max \left\{ 0, \omega_4 - g(\delta, \omega_4) \int_0^t a(s)\, ds \right\} dt. \tag{3.41}$$

It follows from (2.1)–(2.3) and (3.2) that one can find such numbers τ_0, ρ and β independent of δ.

Now we choose the number μ so that

$$\mu \in (0, 1), \quad \mu < a(t)g'(\delta, \omega_3), \quad t \in (0, \tau_0]. \tag{3.42}$$

This implies in view of (3.12) and the fact that $q_*(t) \geqq a(T-t)$ by (3.9) that

$$x_*(t) + u_*(t) > \omega_3 \quad \text{on } [T-\tau_0, T). \tag{3.43}$$

Assume that $\delta \in (0, \rho) \cap (0, \delta_0)$ and that T is so large that $T > \tau_0$ and $g(\delta, \omega_4) \int_0^T a(t)\, dt > \omega_4$. Define $t_0 = \inf \{t > T | x_*(t) < \omega_3\}$ and assume that t_0 is finite. If this is not the case, then we are done, since $\omega_3 > \omega_1$. It follows from (3.43) and the fact that $g(\delta, \cdot)$ is nonincreasing that

$$\int_0^{t_0} a(t_0 - s) \min \{0, g(\delta, x_*(s) + u_*(s))\}\, ds \leqq \min \{\omega_3, x_*(T)\} - g(\delta, \omega_3) \int_0^{\tau_0} a(t)\, dt$$

and since a is nonincreasing and $g(\delta, x_*(t) + u_*(t)) > 0$ on $[T-\tau_0, t_0]$ we also have

$$\int_0^{T-\tau_0} a(T-s) \min \{0, g(\delta, x_*(s) + u_*(s))\}\, ds \leqq \omega_3 - g(\delta, \omega_3) \int_0^{\tau_0} a(t)\, dt. \tag{3.44}$$

But then it follows from (2.1), (3.3), (3.38), (3.44) and the fact that $g(\delta, \omega) \leqq \omega_2 + \rho$, that $x_*(T) \leqq \omega_4$. Since u_* is the solution of Problem (T, δ) we must therefore have

$$\int_0^T u_*(t)\, dt \leqq \eta, \tag{3.45}$$

because if we take $u(t) = \max \{0, \omega_4 - g(\delta, \omega_4) \int_0^t a(s)\, ds\}$, then $x(T) = G(\delta, u)(T) \geqq \omega_4$. By the same reasoning we deduce that $x_*(T) \geqq \omega_4 - \eta$.

Let $\beta_0 = \inf \{x_*(t) | t \in [0, T]\}$. We will derive a contradiction from the assumption that $\beta_0 \leqq \beta$. Let $t_1 \in [0, T]$ be such that $x(t_1) = \beta_0$. Since $x_*(T) \geqq \omega_4 - \eta$, it follows from (3.40) that $t_1 < T$. By (2.1) we have

$$\begin{aligned} &\int_0^{t_1} a(t-s) g(\delta, x_*(s) + u(s))\, ds \\ &\qquad \leqq \min \left\{ \beta_0 + |g(\beta_0)| \int_0^{t-t_1} a(s)\, ds, \omega_2 + \delta_0 \right\}, \qquad t \geqq t_1. \end{aligned} \tag{3.46}$$

On the other hand we deduce, since $g(\delta, \cdot)$ is concave and $u_*(t) \geqq 0$, that

$$g(\delta, x_*(t) + u_*(t)) \leqq g(\delta, \beta_0) + g'(\delta, \beta_0)(x_*(t) + u_*(t) - \beta_0). \tag{3.47}$$

Using (2.1), (3.3) and (3.45)–(3.47) we get the following inequality for $y(t) = x_*(t + t_1) - \beta_0$:

$$\begin{aligned} y(t) \leqq \min &\left\{ 0, \omega_2 + \delta_0 - \beta_0 - |g(\delta, \beta_0)| \int_0^t a(s)\, ds \right\} \\ &+ a(0+) \eta g'(\delta, \beta_0) + \int_0^t a(0+) g'(\delta, \beta_0) y(s)\, ds. \end{aligned}$$

If we apply Gronwall's inequality, then we obtain

$$y(t) \leqq \min\left\{0, \omega_2 + \delta_0 - \beta_0 - |g(\delta, \beta_0)| \int_0^t a(s)\, ds\right\} + a(0+)\eta g'(\delta, \beta_0)\, e^{a(0+)g'(\delta,\beta_0)t}, \qquad t \in [0, T - t_1]. \tag{3.48}$$

By the definition of β_0 we have $y(t) \geqq 0$ and since $x_*(T) \geqq \omega_4 - \eta$ we have $y(T - t_1) \geqq \omega_4 - \eta - \beta_0$. But then we get a contradiction from (3.40) and (3.48) if $\beta_0 \leqq \beta$.

Since we have now proved that $x_*(t) + u_*(t) > \beta$ on $[0, T]$ we get a contradiction from (3.39) and (3.44). This completes the proof of Lemma 4.

Next we consider what happens when $\delta \to 0$.

LEMMA 5. *If $T > 0$, then $u_*(T, \delta, \cdot) \to u_*(T, 0, \cdot)$ and $x_*(T, \delta, \cdot) \to x_*(T, 0, \cdot)$ uniformly on $[0, T]$ as $\delta \to 0$. Moreover,*

$$u_*(T, 0, t) = \max\{0, Q(0, q_*(T, 0, t) - x_*(t))\} \tag{3.49}$$

where $Q(0, \cdot)$ is defined by (3.11) when we let $g'(0, \omega_2) = [0, g'(\omega_2)]$ (i.e., a set-valued map) and

$$q_*(T, 0, t) = a(T - t) + \int_t^T a(s - t)k(T, s)\, ds, \tag{3.50}$$

where $k(T, \cdot)$ is a measurable function satisfying $0 \leqq k(T, s) \leqq \mu$. Moreover, if T is sufficiently large and $u_*(T, 0, t) = 0$, $t > T$, then

$$x_*(T, 0, t) \geqq \omega_1, \qquad t \geqq T. \tag{3.51}$$

Proof. Proceeding in the same way as in the proof of Lemma 4 we can show that there exists a constant C_0 independent of T and δ such that

$$x_*(T, \delta, t) \geqq C_0, \qquad \delta \in [0, \delta_0], \quad t \in [0, T]. \tag{3.52}$$

(For the proof in Lemma 4 to work we need an upper bound on $\int_0^T u_*(T, \delta, t)\, dt$, and this is easy to find since this bound may now depend on μ.)

Choose a sequence $\{\delta_j\}_{j=1}^{\infty}$ converging to 0. Since we know that $\min\{\mu, g'(\delta, x_*(T, \delta, \cdot))q_*(T, \delta, \cdot)\}$ and $g(\delta, \max\{x_*(T, \delta, \cdot), Q(\delta, q_*(T, \delta, \cdot))\})$ are uniformly bounded we conclude from (2.1), (3.9) and (3.10) that $x_*(T, \delta, \cdot)$ and $q_*(T, \delta, \cdot) - a(T - \cdot)$ are uniformly bounded and equicontinuous. It follows that we can choose a subsequence also denoted by $\{\delta_j\}$ such that there exist functions $q_*(T, 0, \cdot)$, y and $k(T, \cdot)$, $0 \leqq k(T, s) \leqq \mu$, such that

$$\begin{aligned} &q_*(T, \delta_j, \cdot) \to q_*(T, 0, \cdot),\ x_*(T, \delta_j, \cdot) \to y(\cdot) \quad \text{uniformly on } [0, T] \text{ and} \\ &\min\{\mu, g'(\delta_j, x_*(T, \delta_j, \cdot))q_*(T, \delta_j, \cdot)\} \to k(T, \cdot) \quad \text{weakly in } L^2(0, T). \end{aligned} \tag{3.53}$$

Thus we obtain (3.50). If we define

$$v(t) = \max\{0, Q(0, q_*(T, 0, t)) - y(t)\} \tag{3.54}$$

then we see from (2.2), (2.3), (3.2), (3.11), (3.12) and (3.53) that

$$u_*(T, \delta_j, \cdot) \to v(\cdot) \quad \text{uniformly on } (0, T]. \tag{3.55}$$

But it also follows from (3.10), (3.12), (3.53) and (3.55) that $y = G(0, v)$. Thus, if $v(\cdot)$ and $u_*(T, 0, \cdot)$ differ on a set of positive measure, then we must, by Lemma 1, have

$$J(T, 0, v) > J(T, 0, u_*(T, 0, \cdot)). \tag{3.56}$$

Since $g(0, \omega) = g(\omega_2)$ when $\omega > \omega_2$ and $\int_0^\infty a(t)\,dt = 1$ we have $x_*(T, 0, t) + u_*(T, 0, t) \leqq \omega_2$ and hence by (3.2) $J(T, 0, u_*(T, 0, \cdot)) = J(T, \delta, u_*(T, 0, \cdot)) \geqq J(T, \delta, u_*(T, \delta, \cdot))$. Because it follows from (3.53) and (3.55) that $J(T, \delta_j, u_*(T, \delta_j, \cdot)) \to J(T, 0, v)$ we obtain a contradiction from (3.56). This contradiction shows that $u_*(T, 0, \cdot) = v(\cdot)$ a.e. and hence that $x_*(T, 0, \cdot) \equiv y(\cdot)$.

As the sequence $\{\delta_j\}$ was arbitrary we obtain the first part of the assertion.

It is easy to see that if we define $u_*(T, \delta, t) = 0$, $t \geqq T$, $\delta \in [0, \delta_0)$, then $x_*(T, \delta, \cdot) \to x(T, 0, \cdot)$ uniformly on compact subsets of $\mathbb{R}$ as $\delta \to 0$ and therefore we obtain (3.51) from Lemma 4. This completes the proof of Lemma 5.

Now we proceed to consider (3.1) and we define

$$\begin{aligned} U_0 = \{u : \mathbb{R}^+ \to \mathbb{R}^+ \mid & u \text{ is measurable, } \int_0^\infty u(s)\,ds < \infty; \\ & \lim_{t\to\infty} x(t) \geqq \omega_1 \text{ and } \inf\{t > 0 \mid x(t) > 0\} = 0, \text{ where } x = G(u)\}. \end{aligned} \tag{3.57}$$

It is clear that U_0 is nonempty. As we already noted above $\lim_{t\to\infty} G(u)(t)$ exists if $u(t) \geqq 0$ and $\int_0^\infty u(t)\,dt < \infty$. Next we establish a useful technical result.

LEMMA 6. *If* $u_0 \in U_0$, $x_0 = G(u_0)$, $\omega \in (\omega_1, \omega_2)$, $g'(\omega) < 1$ *and* mes $(\{t \geqq 0 \mid x_0(t) + u_0(t) > \omega, u_0(t) > 0\}) > 0$ *or* mes $(\{t \geqq 0 \mid x_0(t) + u_0(t) \leqq \omega_0\}) > 0$ *or* $\lim_{t\to\infty} x_0(t) > \omega_1$, *then*

$$\int_0^\infty u_0(t)\,dt > \inf\left\{\int_0^\infty u(s)\,ds \,\middle|\, u \in U_0\right\}.$$

Proof. Let us first assume that the set $\{t \geqq 0 \mid x_0(t) + u_0(t) > \omega, u_0(t) > 0\}$ has positive measure. Define the function v as follows

$$v(t) = \begin{cases} u_0(t), & x_0(t) + u_0(t) \leqq \omega, \\ \max\{0, \omega - x_0(t)\}, & \text{otherwise} \end{cases} \tag{3.58}$$

and let

$$w(t) = \int_0^t a(t-s)(g(x_0(s) + u_0(s)) - g(x_0(s) + v(s)))\,ds. \tag{3.59}$$

If we define y by

$$y(t) = \int_0^t a(t-s) g(x_0(s) + v(s))\,ds, \tag{3.60}$$

then we see from (3.1) and (3.59) that

$$y(t) = \int_0^t a(t-s) g(y(s) + v(s) + w(s))\,ds. \tag{3.61}$$

Because $\lim_{t\to\infty} x_0(t) \geqq \omega_1$, $g(\omega_1) = \omega_1$, g nondecreasing, $v(t) \geqq 0$ and $\int_0^\infty a(s)\,ds = 1$, it follows from (3.60) that $\lim_{t\to\infty} y(t) \geqq \omega_1$. Hence $v + w \in U_0$ by (3.58), (3.60), (3.61) and the fact that $w(t) \geqq 0$. From the definitions (3.58) and (3.59) we get, since g is concave,

$$\int_0^\infty (w(t) + v(t))\,dt \leqq \int_0^\infty v(t)\,dt + \int_0^\infty g'(\omega)(u_0(t) - v(t))\,dt < \int_0^\infty u_0(t)\,dt,$$

because $g'(\omega) < 1$ and $\int_0^\infty (u_0(t) - v(t))\,dt > 0$ by assumption and (3.58). Thus we get the desired assertion.

Next we assume that the set $\{t \geqq 0 \mid x_0(t) + u_0(t) \leqq \omega_0\}$ has positive measure. By the previous result we may assume that $u_0(t) = 0$ if $x_0(t) + u_0(t) > \omega_0$ and $g'(x_0(t) +$

$u_0(t))<1$. Let $\chi(t)=0$ if $x_0(t)+u_0(t)\leqq\omega_0$ and $\chi(t)=1$ otherwise. Since $u_0\in U_0$, we must have $\int_0^\infty \chi(t)\,dt=\infty$. Define the functions v, y and z by

$$v\Big(\int_0^t \chi(s)\,ds\Big)=u_0(t),\quad y=G(v),\quad z(t)=y\Big(\int_0^t \chi(s)\,ds\Big). \tag{3.62}$$

A change of variables shows by (3.1) and (3.62) that z satisfies the equation

$$z(t)=\int_0^t a\Big(\int_s^t \chi(\tau)\,d\tau\Big)\,g(z(s)+u_0(s))\chi(s)\,ds. \tag{3.63}$$

Let

$$w(t)=\int_0^t a\Big(\int_s^t \chi(\tau)\,d\tau-a(t-s)\Big)\,g(x_0(s)+u_0(s))\chi(s)\,ds. \tag{3.64}$$

By assumption $\chi(t)=0$ on a set of positive measure and since a is nonincreasing and nonconstant on $(0,\infty)$ and the set where $\chi(t)=0$ does not contain a nonempty set of the form $[0,\tau)$ (see the definition of U_0), it follows that

$$w(t)\geqq 0,\qquad w(t)\not\equiv 0. \tag{3.65}$$

By (3.1), (3.63) and (3.64) we have

$$z(t)-x_0(t)=w(t)+\int_0^t a\Big(\int_s^t \chi(\tau)\,d\tau\Big)\,(g(z(s)+u_0(s))-g(x_0(s)+u_0(s)))\chi(s)\,ds. \tag{3.66}$$

From (2.1), (2.2) and (3.65) we see that $z(t)\geqq x_0(t)$, and that strict inequality holds on a set of positive measure. Since $\lim_{t\to\infty} x_0(t)\geqq\omega_1$, there exists a number t_0 such that $x_0(t)\geqq(\omega_1+\omega_0)/2$, $t\geqq t_0$. We claim that

$$z(t)>x_0(t)\quad\text{on } [t_0,\infty). \tag{3.67}$$

Suppose that (3.67) does not hold but that there exists a number $t_1\geqq t_0$ such that $z(t_1)>x_0(t_1)$. Let $t_2=\sup\{t\,|\,z(\tau)>x_0(\tau),\ \tau\in[t_1,t]\}$. Then we have $z(t_2)=x_0(t_2)$ but since $t_1\geqq t_0$ and $x_0(t)+u_0(t)\leqq\omega_2$ we have $g(z(t)+u_0(t))>g(x_0(t)+u_0(t))$ on (t_1,t_2) by (2.3) and we get a contradiction from (3.65) and (3.66). This implies that there exist numbers t_3 and t_4 such that $z(t)=x_0(t)$, $t\geqq t_4$, $z(t)>x_0(t)$, $t_3<t<t_4$. Unless $\chi(t)=0$ a.e. on (t_3,t_4) we get a contradiction from (3.66) and if $\chi(t)=0$ a.e. on (t_3,t_4) then it follows from (3.64) that $w(t)$ must be positive at some point in (t_4,∞) and this gives a contradiction too. Thus (3.67) holds.

It follows from the results in [5] and our assumptions on g and a that $\lim_{t\to\infty} y(t)=\lim_{t\to\infty} z(t)$ must be one of the numbers 0, ω_1 or ω_2 and since $z(t)\geqq x_0(t)$ and $u_0\in U_0$ there remains the possibilities ω_1 and ω_2. We claim that

$$\lim_{t\to\infty} z(t)=\lim_{t\to\infty} y(t)=\omega_2. \tag{3.68}$$

Suppose that this is not the case, but that $\lim_{t\to\infty} z(t)=\omega_1$. Then there exists a number $t_5>t_0$ and a number $\omega_5\in(\omega_1,\omega_2)$ so that $g'(\omega_5)>1$ and

$$z(t)\leqq\omega_5,\qquad t\geqq t_5. \tag{3.69}$$

If $m\{t>t_5\,|\,u_0(t)>0\}>0$, then we can by (3.67) choose a function $u_1(t)\geqq 0$ such that $\int_0^\infty u_1(t)\,dt<\int_0^\infty u_0(t)\,dt$ and if we define v_1 by $v_1(\int_0^t \chi(s)\,ds)=u_1(s)$, and $y_1=G(v_1)$, $z_1(t)=y_1(\int_0^t \chi(\tau)\,d\tau)$, then $z_1(t)\geqq x_0(t)$, $t\in\mathbb{R}^+$. This would give the assertion of the Lemma and hence we have to consider the case that $u_0(t)=0$, $t\geqq t_5$. But this

implies, by (3.69), that

$$g(z(t)+u_0(t))-g(x_0(s)+u_0(s)) \geqq g'(\omega_5)(z(t)-x_0(t)), \qquad t \geqq t_5.$$

Thus, using (2.1), (3.65)–(3.67) and the fact that z and x_0 are continuous we deduce a contradiction by taking a sequence $\{\tau_j\}_{j=1}^{\infty}$ tending to $+\infty$ such that $z(\tau_j)-x_0(\tau_j) = \inf_{\tau\in(t_5,\tau_j)}(z(\tau)-x_0(\tau))>0$ and invoking the fact that $g'(w_5)\int_0^{\infty} a(s)\,ds>1$. This contradiction shows that (3.68) holds.

It is easy to check, using the Lipschitz-continuity of g that if $\{v_j\}_{j=1}^{\infty}$ is a sequence of nonnegative measurable functions on $\mathbb{R}^+$ such that $v_j \to v$ in $L^1(\mathbb{R}^+)$ as $j\to\infty$ and $\int_0^{\infty} v_j(t)\,dt<\int_0^{\infty} v(t)\,dt$ then $y_j = G(v_j)$ converges to y, uniformly on compact subsets of $\mathbb{R}^+$. In view of (3.68), this implies that there exist numbers $\omega_6\in(\omega_1,\omega_2)$, $t_6, t_7\in\mathbb{R}^+$ and $j_0\geqq 1$ such that $y_{j_0}(t)\geqq\omega_6$ on (t_6, t_7) and $g(\omega_6)\int_0^{t_7-t_6} a(s)\,ds>\omega_6$. But then it follows from (3.1) that $y_{j_0}(t)\geqq\omega_6$ for all $t\geqq t_7$, that is, $v_{j_0}\in U_0$. Since $\int_0^{\infty} v_{j_0}(t)\,dt<\int_0^{\infty} v(t)\,dt\leqq\int_0^{\infty} u_0(t)$, we get the desired assertion. Since the last case of Lemma 6 follows from the argument above, the proof of Lemma 6 is completed.

Now we are in a position to prove the next lemma.

LEMMA 7. *There exists a unique function $u^*\in U_0$ such that $\int_0^{\infty} u^*(t)\,dt = \inf\{\int_0^{\infty} u(s)\,ds \mid u\in U_0\}$. Moreover, if $x^* = G(u^*)$, then $x^*(t)+u^*(t)>\omega_0$ and $u^*(t)=0$ if $g'(x^*(t)+u^*(t))<1$, a.e., $t\in\mathbb{R}^+$ and $\lim_{t\to\infty} x^*(t)=\omega_1$.*

Proof. Let $\{u_j\}_{j=1}^{\infty}\subset U_0$ be a sequence such that $\int_0^{\infty} u_j(t)\,dt\to\inf\{\int_0^{\infty} u(t)\,dt \mid u\in U_0\}$ as $j\to\infty$ and let $x_j = G(u_j)$. By Lemma 6 we may assume that $x_j(t)+u_j(t)>\omega_0$ and $u_j(t)=0$ if $g'(x_j(t)+u_j(t))<1$ for a.e. $t\in\mathbb{R}^+$. This implies by (2.2) and (2.3) that $x_j(t)+u_j(t)\leqq\omega_2$ and hence it follows from (2.1) and (3.1) that the sequence $\{x_j\}_{j=1}^{\infty}$ is uniformly bounded and equicontinuous and that the sequence $\{u_j\}$ is bounded in $L^2(\mathbb{R}^+)$. If we proceed in the same way as in the proof of Lemma 1 and note that g is concave on (ω_0,∞), then we conclude that there exist subsequences, again denoted by $\{x_j\}$ and $\{u_j\}$ and functions y and v such that

$$\begin{aligned} &x_j\to y \quad \text{uniformly on compact subsets of } \mathbb{R}^+,\\ &u_j\to v \quad \text{weakly in } L^2(\mathbb{R}^+) \text{ as } j\to\infty \end{aligned} \tag{3.70}$$

and

$$x(t)\geqq y(t), \quad t\in\mathbb{R}^+, \quad \text{where } x = G(v) \text{ and } \int_0^{\infty} v(t)\,dt\leqq\inf\left\{\int_0^{\infty} u(s)\,ds \mid u\in U_0\right\}. \tag{3.71}$$

Suppose that $v\notin U_0$. If this is a consequence of the fact that $x(t)=0$ on $[0,\tau]$, $\tau>0$ then we can replace v by $v_\tau(t) = v(t-\tau)$ and proceed. Thus we must have $\lim_{t\to\infty} x(t)<\omega_0$ and then we see from (2.2), (2.3) and the results in [5] that

$$\lim_{t\to\infty} x(t) = 0.$$

By (3.70) and (3.71) this implies that there exist numbers j_0, t_1 and t_2 such that

$$x_{j_0}(t)\leqq\frac{\omega_0}{2} \quad \text{on } (t_1,t_2) \quad \text{and} \quad \frac{(t_2-t_1)\omega_0}{2}>\int_0^{\infty} u_{j_0}(t)\,dt. \tag{3.72}$$

Since we assumed that $x_{j_0}(t)+u_{j_0}(t)>\omega_0$ a.e. we have $u_{j_0}(t)>\omega_0/2$ a.e. on (t_2-t_1) and therefore we get a contradiction from the second part of (3.72). Thus we must have $v\in U_0$ and by the second part of (3.71), v is an element in U_0 that minimizes the L^1-norm.

To prove that this optimal element is unique we proceed in the same way as in Lemma 1, using also the results in Lemma 6 and its proof. The last part of the assertion of Lemma 7 follows directly from Lemma 6. This completes the proof of Lemma 7.

There is one additional result we have to establish and then the proof of the theorem is complete.

LEMMA 8. *$x_*(T, 0, \cdot) \to x^*(\cdot)$ and $u_*(T, 0, \cdot) \to u^*(\cdot)$ uniformly on compact subsets of $\mathbb{R}^+$ as $T \to \infty$. Moreover, u^* is continuous.*

Proof. Let $\{T_j\}_{j=1}^{\infty}$ be a sequence tending to $+\infty$. By (2.1), (3.2), (3.3) and (3.50)–(3.52) the functions $x_*(T_j, 0, \cdot)$ and $q_*(T_j, 0, \cdot) - a(T_j - \cdot)$ are uniformly bounded and equicontinuous on $\mathbb{R}^+$ (we let $q_*(T_j, 0, t) = 0$, $t > T_j$). Hence we can choose a subsequence, again denoted by $\{T_j\}$ such that there exist continuous functions y and q so that

$$x_*(T_j, 0, \cdot) \to y(\cdot), q_*(T, 0, \cdot) \to q(\cdot) \text{ uniformly on compact subsets of } \mathbb{R}^+ \text{ as } j \to \infty. \tag{3.73}$$

Moreover, by (3.49) we conclude that if we define v by

$$v(t) = \max\{0, Q(0, q(t)) - y(t)\} \tag{3.74}$$

then v is continuous and

$$u_*(T_j, 0, \cdot) \to v(\cdot) \text{ uniformly on compact subsets of } \mathbb{R}^+ \text{ as } j \to \infty. \tag{3.75}$$

From (3.75) it follows that

$$\int_0^\infty v(t)\, dt \leqq \liminf_{j \to \infty} \int_0^\infty u_*(T_j, 0, t)\, dt \tag{3.76}$$

and by (3.3), (3.73) and (3.75) we see that $y = G(0, v)$. We claim that

$$\lim_{t \to \infty} y(t) \geqq \omega_1. \tag{3.77}$$

By (3.51), (3.52), (3.73) and the fact that $x_*(T, 0, t) \leqq \omega_2$ we see that y is bounded and hence it follows, again from (2.1), (3.2), (3.3), (3.76) and [5], that $\lim_{t \to \infty} y(t)$ exists and must be one of the points ω_1 or ω_2 (the number 0 is excluded since $g(0, 0) < 0$). Thus we have (3.77).

Since $x_*(T, 0, t) + u_*(T, 0, t) \leqq \omega_2$ we must by (3.73) and (3.75) also have $y(t) + v(t) \leqq \omega_2$ and this means that $g(0, y(t) + v(t)) \leqq g(y(t) + v(t))$. Therefore, if $z = G(v)$, then we have $z(t) \geqq y(t)$, so that (3.77) implies that $v \in U_0$. Suppose that

$$\int_0^\infty v(t)\, dt > \int_0^\infty u^*(t)\, dt. \tag{3.78}$$

It follows from the same argument that was used in the proof of Lemma 6, that if $u_0(t) \geqq u^*(t)$,

$$\int_0^\infty u^*(t)\, dt < \int_0^\infty u_0(t)\, dt < \int_0^\infty v(t)\, dt \tag{3.79}$$

and $x_0 = G(u_0)$, then

$$\lim_{t \to \infty} x_0(t) = \omega_2. \tag{3.80}$$

Since $x_0(t) + u_0(t) \geqq x^*(t) + u^*(t) > \omega_0$ by Lemma 6 and because we can choose u_0 so that $x_0(t) + u_0(t) \leqq \omega_2$ it follows from (3.2) that we also have $x_0 = G(0, u_0)$. Since

$x_*(T, 0, T) \leqq \omega_2$ we conclude from (3.76), (3.79) and (3.80) that if j is sufficiently large, then $J(T_j, 0, u_0) < J(T_j, 0, u_*(T_j, 0, \cdot))$ and this is a contradiction. Hence we have $\int_0^\infty v(t)\,dt \leqq \int_0^\infty u^*(t)\,dt$ and by the uniqueness result in Lemma 7 we get $v = u^*$. But then we also have $y = x^*$ and since v is continuous by (3.75) and the sequence $\{T_j\}$ was arbitrary, we obtain the assertion of Lemma 8, and the proof is completed.

REFERENCES

[1] V. L. BAKKE, *Optimal fields for problems with delay*, J. Optim. Theory Appl., 33 (1981), pp. 69–84.

[2] G. GRIPENBERG, *On the resolvents of nonconvolution Volterra kernels*, Funkcial. Ekvac., 23 (1980), pp. 83–95.

[3] A. HALANAY, *Optimal controls for systems with time lag*, this Journal, 6 (1968), pp. 215–234.

[4] G. L. HARATIŠILI AND T. A. TADUMADZE, *Nonlinear optimal control systems with variable time lags*, Math. USSR Sb., 35 (1979), pp. 863–881.

[5] S. O. LONDEN, *On the asymptotic behavior of the bounded solutions of a nonlinear Volterra equation*, SIAM J. Math. Anal., 5 (1974), pp. 849–875.

[6] R. K. MILLER, *Nonlinear Volterra Equations*, W. A. Benjamin, Menlo Park, CA, 1971.

[7] W. RUDIN, *Functional Analysis*, McGraw-Hill, New York, 1973.

[8] F. B. VASIL'EV, *Optimality for some classes of systems not solved with respect to the derivative*, Soviet Math. Dokl., 10 (1969), pp. 224–227.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

0363-0129/83/2104-0006 $01.25/0

CANONICAL REALIZATIONS OF SYSTEMS WITH DELAYS*

LUCIANO PANDOLFI†

Abstract. In this paper we consider the properties of the realizations of a class of nonrational transfer functions. The transfer functions of this class can be realized as systems with delays. We introduce the concept of ω-canonical and ∞-canonical realization. This last definition is equivalent to saying that all spectral projections of the realization are controllable and observable systems with no state delay. For a general class of systems, ∞-canonical realizations are shown to be minimal according to a suitable definition.

Key words. linear systems, delay systems, transfer function, canonical realization

1. Introduction. The theory of the realization of transfer functions of linear finite dimensional systems seems to be well understood (see [7], [5]). This problem has been solved with the use of algebraic techniques, which have been successfully extended to the realization problem of transfer functions of a wide class of systems. The problem of the realization of systems with delays can be approached in this framework (see [8], [17], [18] and the references therein). In this paper we study the realization of systems with delays using a different approach.

Let (S) be the control system governed by the equation

$$\text{(1a)} \qquad \dot{x} = \sum_{i=0}^{\nu} A_i x(t-h_i) + \int_{-h}^{0} \bar{A}(s)x(t+s) + \sum_{i=0}^{\nu} B_1 u(t-h_i) + \int_{-h}^{0} \bar{B}(s)u(t+s),$$

$$\text{(1b)} \qquad y(t) = \sum_{i=0}^{\nu} C_i x(t-h_i) + \int_{-h}^{0} \bar{C}(s)x(t+s),$$

where $x \in R^n$, $u \in R^m$, $y \in R^p$ and the matrices A_i, $\bar{A}(s)$ are $n \times n$, B_i, $\bar{B}(s)$ are $n \times m$ and C_i, $\bar{C}(s)$ are $p \times n$, with real elements. We assume that the matrices A_i, B_i, C_i are constant, and that the matrices $\bar{A}(s)$, $\bar{B}(s)$, $\bar{C}(s)$ have square integrable elements. The numbers h, h_i satisfy the conditions $0 = h_0 < h_1 < \cdots < h_\nu \leqq h$.

We denote with the symbol $\mathcal{S}$ the class of control systems just introduced.

In the following, we assume that $u(t)$ is a piecewise continuous function for $t \geqq -h$. The initial data for (1a) will be of the form $u(t) = v(t)$, a piecewise continuous function on $[-h, 0]$, and $x(t) = \varphi(t)$, $t \in [-h, 0)$, $x(0) = x_0 \in R^n$. We assume that the function $\varphi(t)$ is square integrable.

The properties of (1a) that we shall need in this paper are introduced in § 2.

Now we choose matrices $A(t)$, $B(t)$, $C(t)$, whose elements are functions of bounded variation on $[-h, 0]$, and such that we have, for all continuous functions $\varphi(t)$, $v(t)$,

$$\int_{-h}^{0} dA(s)\varphi(s) = \sum_{i=0}^{\nu} A_i \varphi(-h_i) + \int_{-h}^{0} \bar{A}(s)\varphi(s)\, ds,$$

$$\int_{-h}^{0} dB(s)v(s) = \sum_{i=0}^{\nu} B_i v(-h_i) + \int_{-h}^{0} \bar{B}(s)v(s)\, ds,$$

$$\int_{-h}^{0} dC(s)\varphi(s) = \sum_{i=0}^{\nu} C_i \varphi(-h_i) + \int_{-h}^{0} \bar{C}(s)\varphi(s)\, ds.$$

* Received by the editors December 15, 1980, and in final revised form July 13, 1982.
† Istituto di Matematica, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10100 Torino, Italy.

In order to reduce notational complexity, in the following, when the functions $\varphi(t)$, $v(t)$ are continuous, we use the Stieltjes integral notation instead of the right-hand sides of the above formulas.

As usual, if $f(t)$ is a function defined for $t \geqq -h$, f_t denotes $f(t+s)$, $s \in [-h, 0]$ for every $t \geqq 0$.

The state of (S) at the time $t \geqq 0$ may be defined in several ways. We choose to define it as

$$Z(t) = (x(t), x_t, u_t).$$

We consider $Z(t)$ as an element of $M^2 \times L^2$, where $M^2 = R^n \times L^2(-h, 0; R^n)$, $L^2 = L^2(-h, 0; R^m)$. The M^2-component of $Z(t)$ is the pair $(x^0(t), x^1(t)) = (x(t), x_t)$ and will be denoted $X(t)$.

Let $\hat{}$ denote the Laplace transform. A standard calculation proves that, when $z(0) = 0$,

$$\hat{y}(\lambda) = T(\lambda)\hat{u}(\lambda), \qquad \text{where}$$

$$T(\lambda) = \left(\int_{-h}^{0} dC(s) \exp(\lambda s)\right) \Delta^{-1}(\lambda) \left(\int_{-h}^{0} dB(s) \exp(\lambda s)\right),$$

$$\Delta(\lambda) = \lambda I - \int_{-h}^{0} dA(s) \exp(\lambda s).$$

$T(\lambda)$ is called the transfer function of (S).

Assume that $T(\lambda)$ is the transfer function of a system of class $\mathcal{S}$ (Theorem 1.1. will give a necessary and sufficient condition for this). In this paper we want to investigate if $T(\lambda)$ admits a realization which is canonical (the definition is in § 2). The approach that we use in this paper was previously applied to systems with only input delays [13].

Now we observe that if a realization of $T(\lambda)$ is given (i.e., if (1a) and (1b) are given) we can read the numbers n, m, p. The numbers m, p are determined by $T(\lambda)$ (which is a $p \times m$ matrix). The number $n = \dim x$ depends on (S), and will be denoted $n(S)$. The control process (S) is infinite dimensional, but the number $n(S)$ is always finite.

The organization of this paper is as follows. Some preliminary material is introduced in § 2. In particular, ω-canonical and ∞-canonical (shortly, canonical) realizations are defined. In § 3 we give a condition which characterizes those transfer functions which admit ω-canonical realizations, using a Hankel matrix. We shall see that there are systems of class $\mathcal{S}$ whose transfer functions do not admit canonical realizations in class $\mathcal{S}$. In § 4 we shall prove, under special assumptions, that if (S) is a canonical realization of its transfer function, then $n(S)$ is as small as possible. This means that we can associate with (S) the "dimension" n (the rank n, in the terminology of [17]) and that we can say that canonical realizations are minimal.

Now we give a characterization of those transfer functions which admit a realization of class $\mathcal{S}$. We observe the following properties of $T(\lambda)$: $T(\lambda)$ is a meromorphic function, $T(\lambda) = H(\lambda)/d(\lambda)$, where

i. $H(\lambda) = \sum_{i=0}^{n-1} K_i(\lambda)\lambda^i$, where $K_i(\lambda)$ is a $p \times m$ matrix of entire functions,

ii. $d(\lambda) = \det \Delta(\lambda) = \sum_{i=0}^{n} d_i(\lambda)\lambda^i$, where $d_i(\lambda)$ are entire functions, and $d_n(\lambda) = 1$.

THEOREM 1.1. *$T(\lambda)$ is the transfer function of a system $(S) \in \mathcal{S}$ if and only if it is a meromorphic function*

$$T(\lambda) = \frac{H(\lambda)}{d(\lambda)},$$

such that conditions i, ii *are satisfied and, moreover there exist scalars* $l_i(s)$ *and* $p \times m$ *matrices* $C_i(s)$ *of class* $\mathcal{V}$, *such that*

$$\int_{-h}^{0} e^{\lambda s} dl_i(s) = d_i(\lambda), \qquad \int_{-h}^{0} dC_i(s)\, e^{\lambda s} = K_i(\lambda).$$

Proof. The necessity is obvious. To prove sufficiency we observe that a realization of $T(\lambda)$ is given by

$$A(s) = \begin{bmatrix} 0 & l_n(s)I & 0 & 0 & \cdots & 0 \\ 0 & 0 & l_n(s)I & 0 & \cdots & 0 \\ -l_0(s)I & & \cdots & & & -l_{n-1}(s)I \end{bmatrix}, \qquad B(s) = \begin{bmatrix} I \\ l_n(s)I \end{bmatrix}$$

(I, 0 are the $m \times m$ identity and zero matrices), while $C(s)$ is

$$C(s) = [C_0(s), \cdots, C_{n-1}(s)].$$

□

Of course the realization of $T(\lambda)$ which has been constructed in the above theorem may not be canonical in any acceptable sense.

2. Known facts and preliminaries. We recall some properties about the equation

$$\dot{x} = \sum_{i=0}^{\nu} A_i x(t-h_i) + \int_{-h}^{0} \bar{A}(s) x(t+s) + f(t), \qquad t \geqq 0 \tag{2}$$

(see [3], [6]). Let I, 0 denote the identity and zero matrices or operators, as deduced by the context.

The spectrum of (2) is the set σ of complex numbers

$$\sigma = \{\lambda, \det \Delta(\lambda) = d(\lambda) = 0\}.$$

The set σ is never empty and for any real α, the set $\sigma \cap \{\lambda, \operatorname{Re} \lambda > \alpha\}$ is finite (may be empty).

In this paper ω will always denote a number such that if $|\lambda| = \omega$, then $\lambda \notin \sigma$. The set

$$\sigma_\omega = \sigma \cap \{\lambda, |\lambda| < \omega\}$$

is always finite (may be empty), because $d(\lambda)$ is an entire function.

If $f(t)$ is identically zero, (2) defines a strongly continuous semigroup of bounded operators $S(t)$ on $M^2 = R^n \times L^2(-h, 0; R^n)$. Let A be the infinitesimal generator of $S(t)$. Then $\sigma(A) = \sigma$. If $\sigma_\omega \neq \varnothing$, there exists a finite dimensional subspace $N_\omega \subseteq M^2$ which reduces $S(t)$, and

$$N_\omega = \bigoplus_{\lambda \in \sigma_\omega} \bigcup_{n>0} (\lambda I - A)^n.$$

Let $m(\lambda_0)$ be the multiplicity of the root λ_0 of $d(\lambda)$. Then

$$n_\omega = \dim N_\omega = \sum_{\lambda \in \sigma_\omega} m(\lambda).$$

P_ω will be the projection of M^2 on N_ω.

If $u(t)$ is a given piecewise continuous function, $\tilde{B}u_t$ denotes the M^2-valued function

$$\tilde{B}u_t = \begin{pmatrix} \sum_{i=0}^{\nu} B_i u(t-h_i) + \int_h^0 \bar{B}(s) u(t+s) \\ 0 \end{pmatrix}.$$

On $N_\omega \times L^2$ we consider the control process given by

$$X(t) = S(t)\phi + \int_0^t S(t-s)P_\omega B u_s \, ds, \qquad \phi \in N_\omega. \tag{3}$$

(observe that, according to the conditions given in § 1, the initial datum is an element $(\phi, v) \in N_\omega \times L^2$, with a piecewise continuous $v(t)$).

Let $B(\lambda) = \int_h^0 dB(s)\, e^{\lambda s}\, ds$.

DEFINITION 2.1. We say that the system (S) is ω-controllable when

$$\operatorname{rank}[\Delta(\lambda), B(\lambda)] = n, \quad \forall \lambda, \quad |\lambda| < \omega. \tag{4}$$

and that (S) is ∞-*controllable* (or *spectrally controllable*, or *fully stabilizable*), when it is ω-controllable for every ω.

We know from [11] that, if σ_ω is not empty, (S) is ω-controllable if and only if system (3) is reachable, i.e., if and only if there exists $T > h$ such that for every $X_1 \in N_\omega$, and continuous $v \in L^2$ we can find a control $u(t)$ such that $u_0 = 0$, $u_T = v$, and such that the solution of (3) with $X(0) = 0$ satisfies $X(T) = X_1$.

It is well known that condition (4) plays an important role in the theory of delayed systems [12].

Now we consider the observability properties of (S). Let us assume that $u(t) = 0$ for $t \geqq -h$. Let σ_ω be nonempty. If the initial data $X(0)$ belongs to N_ω, it is known [16] that the condition

$$\operatorname{rank}[\Delta^*(\lambda), C^*(\lambda)] = n, \quad \forall \lambda, \quad |\lambda| < \omega, \tag{5}$$

is necessary and sufficient for $y(t)$ to be zero if and only if $X(0)$ is zero (i.e., so that the projection of (S) on N_ω be observable). This observation suggests the following definition:

DEFINITION 2.2. (S) is ω-*observable* if (5) holds, and is ∞-observable when it is ω-observable for every ω.

Remark 2.1. Our definition of ω-observability should be compared with the definition of ω-detectability in [11].

The system (S) that we are studying has also input delays, so that interpretations of (5) should be given, which take into account the fact that the initial data for (1a) is not $X(0)$, but $Z(0)$. A simple interpretation is the following one. Assume that $u(t)$ is zero for $t \geqq 0$. Observe that, if $u(t)$ is differentiable, the solution of (1a) is also a solution of the equation

$$\dot{x} = \sum_{i=0}^{\nu} A_i x(t-h_i) + \int_{-h}^{0} A(s)x(t+s)\, ds + \sum_{i=0}^{\nu} B_i u(t-h_i) + \int_{-h}^{0} B(s)u(t+s)\, ds,$$

$$\dot{u} = v.$$

With this equation we associate the output (1b) and we call the obtained system (S'). Of course the solutions of (S) and (S') are the same only if $u(0) = 0$. (S') is ω-observable when the matrix

$$\begin{bmatrix} \Delta^*(\lambda) & 0 & C^*(\lambda) \\ B^*(\lambda) & \lambda I & 0 \end{bmatrix}$$

is of full rank for $|\lambda| < \omega$. For $\lambda \neq 0$ this matrix is of full rank if and only if (5) holds, while for $\lambda = 0$ it may be possible to find initial conditions $x(t) = x_0$, $u(t) = u_0$ for $t \in [-h, 0]$ which give zero output. However, we assumed that $u(0) = 0$, so that $u_0 = 0$ and, if (5) holds, $x_0 = 0$. Hence (5) is a necessary and sufficient condition for the ω-observability of (S'), if we impose the (natural) condition $u(0) = 0$.

Another condition which is equivalent to (5) and which should be compared with [11, Thm. 3], is expressed by the following theorem:

THEOREM 2.1. *Let $u(t)$ be zero for $t>0$. Let γ be any number such that when $\lambda \in \sigma, \lambda \notin \sigma_\omega$, then $\operatorname{Re}\lambda < -\gamma$. (S) is ω-observable if and only if when $\lim_{t\to+\infty} y(t)e^{\gamma t} = 0$, then $\lim_{t\to+\infty} e^{\gamma t}x(t) = 0$.*

Proof. Let $t = h + r$. Then

$$X(h+r) = S(h+r)X(0) + \int_0^h S(h+r-s)(Bu_s)\,ds$$

$$= S(r)\left\{S(h)X(0) + \int_0^h S(h-s)(Bu_s)\,ds\right\} = S(r)\bar{X}(0).$$

This shows that the first component of $X(h+r)$ is a solution of (1a) with $u(t)=0$ for every t. If condition (5) holds and if $\exp(\gamma t)y(t)\to 0$ for $t\to+\infty$, then $X(h+r)$ has a zero component on the eigenspaces relative to eigenvalues with $\operatorname{Re}\lambda > -\gamma$ and also $\exp(\gamma t)x(t)$ tends to zero. The converse is obvious. □

Now we can define canonical systems.

DEFINITION 2.3. We say that (S) is an *ω-canonical realization of its transfer function when* (S) is ω-controllable and ω-observable. We say that (S) is an *∞-canonical realization of its transfer function* when it is ω-canonical for every ω.

Remark 2.2. Other definitions of controllability (and observability) for delay systems have been proposed. For example, the definitions of L^2, M^2, F-controllability [3] have been widely studied. All of them imply spectral controllability. An analogous observation holds for the observability property.

Let ω be such that $\sigma_\omega \neq \varnothing$. The elements of N_ω are equivalence classes of pairs (x^0, x^1) with x^1 in $L^2(-h, 0; R^n)$. In each class there is an element whose second component x^1 is continuous. We choose this element to represent its equivalence class. (S_ω) will be the control system on $N_\omega \times L^2$, given by (3), with the output $y(t) = \tilde{C}(X(t))$ where $\tilde{C}: N_\omega \to R^p$ is defined as

$$\tilde{C}(X(t)) = C_0 x^0(t) + C_1 x^1(t-h_1) + \cdots + C_\nu x^1(t-h_\nu) + \int_{-h}^0 C(s)x^1(t+s)\,ds.$$

Let us observe that $\tilde{C}$ is a linear operator which is bounded, since its domain is the finite dimensional space N_ω.

Remark 2.3. In the following (§§ 3, 4), when there is no ambiguity, we drop the symbol $\infty-$, and we say simply controllable, observable and canonical realizations.

3. Canonical realizations. In this section we study the properties of ω-canonical realizations. We follow the same lines as in [13]. Let ω be fixed. There are numbers $r<\omega<R$ such that $T(\lambda)$ has a Laurent expansion in the region

$$\Omega = \{\lambda : r < |\lambda| < R\}.$$

Let

$$T(\lambda) = f(\lambda) + \sum_{k=1}^{\infty} L_k \lambda^{-k}, \qquad f(\lambda) \text{ analytic.}$$

From [19]

$$L_k = \frac{1}{2\pi i}\int_{|\zeta|=r'} T(\zeta)\zeta^{k-1}\,d\zeta$$

(r' is any number such that $r<r'<R$).

Let H_ω be the matrix

$$\begin{bmatrix} L_1 & L_2 & L_3 & \cdots \\ L_2 & L_3 & L_4 & \cdots \\ L_3 & \vdots & \vdots & \cdots \\ \vdots & & & \end{bmatrix}.$$

H_ω is the Hankel matrix of the principal part of $T(\lambda)$ in Ω. Let m_ω be the rank of H_ω.

Let $T_\omega(\lambda)$ be the transfer function of (S_ω), and $R(\lambda)=(\lambda I-A)^{-1}$ be the resolvent operator of $S(t)$. We compute $T_\omega(\lambda)$. Let $Z(0)=0$.

$$\begin{aligned}
\hat{y}(\lambda)=\int_0^{+\infty} e^{-\lambda t}y(t)\,dt &= \int_0^{+\infty} e^{-\lambda t}\tilde{C}\left[\int_0^t S(t-s)P_\omega(\tilde{B}u_s)\,ds\right]dt \\
&= \tilde{C}\int_0^{+\infty} e^{-\lambda t}\int_0^t S(t-s)P_\omega(\tilde{B}u_s)\,ds\,dt \\
&= \tilde{C}\left[\int_0^{+\infty} R(\lambda)\,e^{-\lambda s}P_\omega(\tilde{B}u_s)\,ds\right] \\
&= \tilde{C}\left[R(\lambda)P_\omega\int_0^{+\infty} e^{-\lambda s}\begin{pmatrix}\sum_{i=0}^{\nu}B_iu(t-h_i)+\int_{-h}^0\bar{B}(s)u(t+s)\\ 0\end{pmatrix}ds\right] \\
&= \tilde{C}\left[R(\lambda)P_\omega\begin{pmatrix}\int_h^0 dB(r)\,e^{\lambda r}\\ 0\end{pmatrix}\hat{u}(\lambda)\right].
\end{aligned}$$

From [10, p. 178]

$$R(\lambda)P_\omega=\frac{1}{2\pi i}\int_{|\zeta|=r'}R(\zeta)\frac{1}{\lambda-\zeta}\,d\zeta,\qquad |\lambda|>r'.$$

A standard calculation shows that

$$R(\zeta)\psi=\begin{pmatrix}\Delta^{-1}(\zeta)v\\ \Delta^{-1}(\zeta)e^{\zeta\theta}v-\int_0^\theta e^{\zeta(\theta-s)}\psi^1(s)\,ds\end{pmatrix},$$

where

$$v=\psi^0-\int_{-h}^0 dA(s)\left(\int_0^s e^{\zeta(s-r)}\psi^1(r)\,dr\right)\quad\text{for }\psi=\begin{pmatrix}\psi^0\\ \psi^1\end{pmatrix}\in M^2.$$

Hence

$$R(\zeta)P_\omega\begin{pmatrix}\int_{-h}^0 dB(r)\,e^{\lambda r}\\ 0\end{pmatrix}=\begin{pmatrix}\Delta^{-1}(\zeta)\int_{-h}^0 dB(r)\,e^{\lambda r}\\ \Delta^{-1}(\zeta)e^{\zeta\theta}\int_{-h}^0 dB(r)\,e^{\lambda r}\end{pmatrix}.$$

Hence, recalling the definition of $\tilde{C}$,

$$\begin{aligned}
T_\omega(\lambda) &= \int_{-h}^0 dC(\theta)\frac{1}{2\pi i}\int_{|\zeta|=r'}\left[\frac{e^{\zeta\theta}\Delta^{-1}(\zeta)}{(\lambda-\zeta)}\right]d\zeta\int_{-h}^0 dB(r)\,e^{\lambda r} \\
&= \frac{1}{2\pi i}\int_{|\zeta|=r'}\int_{-h}^0 dC(\theta)\,e^{\zeta\theta}\frac{\Delta^{-1}(\zeta)}{(\lambda-\zeta)}\int_{-h}^0 dB(r)\,e^{\lambda r}\,d\zeta\quad (r<r'<|\lambda|<R).
\end{aligned}$$

We call $\tilde{T}(\lambda)$ the function

$$\begin{aligned}
\tilde{T}(\lambda) &= \frac{1}{2\pi i}\int_{|\zeta|=r'}\frac{1}{\lambda-z}T(z)\,dz \\
&= \frac{1}{2\pi i}\int_{|z|=r'}\left[\frac{1}{\lambda-z}\left(\int_{-h}^0 dC(\theta)\,e^{z\theta}\right)\Delta^{-1}(z)\int_{-h}^0 dB(s)\,e^{zs}\right]dz.
\end{aligned}$$

Let r'' be any number such that $r'<r''<|\lambda|$. $\tilde{T}_\omega(\lambda)$ is the function

$$\begin{aligned}\tilde{T}_\omega(\lambda)&=\frac{1}{2\pi i}\int_{|z|=r''}\frac{T_\omega(z)}{\lambda-z}\,dz\\&=\frac{1}{2\pi i}\int_{|z|=r''}\frac{1}{\lambda-z}\left\{\frac{1}{2\pi i}\int_{|\zeta|=r'}\int_{-h}^{0}dC(\theta)\,e^{\zeta\theta}\frac{\Delta^{-1}(\zeta)}{(z-\zeta)}\,d\zeta\int_{-h}^{0}dB(r)\,e^{zr}\right\}dz\\&=\frac{1}{2\pi i}\int_{|\zeta|=r'}\left\{\frac{1}{2\pi i}\int_{|z|=r''}\frac{1}{\lambda-z}\int_{-h}^{0}dC(\theta)\,e^{\zeta\theta}\Delta^{-1}(\zeta)\int_{-h}^{0}dB(r)\,e^{zr}\frac{1}{(z-\zeta)}\,dz\right\}d\zeta\\&=\frac{1}{2\pi i}\int_{|\zeta|=r'}\int_{-h}^{0}dC(\theta)\,e^{\zeta\theta}\Delta^{-1}(\zeta)\left\{\frac{1}{2\pi i}\int_{|z|=r''}\frac{1}{\lambda-z}\int_{-h}^{0}dB(r)\,e^{zr}\frac{1}{z-\zeta}\,dz\right\}d\zeta.\end{aligned}$$

Observe that $z\to(1/(\lambda-z))\int_{-h}^{0}dB(r)\,e^{zr}$ is analytic for $|z|<r''+\varepsilon$, for some positive ε. Hence

$$\tilde{T}_\omega(\lambda)=\frac{1}{2\pi i}\int_{|\zeta|=r'}\int_{-h}^{0}dC(\theta)\,e^{\zeta\theta}\Delta^{-1}(\zeta)\frac{1}{\lambda-\zeta}\left(\int_{-h}^{0}dB(r)\,e^{\zeta r}\right)d\zeta=\tilde{T}(\lambda).$$

Observe now that $\tilde{T}(\lambda)$ and $\tilde{T}_\omega(\lambda)$ are the principal parts of $T(\lambda)$ and $T_\omega(\lambda)$ for $r<|\lambda|<R$.

Now we calculate $\tilde{H}_\omega$, the Hankel matrix of the coefficients of the principal parts of $T_\omega(\lambda)$ (i.e. of $\tilde{T}_\omega(\lambda)$).

From § 2, (S) is ω-controllable if

$$\operatorname{rank}[\Delta(\lambda),B(\lambda)]=n,\qquad|\lambda|<\omega.$$

Let A_ω be the infinitesimal generator of $P_\omega S(t)P_\omega$. Obviously (see [12]) the above relation holds if and only if

$$\operatorname{rank}[\lambda I-A_\omega,P_\omega B(\lambda)]=n_\omega \tag{6}$$

($n_\omega=\dim N_\omega$). Let $B_1(\lambda)$ be such that

$$P_\omega B(\lambda)=\Phi_\omega(\theta)B_1(\lambda)=\Phi_\omega(\theta)\int_{-h}^{0}dB_1(s)\,e^{\lambda s}$$

where $\phi_\omega(\theta)$ is a matrix whose columns are a basis of N_ω. Condition (6) is equivalent to

$$\operatorname{rank}[B_1(A_\omega),A_\omega B_1(A_\omega),\cdots,A_\omega^{n_\omega-1}B_1(A_\omega),\cdots]=n_\omega,$$

where $B_1(A_\omega)$ is defined as

$$B_1(A_\omega)=\int_{-h}^{0}\exp(A_\omega s)\,dB_1(s).$$

The matrix above in square brackets will be called the ω-controllability matrix of (S) and denoted with the symbol $\mathcal{R}_\omega$. In an analogous way we see that (S) is ω-observable if and only if the rank of the matrix

$$\mathcal{O}_\omega^*=[C_1^*(A_\omega),A_\omega^*C_1^*(A_\omega),\cdots,(A_\omega^{n_\omega-1})^*C_1(A_\omega),\cdots]$$

is n_ω. Here

$$C_1^*(A_\omega)=\int_{-h}^{0}\exp(A_\omega^* s)\,dC_1^*(s)$$

and $C_1(s)$ is given by $P_\omega \int_{-h}^{0} e^{\lambda s}\, dC^*(s) = \Phi_\omega(\theta) \int_{-h}^{0} e^{\lambda s}\, dC_1^*(s)$.

LEMMA 3.1. *$H_\omega = \tilde{H}_\omega = K_\omega$, where*

$$K_\omega = \mathcal{O}_\omega \mathcal{R}_\omega.$$

Proof. We have already shown that $\tilde{T}(\lambda) = \tilde{T}_\omega(\lambda)$. Hence $H_\omega = \tilde{H}_\omega$, since H_ω is the Hankel matrix of the principal part of $T(\lambda)$ for $|\lambda| < \omega$, i.e. of $\tilde{T}(\lambda)$. Now we show that $\tilde{H}_\omega = K_\omega$. $\tilde{H}_\omega$ is the Hankel matrix associated with $T_\omega(\lambda)$, which is the transfer function of a system with delays acting only on the input and output variables. Hence there exists matrices $L(s)$, $M(s)$, $\tilde{A}$ such that

$$T_\omega(\lambda) = \left(\int_{-h}^{0} dL(s)\, e^{\lambda s} \right) (\lambda I - \tilde{A})^{-1} \int_{-h}^{0} dM(s)\, e^{\lambda s}.$$

For $|\lambda| > \max\{|\mu|,\ \mu \in \sigma(\tilde{A})\}$ we have that

$$\begin{aligned} T_\omega(\lambda) &= \frac{1}{\lambda} \int_{-h}^{0} dL(s)\, e^{\lambda s} \left(I + \frac{\tilde{A}}{\lambda} \right)^{-1} \int_{-h}^{0} dM(s)\, e^{\lambda s} \\ &= \frac{1}{\lambda} \left\{ \sum_{n=0}^{\infty} \int_{-h}^{0} dL(s) \frac{s^n}{n!} \lambda^n \right\} \sum_{r=0}^{\infty} \frac{\tilde{A}^r}{\lambda^r} \left\{ \sum_{k=0}^{\infty} \int_{-h}^{0} dM(s) \frac{s^k}{k!} \lambda^k \right\} \\ &= \left\{ \sum_{n=0}^{\infty} \int_{-h}^{0} dL(s) \frac{s^n \lambda^n}{n!} \right\} \left\{ \sum_{k=1}^{\infty} \lambda^{-k} \tilde{A}^{k-1} \right\} \int_{-h}^{0} \exp(\tilde{A}r)\, dM(r) + f(\lambda) \end{aligned}$$

as in [12, p. 33], where $f(\lambda)$ is analytic. Hence

$$\begin{aligned} T_\omega(\lambda) = g(\lambda) &+ \frac{1}{\lambda} \left\{ \int_{-h}^{0} dL(s) + \int_{-h}^{0} dL(s) \frac{s}{1!} \tilde{A} + \int_{-h}^{0} dL(s) \frac{s^2}{2!} \tilde{A}^2 + \cdots \right\} \\ &\qquad \times \int_{-h}^{0} \exp(\tilde{A}r)\, dM(r) \\ &+ \frac{1}{\lambda^2} \left\{ \int_{-h}^{0} dL(s) \tilde{A} + \int_{-h}^{0} dL(s) \frac{s}{1!} \tilde{A}^2 + \int_{-h}^{0} dL(s) \frac{s^2}{2!} \tilde{A}^3 + \cdots \right\} \\ &\qquad \times \int_{-h}^{0} \exp(\tilde{A}r)\, dM(r) \\ &+ \cdots \\ = l(\lambda) &+ \sum_{k=1}^{\infty} \frac{1}{\lambda^k} \left(\int_{-h}^{0} dL(s) \exp(\tilde{A}s) \right) \tilde{A}^{k-1} \int_{-h}^{0} \exp(\tilde{A}r)\, dM(r) \end{aligned}$$

and $l(\lambda)$ is an analytic function.

Hence $\tilde{H}_\omega = \tilde{K}_\omega = \tilde{\mathcal{O}}_\omega \tilde{\mathcal{R}}_\omega$,

$$\tilde{\mathcal{O}}_\omega^* = \left[\left(\int_{-h}^{0} dL(s) \exp(\tilde{A}s) \right)^*, \tilde{A}^* \left(\int_{-h}^{0} dL(s) \exp(\tilde{A}s) \right)^*, \cdots \right],$$

$$\tilde{\mathcal{R}}_\omega = \left[\int_{-h}^{0} \exp(\tilde{A}r)\, dM(r), \tilde{A} \int_{-h}^{0} \exp(\tilde{A}r)\, dM(r), \cdots \right].$$

Now we prove that $\tilde{\mathcal{R}}_\omega = \mathcal{R}_\omega$. Observe that $\tilde{A}$ and $M(s)$ satisfy ($\mathscr{L} = \hat{} =$ Laplace transform)

$$
\begin{aligned}
&(\lambda I - \tilde{A})^{-1}\left(\int_{-h}^{0} dM(s)\, e^{\lambda s} u(s) + \varphi_0\right) \\
&\quad = \mathscr{L}\left(S(t)P_\omega \int_{-h}^{0} dB(s) u(t+s) + S(t)\varphi_0\right) \\
&\quad = (\lambda I - A_\omega)^{-1}\left\{P_\omega \int_{-h}^{0} dB(s)\, e^{\lambda s}\hat{u}(\lambda) + \varphi_0\right\} \\
&\quad = (\lambda I - A_\omega)^{-1}\left\{\phi_\omega(\theta)\int_{-h}^{0} dB_1(\theta)\, e^{\lambda\theta}\hat{u}(\lambda) + \varphi_0\right\} \quad \text{for } \gamma_0 \in N_\omega.
\end{aligned}
$$

Hence $\tilde{A} = A_\omega$, $B_1(A_\omega) = \int_{-h}^{0} \exp(\tilde{A}r)\, dM(r)$. In an analogous way we see that $\mathcal{O}_\omega = \tilde{\mathcal{O}}_\omega$. In fact

$$
\int_{-h}^{0} dL(s)\, e^{\lambda s}(\lambda I - \tilde{A})^{-1}\varphi(\cdot) = \mathscr{L}(\tilde{C}(S(t+s)P_\omega\varphi)) = \int_{-h}^{0} dC(s)\, e^{\lambda s} P_\omega (\lambda I - A_\omega)^{-1} P_\omega \varphi.
$$

Since $\tilde{A} = A_\omega$, we have that

$$
\left(\int_{-h}^{0} dL(s)\exp(\tilde{A}s)\right) = \left(\int_{-h}^{0} dC(s) P_\omega \exp(A_\omega s)\right) = \phi_\omega(\theta)\int_{-h}^{0} \exp(A_\omega^* s)\, dC_1^*(s).
$$

This finishes the proof. □

Let us recall that $m_\omega = \operatorname{rank} H_\omega$, $n_\omega = \dim N_\omega$, and put $\tilde{m}_\omega = \operatorname{rank} \tilde{H}_\omega$.

LEMMA 3.2. *$\tilde{m}_\omega = m_\omega$ (so that it does not depend on the realization of $T(\lambda)$), and $m_\omega \leqq n_\omega$.*

Proof. $m_\omega = \operatorname{rank} H_\omega = \operatorname{rank} \tilde{H}_\omega = \tilde{m}_\omega$. Hence $\tilde{m}_\omega$ is not going to change with the realization of $T(\lambda)$.

$$
m_\omega = \tilde{m}_\omega = \operatorname{rank} \tilde{H}_\omega = \operatorname{rank} \mathcal{O}_\omega \mathcal{R}_\omega \leqq n_\omega,
$$

since $\operatorname{rank} \mathcal{R}_\omega \leqq n_\omega$ and $\operatorname{rank} \mathcal{O}_\omega \leqq n_\omega$. □

If (S) is an ω-canonical realization of its transfer function, we have that

$$
\operatorname{rank} \mathcal{O}_\omega^* = n_\omega, \qquad \operatorname{rank} \mathcal{R}_\omega = n_\omega,
$$

so that $m_\omega = \operatorname{rank} \mathcal{O}_\omega \mathcal{R}_\omega = n_\omega$.

DEFINITION 3.1. *(S) is an ω-minimal realization of $T(\lambda)$* if the dimension of N_ω is as small as possible, i.e., it is m_ω. The above considerations imply that, if (S) is an ω-canonical realization of its transfer function, it is also an ω-minimal realization. The *ω-dimension* of (S) is the number $\dim N_\omega$.

THEOREM 3.1. *$T(\lambda)$ has an ω-canonical realization if and only if it has a realization with*

$$
n_\omega = \operatorname{rank} H_\omega.
$$

Proof. The necessity part has already been observed. Let (S) satisfy

$$
n_\omega = \dim N_\omega = \operatorname{rank} H_\omega = \operatorname{rank} \mathcal{O}_\omega \mathcal{R}_\omega.
$$

Since $\operatorname{rank} \mathcal{O}_\omega^* \leqq n_\omega$, $\operatorname{rank} \mathcal{R}_\omega \leqq n_\omega$, then $\operatorname{rank} \mathcal{O}_\omega \mathcal{R}_\omega \leqq n_\omega$, so that $n_\omega = \operatorname{rank} \mathcal{O}_\omega R_\omega \leqq n_\omega$, and $\operatorname{rank} \mathcal{O}_\omega^* = n_\omega$, $\operatorname{rank} \mathcal{R}_\omega = n_\omega$. Hence, (S) is an ω-canonical realization of its transfer function. □

Theorem 3.1 is a criterion which might be used to test whether a given realization of $T(\lambda)$ is canonical. Of course, in concrete examples, it is not easy to compute the number n_ω. However, Theorem 3.1 allows us to obtain some interesting consequences. First of all observe that $n_\omega = n = \dim x$ when $T(\lambda)$ has been obtained from a system which contains only input delays. Hence in this case it reduces to [13, Thm. 3.1]. Theorem 3.1. also gives some information about canonical realizations.

THEOREM 3.2. *A realization* (S) *of a transfer function* $T(\lambda) \in \mathcal{S}$ *is* ∞*-canonical if and only if its* ω*-dimension is equal to* m_ω *for each* ω. *In this case,* (S) *is* ω*-minimal for each* ω.

Another interesting consequence of Theorem 3.1 is the following:

THEOREM 3.3. *There exist systems like* (S) *whose transfer functions have no* ω*-canonical realization in the class of the systems whose dynamics are given by* (1a), (1b), *when* $\sigma_\omega \neq \varnothing$. *In particular, they have no* ∞*-canonical realization.*

Proof. If $\sigma_\omega \neq \varnothing$, the ω-dimension of (S) is at least 1, since a functional differential equation has only eigenvalues. Consider the system

$$\dot{x} = u(t) - u(t-1) \qquad y(t) = x(t). \tag{7}$$

Its transfer function $T(\lambda) = (1 - e^{-\lambda})/\lambda$ is analytic, so that $H_\omega = 0$ for every ω. Hence the transfer function of (7) has no ω-canonical realization for those ω such that $\sigma_\omega \neq \varnothing$. The spectrum of a functional differential equation is never empty. Hence numbers ω such that $\sigma_\omega \neq \varnothing$ always exist, so that $T(\lambda)$ cannot have ∞-canonical realizations. □

Remark 3.1. Let $T(\lambda)$ be given. Assume that $T(\lambda)$ has no ∞-canonical realization. In § 5 we shall show that it can be realized (in a class of systems different from $\mathcal{S}$) by a system which is "canonical" according to some suitable definition.

Remark 3.2. There are functions $T(\lambda)$ which have ω-canonical realizations for every $\omega > 0$, but which do not have ∞-canonical realizations. For example,

$$T(\lambda) = \sum_{k=0}^{\infty} \frac{1}{(\lambda - k^k)}.$$

Remark 3.3. One could guess that only transfer functions of the type

$$T(\lambda) = \sum_{k=1}^{\infty} L_k \lambda^{-k}$$

can have canonical realizations. The next example shows that this is not true.

$$T(\lambda) = \frac{1 - e^{-\lambda}}{\lambda} \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{1}{\lambda} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \frac{1}{\lambda^2} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

has the canonical realization

$$\dot{x}_1 = x_2(t) + u(t) - u(t-1), \qquad y_1(t) = x_1(t),$$
$$\dot{x}_2 = u(t), \qquad y_2(t) = x_2(t).$$

Remark 3.4. If $\omega_1 \leqq \omega_2$, then $m_{\omega_1} \leqq m_{\omega_2}$ for every realization (S) of $T(\lambda)$. Hence, if (S) is canonical, then from Theorem 3.1,

$$\dim H_{\omega_1} = m_{\omega_1} \leqq m_{\omega_2} = \dim H_{\omega_2}.$$

4. Minimality properties of canonical realizations. Recall that $n(S) = n = \dim x(t)$.

The following result is easily proved:

THEOREM 4.1. *Assume that*

$$\operatorname{rank} H_\omega = n$$

for large ω. *If* (S) *is a canonical realization of* $T(\lambda)$, *then* $n(S) = n$, *and* $n(S)$ *is as small as possible.*

Proof. It is clear that $\dim N_\omega = n$, for large ω. Hence σ, the spectrum of a canonical realization of $T(\lambda)$, is finite, i.e., $d(\lambda)$ is a polynomial of degree n. Hence, $n = \dim N_\omega = \dim x = n(S)$, and $n(S)$ is as small as possible, since $\dim N_\omega$ is minimal (Theorem 3.2.). □

Now we recall that a system is called spectrally minimal when the spectrum of the infinitesimal generator coincides with the set of nonanalyticity of $T(\lambda)$ (with the set of poles of $T(\lambda)$, in our case) [1]. The next theorem implies that a canonical realization of class $\mathcal{S}$ is spectrally minimal. In fact we can prove a stronger statement. Let $T(\lambda)$ be given. Let λ_0 be a pole of $T(\lambda)$. If $u_i(\lambda)$ are entire functions, we can write

$$T(\lambda)u_i(\lambda) = \frac{v_i}{(\lambda - \lambda_0)^{r_i}} + f_i(\lambda)$$

for some vector v_i, some entire number r_i and some function $f_i(\lambda)$ such that

$$\lim_{\lambda \to \lambda_0} f_i(\lambda)(\lambda - \lambda_0)^{(r_i \operatorname{sgn} r_i)} = 0.$$

Let $\mathcal{U} = \{u_1(\lambda), \cdots, u_s(\lambda)\}$ be a set of entire functions such that the corresponding vectors v_i are independent, and the numbers r_i are positive. Let $m(\mathcal{U})$ be the sum of the exponents r_i. We say that the number

$$m_T(\lambda_0) = \max_{\mathcal{U}} \{m(\mathcal{U})\}$$

is the order of the pole λ_0 of $T(\lambda)$ (compare [15, Ch. 2]).

THEOREM 4.2. *Let* (S) *be a canonical realization of* $T(\lambda)$. *If* λ_0 *is an eigenvalue of multiplicity* m_0, *then* m_0 *is equal to* $m_T(\lambda_0)$.

Proof. Let A_0 be the restriction of A to the generalized eigenspace of λ_0. We have

$$T(\lambda)u(\lambda) = \{\tilde{C}(\lambda I - A_0)P_{\lambda_0}\tilde{B} + \tilde{C}(\lambda I - A)(I - P_{\lambda_0})\tilde{B}\}\hat{u}.$$

$\tilde{C}(\lambda I - A)(I - P_{\lambda_0})\tilde{B}$ is bounded near λ_0. Hence we must study

$$T_0(\lambda) = \tilde{C}(\lambda I - A_0)P_{\lambda_0}\tilde{B},$$

which is the transfer function of a system without state delays, and with $\sigma(A_0) = \{\lambda_0\}$.

Let $v_1, \cdots, v_s$ be the eigenvectors of A_0. The equation

$$\dot{x} = A_0 x$$

has the solutions

$$x_i(t) = (v_i t^{r_i - 1} + p_i(t)) \exp(\lambda_0 t).$$

$p_i(t)$ is a polynomial of degree less than $r_i - 1$, and $\sum_{i=1}^{s} r_i = m_0$. Hence $\hat{x}_i(\lambda)$ has a pole in λ_0, of order r_i. We assume that the control system of transfer function $T_0(\lambda)$ is controllable, and we prove that, if $m_T(\lambda_0) < m_0$, then it cannot be observable, so that (S) is not a canonical realization of $T(\lambda)$. Since (S) is controllable, there exists a control such that $x(T, u) = p_i(0)$, for some T, and $u_T(\cdot) = 0$. Call $u_i(t)$ the extension of

this control which is zero for $t > T$. It gives an output

$$\hat{y}(\lambda) = T_0(\lambda)\hat{u}_i(\lambda) = C(\lambda)g(\lambda) + \frac{e^{\lambda T}C(\lambda)v_i}{(\lambda - \lambda_0)^{r_i}} + h^i(\lambda).$$

$g(\lambda)$ is the Laplace transform of the function

$$\begin{cases} x(t, u), & 0 \leqq t \leqq T, \\ 0, & t > T. \end{cases}$$

Hence $g(\lambda)$ and $u_i(\lambda)$ are entire functions. $h^i(\lambda)$ has a pole of order less than r_i. The vectors v_i are independent. If $m_T(\lambda_0) < m_0$, then for at least one index i, the corresponding $T_0(\lambda)u_i(\lambda)$ has a pole of order less than r_i, i.e. $C(\lambda_0)v_i = 0$. Since v_i is an eigenvector of A_0, we deduce that the system under study is not observable. □

Now we can study the question of the minimality of the number $n(S)$. We proved already that $n(S)$ is minimal, when

$$\dim H_\omega \leqq n$$

for every ω, without further assumptions.

In the following, we need an asymptotic estimate of the roots of $d(\lambda)$, which is given in [14], under the following special assumptions on $A(s)$:

Condition 1. $\bar{A}(s)$ is a.e. differentiable.

Condition 2. There exists a positive number γ such that $\bar{A}(r) = 0$ on $[-\gamma, 0]$.

THEOREM 4.3. *Let* (S) *be a canonical realization of* $T(\lambda)$. *If Conditions* 1 *and* 2 *hold, then* $n(S)$ *is minimal.*

Proof. Let (S') be a realization of $T(\lambda)$, with $n' = n(S') < n(S) = n$. Let $d(\lambda)$ and $d'(\lambda)$ be the characteristic polynomials of (S) and (S').

$$d'(\lambda) = \lambda^{n'} + \cdots + d'_0(\lambda).$$

(S) is canonical. Hence the poles of $T(\lambda)$ are the poles of $(d(\lambda))^{-1}$ with the same multiplicities. Hence $d'(\lambda)$ has at least all the zeros of $d(\lambda)$, with at least equal multiplicities. We can see that this is impossible.

Let us consider the zeros of $d(\lambda)$. We recall, from [14, Ch. 3], that there exists a finite number of algebraic equations

$$\sum_{k=0}^{s} E_k Z^k = 0 \tag{8}$$

such that:

a. $s \leqq n$, and $s = n$ for one of these equations.

b. Let $Z_1, \cdots, Z_s$ be the s roots of (8) (some of them may coincide). Hence, either $Z_1, \cdots, Z_s$, are roots of $d(s)$, or, for every $Z \in \{Z_1, \cdots, Z_s\}$, $d(s)$ has the following chain of roots:

$$z_p = -(\alpha) \ln \left| 2\pi p \left(\frac{\alpha}{Z} \right) \right| + i\alpha \left[\pm \left\{ 2p - \frac{1}{2} \operatorname{sgn} \alpha \right\} \pi + \arg Z \right] + f(p)$$

where $\lim_{p \to +\infty} f(p) = 0$.

α is a number which depends only on the equations that have been chosen, among those in (8). Hence there is a number α which corresponds to n chains of roots (which may not be distinct).

The same argument applied to the function $d'(\lambda)$ shows that the equation $d'(\lambda) = 0$ has at most n' chains of roots corresponding to the choice of any number α. Hence $d'(\lambda)$ cannot have all the roots as $d(\lambda)$, i.e., (S') is not a realization of $T(\lambda)$.

This finishes the proof. □

Remark 4.1. We conjecture that the above result can be proved without using Conditions 1 and 2.

5. Final remarks and comparison with previous results. Now we compare several different approaches to the problem of realization. In this section we never drop the symbol $\infty-$, and the word "canonical" will be used according to the definitions that will be specified later on.

First of all we consider the question raised in Remark 3.1. It is well known [3] that any system with delays can be written as an abstract system in a Hilbert space X

$$\dot{v} = Av + Bu, \qquad y = Cv, \tag{9}$$

where A is the generator of a strongly continuous semigroup of bounded operators $S(t)$ on X, while B, C are linear operators, $B : R^m \to X$, $C : X \to R^p$. C is bounded if there are no state delays, or if there are no discrete output delays. (If input delays act on the system, then $v(t)$ is somewhat more involved than $Z(t)$. See [3] for details.) Abstract systems in Hilbert spaces are (weakly) reachable when

$$\overline{\operatorname{span}}\, S(t)B = X$$

and (weakly) observable when $\overline{\operatorname{span}}\, S^*(t)C^* = X$. It is known [4] that (9) can be reduced to a canonical (i.e., weakly reachable and weakly observable) system without changing its transfer function. Now consider the following example: Let

$$T(\lambda) = \frac{e^{-\lambda} - e^{-2\lambda} - \lambda}{\lambda^2},$$

which is the transfer function of the system (S')

$$\dot{x} = \int_{-2}^{-1} u(t+s)\, ds - u(t), \qquad y = x(t). \tag{10}$$

$T(\lambda)$ is an entire function. Hence it has no ∞-canonical realization. However (S') may be written as a system in Hilbert space (in a very simple way, since it does not contain discrete delays), and an abstract canonical realization of $T(\lambda)$ (not as a system of class $\mathcal{S}$, of course) can be constructed. This example explains the observation in Remark 3.1.

The problem of realization of systems with delays has been previously approached via algebraic methods (see for example [8], [17], [18]). In the approach of [8] no canonicity question is raised, but methods to get minimal realizations are presented. This method can be shortly described as follows: We write (1a), (1b) as convolution equations,

$$p * x = A * x + B * u, \qquad y = C * X$$

where $*$ is the convolution product, p is the derivative of the δ function, A, B, C are matrices of distributions with support bounded on the left. Only a finite number of distributions $\Theta = \{\theta_1, \cdots, \theta_q\}$ need to be specified, so that the matrices A, B, C have elements in the smallest ring $R[\theta]$ which contains Θ and $R\delta$. The operational transfer function of this system is $W = C(pI - A)^{-1}B$ whose elements are in the smallest field which contains p and $R[\theta]$. For example, (7) can be written

$$p * x = (\delta - \delta_1) * x, \qquad y = \delta * x \quad (\delta_1 = \delta(t-1)),$$

and $W = (\delta - \delta_1)/p$. In [8] a method is given which provides decompositions of any

given W of the form $W = C(pI - A)^{-1}B$ with the size of A as small as possible. This method can be used to try to find ∞-canonical realizations as follows if a transfer function $T(\lambda)$ is given:

a. Write $W(p)$ (this can be done by inspection in simple cases, or by writing any realization of $T(\lambda)$, for example the one given in Theorem 1.1).

b. Use the method of [8] to construct minimal realizations. If we are lucky enough, this realization might be an ∞-canonical realization. We shall give an example later on. We note that at the moment we do not have any better method of constructing ∞-canonical realizations.

The approach of [17] can be applied when a system has only discrete delays, and no continuous delay. In this case, μ_i is the operator $(\mu_i x) = x(t - h_i)$: $C(-h_i, +\infty) \to C(0, +\infty)$. The system (S) can be identified by a triple of matrices (A, B, C) with elements in the ring of polynomials in the symbols μ_i, which is denoted $R[\mu]$. The pair (A, B) is reachable when

$$R[\mu] = \text{span}\,[B, AB, \cdots, A^{n-1}B].$$

System (S) is reachable when (A, B) is reachable, and coreachable or strongly observable when (A^*, C^*) is reachable [18], [9]. It is observable when

$$\bigcap_{i=0}^{n-1} \text{Ker}\, CA^i = \{0\}.$$

The transfer function of (S) is now defined to be

$$H(\lambda) = H(\lambda, \mu) = C(\lambda I - A)^{-1}B,$$

a rational matrix of λ, μ. (S) is a canonical realization when it is observable and reachable. A system which is reachable, is ∞-controllable, but a system which is observable, need not to be ∞-observable. (Again, this is an example of the situation referred to in Remark 3.1.) Since reachability and observability are not dual definitions, it seems that reachable and coreachable realizations are more interesting. Of course, a reachable and coreachable realization is ∞-canonical. Since it is of minimal rank, any algorithm that provides minimal realizations can be helpful investigating the existence of ∞-canonical realizations, exactly as above, but in a simpler way, although a given realization of $T(\lambda)$ can be minimal in the class of realizations with discrete delays, but not in the class $\mathcal{S}$ (observation due to E. Sontag).

We finish this paper with an example which illustrates the above arguments. We want to find an ∞-canonical realization of the transfer function

$$T(\lambda) = \frac{1 - e^{-\lambda} - \lambda^2 e^{-\lambda}}{\lambda^4}.$$

We can observe that $T(\lambda) = H(\lambda, e^{-\lambda})$, when

$$H(\lambda, \mu) = \frac{1 - \mu - \mu\lambda^2}{\lambda^4}$$

so that we can use the method of [17] to construct a realization with only one lag. In this case the Hankel matrix has rank 4, so that the rank of a minimal realization is 4. A realization of minimal rank is

$$\dot{x}_i = x_{i+1}, \qquad 1 \leqq i \leqq 3,$$

$$\dot{x}_4 = u,$$

$$y(t) = x_1(t) - x_1(t-1) - x_3(t-1),$$

which is not ∞-observable. At this point we cannot say that no ∞-canonical realization exists, because it might be possible to realize $T(\lambda)$ in a ring that is richer than the polynomials in one symbol. We observe that the method of [8] applied to the ring generated by δ, δ_1 gives a realization of rank 4. Now we note that

$$\frac{1-e^{-\lambda}}{\lambda}=\int_{-1}^{0} e^{\lambda s}\,ds$$

so that we can consider the operational transfer function

$$W(p)=\frac{g(s)}{p^3}-\frac{(\delta_1)}{p^2},$$

where $g(s)=1$ when $0\leqq s\leqq 1$, and zero otherwise. Observe that

$$\lambda^3 T(\lambda)=\langle p^3 W(p), e^{\lambda t}\rangle.$$

The Hankel matrix of $W(p)$ is now

$$H=\begin{bmatrix} 0 & -\delta_1 & g & 0 & 0 & \vdots & \vdots \\ -\delta_1 & g & 0 & 0 & \vdots & & \\ g & 0 & 0 & \vdots & & & \\ 0 & 0 & \vdots & & & & \\ \vdots & \vdots & & & & & \end{bmatrix}$$

which has rank 3. Hence $W(p)$ admits a realization with $n(S)=3$. Using known algorithms it is easy to find the realization

$$\dot{x}_1=u,\quad \dot{x}_2=x_1,\quad \dot{x}_3=x_2,$$

$$y(t)=-x_2(t-1)+\int_{-1}^{0} x_3(t+s)\,ds,$$

and we were lucky enough to find an ∞-canonical realization.

Remark 5.1. Let us observe that the minimality of a realization as defined in § 4 is a property of the transfer function, while the minimality of the realization of a transfer function over a ring is a property of the transfer function and of the ring, which is not uniquely determined by the transfer function.

Acknowledgment. The author thanks the referees for the careful reading of this paper. They discovered a mistake in the proof of Theorem 1.1. and suggested the introduction of § 5.

REFERENCES

[1] J. S. BARRAS, R. W. BROCKETT AND P. A. FURHMAN, *State space models for infinite-dimensional systems*, IEEE Trans. Autom. Cont., AC 19 (1974), pp. 693–700.

[2] R. W. BROCKETT, *Finite Dimensional Linear Systems*, John Wiley, New York, New York, 1970.

[3] M. C. DELFOUR, *Status of the state space theory of linear hereditary differential systems with delays in state and control variables*, in Analysis and Optimization of Systems, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, 1980, pp. 83–96.

[4] M. C. DELFOUR AND S. K. MITTER, *Controllability and observability for infinite dimensional systems*, this Journal, 10 (1972), pp. 329–333.

[5] P. A. FURHMAN, *Algebraic system theory: an analyst's point of view*, J. Franklin Institute, 301 (1976), pp. 521–540.

[6] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, Berlin, 1977.
[7] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, New York, 1969.
[8] E. W. KAMEN, *On an algebraic theory of systems defined by convolution operators*, Math. Systems Theory, 9 (1975), pp. 57–74.
[9] P. P. KHARGONEKAR, *On matrix fraction representations for linear systems over commutative rings*, Center for Math. System Theory, Univ. of Florida, July 1980.
[10] T. KATO, *Perturbation Theory for Linear Operators*, Die Grundlehren der Math. Wissensch. in Einzeld., Springer-Verlag, Berlin, 1966.
11. A. W. OLBROT, *Stabilizability, detectability and spectrum assignment for linear systems with general time delays*, IEEE Trans. Automat. Cont., 23 (1978), pp. 887–890.
[12] L. PANDOLFI, *On feedback stabilization of functional differential equations*, Boll. Un. Mat. Ital., 11 (1975), pp. 626–635.
[13] ———, *Canonical realizations of systems with delayed controls*, Ricerche di Automatica, 10 (1980), pp. 27–37.
[14] E. PINNEY, *Ordinary Difference-Differential Equations*, University of California Press, Berkeley and Los Angeles, CA, 1958.
[15] H. H. ROSENBROCK, *State-Space and Multivariable Theory*, Nelson, London, 1970.
[16] D. SALAMON, *Observers and duality between observation and state feedback for time delay systems*, Universität Bremen, Forscungsschwerpunkt Dynamische Systeme, Rept. 5, 1979.
[17] E. D. SONTAG, *Linear systems over commutative rings: a survey*, Ricerche di Automatica, 7 (1977), pp. 1–34.
[18] ———, *On split realizations of response maps over rings*, Inform. and Control, 37 (1978), pp. 23–33.
[19] A. SVESHNIKOV AND A. TIKHONOV, *The Theory of Functions of a Complex Variable*, MIR Publishers, Moscow, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

0363-0129/83/2104-0007 $01.25/0

ADMISSIBLE INPUT ELEMENTS FOR SYSTEMS IN HILBERT SPACE AND A CARLESON MEASURE CRITERION*

L. F. HO† AND D. L. RUSSELL‡

Abstract. We study the control system

$$\dot{x} = Ax + bu, \qquad x \in X, \quad u \text{ scalar},$$

where A generates a semigroup on the Hilbert space X, but, in general, the control input element $b \notin X$. Many boundary value control systems, point control force situations, etc., can be studied in this context. We define and analyze "admissible" input elements b and develop sufficient conditions for b to be admissible in terms of the Carleson measure theorem of H^p-theory.

Key words. distributed parameter control, infinite dimensional systems, unbounded control elements, theory of distributions, Carleson measure

1. Introduction. One commonly studies linear, time invariant control systems in a Banach space X in the form

$$\dot{x} = Ax + Bu, \qquad x \in X, \quad u \in U, \tag{1.1}$$

where A is the generator of a strongly continuous semigroup of bounded operators $\{S(t) | t \geqq 0\}$ on X, and B is a bounded operator from the control space, U, into X. If $u : [0, \infty) \to U$ is locally (Bochner) integrable, generalized (or "mild") solutions of (1.1) corresponding to an initial state

$$x(0) = x_0 \in X$$

can be represented by the "variation of parameters" formula (see, e.g., [3], [11])

$$x(t) = S(t)x_0 + \int_0^t S(t-s)Bu(s)\, ds, \tag{1.2}$$

and a number of properties of $x(t)$ can thereby be deduced.

It is well known, however, that most of the interesting infinite dimensional control systems do not arise this way, because the degree of controllability of a system (1.1) with B bounded is rather restricted if, as is usually the case, U is finite dimensional or for some other reason the operator B is compact. Indeed, most of the mathematically intriguing examples arise in the context of partial differential equations with boundary value control inputs, control forces exerted at isolated points, etc., and in the context of functional equations which involve values of the control of discrete instants, viz. $u(t), u(t-T_1), \cdots, u(t-T_n)$. In each of these cases the formulation (1.1) is inadequate and one must consider input operators B whose range is not restricted to the space X.

A number of authors have addressed the problem of interpretation of (1.1) for operators B of rather general type. We particularly cite the contributions of Curtain and Pritchard [3], Zabczyk [22], Fattorini [6], and Washburn [20]. It seems fair to say that, as brought out in [3], the theory is more extensive and generally applicable in the case of systems of "diffusion type", ordinarily involving holomorphic semigroups, than in systems of "wave" or hyperbolic character.

In the present article we shall restrict our attention to spaces X which are separable Hilbert spaces and to finite dimensional control spaces U. Taking U to be R^m, (1.1)

* Received by the editors March 31, 1982. This research was supported in part by the Air Force Office of Scientific Research under grant AFOSR 79-0018.

† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019.

‡ Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

becomes

$$\dot{x} = Ax + \sum_{j=1}^{m} b_j u^j \tag{1.3}$$

where b_j is the control input element associated with the jth control component u^j. Since every solution of (1.3) is a linear combination of solutions of $x = Ax$ and the individual systems $x = Ax + b_j u^j$, $j = 1, 2, \cdots, m$, we may, without loss of generality, confine our discussion to systems

$$\dot{x} = Ax + bu \tag{1.4}$$

wherein the control u is scalar valued. Much of our theory can be extended to cases wherein U is infinite dimensional, but we will not do that here.

What distinguishes the present study from earlier contributions is the attention which we pay not only to the relationship between the operator, A, and the input element, b, but also to the relationship between b and the semigroup $S(t)$ generated by A. In cases where A has discrete spectrum $\{\lambda_k | k \in K\}$, K being a countable index set, this amounts to a study encompassing the input element b, the eigenvectors $\{\phi_k | k \in K\}$ of A, the corresponding eigenvectors of the dual operator, A', as defined in § 2, and the exponential functions $\exp(\lambda_k t)$, $k \in K$. It is in particular reference to the latter that what is probably the most important idea of this paper is developed. We show that a sufficient condition for b to be an "admissible input element" (definition in § 2) can be given in terms of a measure on Borel subsets of the complex plane whose support is $\{-\lambda_k | k \in K\}$. When that measure turns out to be a *Carleson measure* the input element b is admissible. This result brings out yet again the intimate relationship between the control theory of infinite dimensional linear systems and parallel developments in H^p theory [5], [8], [12] and the related theory of completeness and independence of sets of complex exponentials.

2. Admissible input elements. Let X be a separable Hilbert space and let A be a closed operator on X with domain, $\mathcal{D}(A)$, dense in X, generating a strongly continuous semigroup of bounded operators $S(t)$ on X for $t \geqq 0$. For $b \in X$ the (generalized, or mild) solution of

$$\dot{x} = Ax + bu, \qquad u \in L^2_{\text{loc}}[0, \infty), \tag{2.1}$$

$$x(0) = x_0 \in X, \tag{2.2}$$

is given by the "variation of parameters" formula

$$x(t) = S(t)x_0 + \int_0^t S(t-s)bu(s)\, ds, \tag{2.3}$$

and may be seen to be a continuous function $x : [0, \infty) \to X$. Whether $\dot{x}(t)$ is defined for each $t \geqq 0$ and (2.1) holds is more complicated: sufficient conditions are that $b \in \mathcal{D}(A)$ or that u is differentiable as a function of t [3], [11].

In this paper we wish to consider (2.1), (2.2) in certain cases where b does not lie in X and to provide, for such b, a formula parallel to (2.3). Our approach is similar to that used in [14].

Identifying X with its dual X', we denote the duality relationship by $\langle x, y \rangle$, $x \in X$, $y \in X$, linear in both x and y. Where X is the complexification of a real Hilbert space X_0 the conjugate element $\bar{y}$ is well defined for each $y \in X$ and, with $(\cdot, \cdot)$ denoting

the inner product in X,

$$(x, y) = \langle x, \bar{y} \rangle, \langle x, y \rangle = (x, \bar{y}).$$

The bilinear form $\langle \cdot, \cdot \rangle$ is symmetric, i.e., $\langle x, y \rangle = \langle y, x \rangle$, $x, y \in X$, and, for all $x \in X$,

$$\|x\|_X = \sup_{\substack{y \in X \\ y \neq 0}} \frac{|\langle x, y \rangle|}{\|y\|_X}. \tag{2.4}$$

The symbol A' will be used to denote the dual of A relative to the bilinear form $\langle \cdot, \cdot \rangle$, that is,

$$\langle Ax, y \rangle = \langle x, A'y \rangle, \qquad x \in \mathcal{D}(A), \quad y \in \mathcal{D}(A').$$

The operator A' is closed with domain $\mathcal{D}(A')$ dense in X. It is known that if A generates a semigroup $S(t)$, then $S(t)'$ is also a semigroup generated by A'. See [4] for details.

Let Y be a dense subspace of X which is a Hilbert space in its own right with norm $\|\cdot\|_Y$ stronger than $\|\cdot\|_X$ so that the injection map

$$j = Y \to X, \qquad j(y) = y, \quad y \in Y,$$

is one-to-one and continuous with dense range $Y \subset X$. We further suppose that Y is invariant under the action of $S(t)' : y \in Y \Rightarrow S(t)'y \in Y$, and that this map is continuous with respect to $\|S(t)'y\|_Y$, $\|y\|_Y$ and the usual topology of $[0, \infty)$.

Let Y' be the dual of Y with respect to X as described, e.g. in [1], [14], [15]. This means that Y' is the closure of X with respect to the norm

$$\|x\|_{Y'} = \sup_{\substack{y \in Y \\ y \neq 0}} \frac{|\langle x, y \rangle|}{\|y\|_Y}. \tag{2.5}$$

It is known that Y', so defined, is a realization of the dual space of Y, and it is easily verified that the bilinear form $\langle x, y \rangle$ may be defined, by continuity, for $x \in Y'$, $y \in Y$ as

$$\langle x, y \rangle = \lim_{k \to \infty} \langle x_k, y \rangle$$

where $\{x_k\}$ is a sequence in X converging to x in $\|\cdot\|_{Y'}$. So defined, $\langle x, y \rangle$ generates, as x ranges over Y', all continuous linear functionals on Y. We have

$$X \subset X \subset Y'.$$

DEFINITION 2.1. In the system (2.1), i.e.,

$$\dot{x} = Ax + bu, \qquad u \in L^2_{\text{loc}}[0, \infty),$$

b is an *admissible input element* if there exist Y. Y', as above, with $b \in Y'$, such that for every $T > 0$ the continuous map

$$L_T : Y \to C[0, T]$$

defined by

$$(Ly)(t) = \langle b, S(t)'y \rangle, \qquad y \in Y, \quad t \in [0, T], \tag{2.6}$$

has a continuous extension to

$$L_T : X \to L^2[0, T].$$

Remark. It is clear that this amounts to the statement that in the dual *observed system*

$$\dot{y} = A'y, \qquad z = \langle b, y\rangle,$$

b is an *admissible observation element*; that is, for $y \in Y$,

$$z(\cdot) = \langle b, S(\cdot)'y\rangle \in C[0, T],$$

this relationship extending continuously to $z(\cdot) \in L^2[0, T]$ for $y \in X$.

To verify that Definition 2.1 enables consistent definition, at least in a generalized sense, of solutions of (2.1), (2.2) when b is an admissible input element and to establish some of the properties of the resulting solution, we present

THEOREM 2.2. *If b is an admissible input element, the formula*

$$\langle x(t), y\rangle = \langle x_0, S(t)'y\rangle + \int_0^t \langle b, S(t-s)'y\rangle u(s)\, ds, \qquad y \in Y, \tag{2.7}$$

defines, for each $t \geqq 0$, a unique element $x(t) \in X$. Given $T > 0$ and $u \in L^2[0, T]$,

$$x(t) = S(t)x_0 + B(t)u, \qquad t \in [0, T], \tag{2.8}$$

where $B(t)$ is the strongly continuous family of bounded operators $B(t): L^2[0, T] \to X$ given by

$$\langle B(t)u, y\rangle = \int_0^t \langle b, S(t-s)'y\rangle u(s)\, ds, \qquad y \in Y. \tag{2.9}$$

Proof. From (2.8) and the fact that Y is dense in X, it is clear that

$$x(t) - S(t)x_0 \equiv \xi(t) \equiv B(t)u$$

where, for $y \in Y$,

$$\langle \xi(t), y\rangle = \int_0^t \langle b, S(t-s)'y\rangle u(s)\, ds.$$

Let $x \in X$ and let $\{y_k\}$ be a sequence in Y converging to x with respect to $\|\cdot\|_X$. Since b is an admissible input element, the corresponding functions h_k defined by

$$h_k(t-s) = \langle b, S(t-s)'y_k\rangle \tag{2.10}$$

converge in $L^2[0, T]$ to a function $h \in L^2[0, T]$. Defining

$$\langle \xi(t), x\rangle = \int_0^t h(t-s)u(s)\, ds,$$

we see that for $t \in [0, T]$

$$\|\langle \xi(t), x\rangle\| \leqq \|h\|_{L^2[0,T]}\|u\|_{L^2[0,T]} \leqq \|L_T\|\|x\|_X\|u\|_{L^2[0,T]}$$

since (cf. (2.6), (2.10)) $h = L_T x$. Hence $\xi(t) \in X' = X$. This also gives

$$\|\xi(t)\|_X \leqq \|L_T\|\|u\|_{L^2[0,T]},$$

showing that for $t \in [0, T]$, $B(t)$ is bounded with

$$\|B(t)\| \leqq \|L_T\|.$$

To establish that $\xi(t)$ is continuous in t for each fixed $u \in L^2[0, T]$ (and, hence, that $B(t)$ is strongly continuous in t), let $0 \leqq t \leqq \hat{t} \leqq T$ and form, for $y \in Y$,

$$\begin{aligned}\langle \xi(\hat{t}) - \xi(t), y\rangle &= \int_0^{\hat{t}} \langle b, S(t-s)'y\rangle u(s)\, ds - \int_0^t \langle b, S(t-s)'y\rangle u(s)\, ds \\ &= \int_0^t \langle b, S(t-\tau)y\rangle u(\tau + (\hat{t}-t))\, d\tau - \int_0^t \langle b, S(t-s)'y\rangle u(s)\, ds \\ &\quad + \int_0^{\hat{t}-t} \langle b, S(t-s)'y\rangle u(s)\, ds \qquad (\text{with } \tau = s - (\hat{t}-t)) \\ &= \int_0^t \langle b, S(t-s)'y\rangle (u(s + (t-t)) - u(s))\, ds \\ &\quad + \int_0^{\hat{t}-t} \langle b, S(t-s)'y\rangle u(s)\, ds \\ &\leqq \|L_T\| \|y\|_X (\|u(\cdot + (\hat{t}-t)) - u\|_{L^2[0,t]} + \|u\|_{L^2[0,\hat{t}-t]}).\end{aligned}$$

Since Y is dense in X and since for fixed $u \in L^2[0, T]$ we have

$$\lim_{\hat{t}-t \to 0} \|u\|_{L^2[0,\hat{t}-t]} = 0,$$

$$\lim_{t \uparrow \hat{t}} \|u(\cdot + (t-t)) - u\|_{L^2[0,t]} = \lim_{\hat{t} \downarrow t} \|u(\cdot + (\hat{t}-t)) - u\|_{L^2[0,t]} = 0.$$

We conclude that for fixed $u \in L^2[0, T]$, and t, $\hat{t}$ as described,

$$\lim_{t \uparrow \hat{t}} \|\xi(\hat{t}) - \xi(t)\|_X = \lim_{\hat{t} \downarrow t} \|\xi(\hat{t}) - \xi(t)\|_X = 0,$$

and thus $\xi(t)$ is continuous in X. This completes the proof of the theorem.

Let H be a separable Hilbert space and let $\{p_k | k \in K\}$ be a sequence in H, K being a countable ordered index set. The p_k are strongly independent if no p_k lies in the closed span of $\{p_l | l \neq k\}$. If, in addition, there is a positive number c such that whenever

$$p = \sum_{K_0} \alpha_k p_k, \tag{2.11}$$

the α_k being complex and K_0 an arbitrary finite subset of K, we have

$$\sum_{K_0} |\alpha_k|^2 \leqq c^2 \|p\|_H^2, \tag{2.12}$$

we say that the p_k are *uniformly l^2-independent*, since (2.12) implies

$$\sum_{K} |\alpha_k|^2 \leqq c^2 \|p\|_H^2 \tag{2.13}$$

whenever $\{\alpha_k\} \in l^2$ and $p = \sum_K \alpha_k p_k$ is convergent in H.

If there is a positive number C such that

$$\|p\|_H^2 \leqq C^2 \sum_{K_0} |\alpha_k|^2,$$

p as in (2.11), we say that the sequence $\{p_k\}$ is *uniformly l^2-convergent* since this property implies that if $\{\alpha_k\} \in l^2$ the series $\sum_K \alpha_k p_k$ is convergent in H and

$$\|p\|_H^2 \leqq C^2 \sum_{K} |\alpha_k|^2. \tag{2.14}$$

Recall that a sequence $\{p_k\}$ in H forms a Schauder basis for H if for every $p \in H$ there are unique coefficients α_k such that the series $\sum \alpha_k p_k$ converges to p in H [21]. A Schauder basis which is, at the same time, both uniformly l^2-independent and uniformly l^2-convergent is a *Riesz basis.* For evident reasons we shall also use, synonymously, the term uniform l^2-basis. If $\{p_k\}$ is a *uniform l^2-basis* for H then every p in H has a unique convergent representation

$$p = \sum_K \alpha_k p_k$$

with (cf. (2.13), (2.14))

$$c^{-2} \sum_K |\alpha_k|^2 \leqq \|p\|_H^2 \leqq C^2 \sum_K |\alpha_k|^2.$$

For the remainder of this section we suppose that:

(i) The operator A with dense domain $\mathcal{D}(A) \subseteq X$ generates the strongly continuous semigroup of bounded operators $S(t)$, $t \geqq 0$;

(ii) $\sigma(A)$, the spectrum of A, consists of discrete, simple eigenvalues λ_k, $k \in K$, and the corresponding normalized eigenvectors ϕ_k, $k \in K$, form a strongly independent, uniformly l^2-convergent Schauder basis for X.

Since the ϕ_k, $k \in K$, are strongly independent and have closed span equal to X, there exist unique biorthogonal elements ψ_k, $k \in K$, such that

$$\langle \psi_k, \phi_l \rangle = \begin{cases} 1, & k = l, \\ 0, & k \neq l, \end{cases} \quad k, l \in K.$$

As is well known, the ψ_k are eigenvectors of the dual operator A' corresponding to the eigenvalues λ_k, $k \in K$. We further assume that:

(iii) The eigenvectors ψ_k of A' have the property

$$\psi_k \in Y \subset X$$

(this is true, for example, if $Y \supset \mathcal{D}((A')^r)$ for some positive integer r).

If $x \in X$, the fact that the ϕ_k form a Schauder basis in X implies the existence of unique ξ_k, $k \in K$, such that

$$x = \sum_K \xi_k \phi_k, \tag{2.15}$$

the series converging in X. From this it is evident that

$$\xi_k = \langle \psi_k, x \rangle, \qquad k \in K.$$

We are not assured, in general, that the ξ_k are square summable, but the uniform l^2-convergence property of the ϕ_k shows the square summability of the sequence $\{\xi_k\}$ to be a sufficient condition for convergence of (2.15).

Since we assume the ψ_k lie in Y, given any element $b \in Y'$ (and this includes $b \in X$) we may define

$$b_k = \langle \psi_k, b \rangle \tag{2.16}$$

and obtain a set of coefficients b_k, $k \in K$, associated with b. In general it is not possible to recover b from the coefficients b_k. (An example is $X = L^2[0, 2\pi]$, $Y = H^1[0, 2\pi]$, $\psi_k(x) = (2\pi)^{-1} e^{ikx}$, $k = 0, 1, 2, \cdots$. The ψ_k $(= \phi_k)$ here form an orthonormal basis for X and belong to Y but there is a nonzero element, namely $\delta_{(0)} - \delta_{(2\pi)}$, in Y' for which all of the b_k are zero. This arises, of course, because the closed span of the ψ_k in Y is not equal to Y.) As a consequence it is not generally meaningful to write $b = \sum_K b_k \phi_k$.

Nevertheless it may be meaningful to consider the initial value problem (2.1), (2.2), i.e.,

$$\dot{x} = Ax + bu, \quad x(0) = x_0 \in X, \quad u \in L^2_{\text{loc}}[0, \infty),$$

for certain $b \in Y'$, namely, those that we have already characterized as admissible input elements. We wish now to show that the class of such admissible input elements can be characterized in terms of the coefficients b_k and the eigenvalues λ_k. If $x(t)$ is the solution of (2.1), (2.2) established by Theorem 2.2 for an admissible input element b, then, in particular, for $t \geqq 0$,

$$\langle x(t), \psi_k \rangle = \langle x_0, S(t)'\psi_k \rangle + \int_0^t \langle b, S(t-s)'\psi_k \rangle u(s)\, ds = e^{\lambda_k t} x_{0,k} + \int_0^t e^{\lambda_k (t-s)} u(s)\, ds$$

where

$$x_0 = \sum_K x_{0,k} \phi_k.$$

We do not know that the numbers $e^{\lambda_k t} x_{0,k}$ are square summable, but the series

$$\sum_K e^{\lambda_k t} x_{0,k} \phi_k$$

must converge to $S(t)x_0$ by virtue of the (assumed) Schauder basis property of the ϕ_k. It follows that a *sufficient* condition for $x(t)$ to belong to X is that the numbers

$$\zeta_k(t) = b_k \int_0^t e^{\lambda_k (t-s)} u(s)\, ds \tag{2.17}$$

should be square summable for each $t \geqq 0$. Equivalently, making a trivial change of independent variable,

$$\zeta_k(t) = b_k \int_0^t e^{\lambda_k s} f(s)\, ds, \qquad f(s) = u(t-s).$$

The necessity of considering an infinite number of values of t can be obviated by taking f to be an element of $L^2[0, T]$, $T > 0$ fixed, and defining $f(s) \equiv 0$ in $[t, T]$ for $t < T$. The map

$$\zeta_k = b_k \int_0^T e^{\lambda_k s} f(s)\, ds, \qquad f \in L^2[0, T], \tag{2.18}$$

so defined may be designated as

$$L'_T : L^2[0, T] \to X, \tag{2.19}$$

$$L'_T(f) = x = \sum_K \zeta_k \phi_k, \tag{2.20}$$

and it is easy to see that L'_T is the dual of $L_T : X \to L^2[0, T]$ as defined by (2.6). Thus the boundedness of L_T, as required in Definition 2.1, may be obtained as an immediate corollary if it is shown that L'_T, defined by (2.18)–(2.19), is bounded. For our present purpose this is the route of choice.

Extending f further via $f(t) = 0$, $t > T$, the Laplace transform of f is the entire function

$$\phi(z) = \int_0^\infty e^{-zt} f(t)\, dt = \int_0^T e^{-zt} f(t)\, dt.$$

In terms of ϕ we clearly have

$$\zeta_k = b_k \phi(-\lambda_k), \qquad k \in K,$$

and the following proposition is evident.

PROPOSITION 2.3. *The operator L_T (equivalently L_T') is bounded just in case, for every $f \in L^2[0, T]$ the Laplace transform of f, ψ, has the property*

$$\sum_K |b_k \phi(-\lambda_k)|^2 < \infty. \tag{2.21}$$

We are fortunate that the inequality can often be established with the use of the concept of a *Carleson measure* and the corresponding Carleson measure theorem as it applies to the space

$$H_\alpha^2 \equiv H^2\{z | \mathrm{Re}\,(z) > \alpha\}, \qquad \alpha \text{ real}. \tag{2.22}$$

The space $H^2\{z|\mathrm{Re}\,(z)>\alpha\}$ consists of those complex functions $\phi(z)$, analytic in $\mathrm{Re}\,(z)>\alpha$, bounded in each half plane $\mathrm{Re}\,(z) \geqq \alpha+\delta$, $\delta>0$, and satisfying

$$\int_{-\infty}^{\infty} |\phi(\xi+i\eta)|^2\, d\eta \leqq M_\phi, \qquad \xi > \alpha, \tag{2.23}$$

where M_ϕ is a positive number depending only on ϕ (and not, in particular, on ξ). It is known (see, e.g., [10]) that each such function has a limiting "boundary" function

$$\phi_\alpha(\eta) = \lim_{\xi \downarrow \alpha} \phi(\xi + i\eta) \tag{2.24}$$

defined almost everywhere in $-\infty < \eta < \infty$ and $\phi_\alpha(\eta)$ is measurable with

$$\int_{-\infty}^{\infty} |\phi_\alpha(\eta)|^2\, d\eta \leqq M_\phi.$$

Each $\phi \in H_\alpha^2$ is the Laplace transform of a unique function $f \in L^2_{\mathrm{loc}}[0, \infty)$ such that

$$\int_0^{\infty} |e^{-\alpha t} f(t)|^2\, dt < \infty.$$

Let μ be a (nonnegative valued) measure defined on the Borel subsets of $\{z|z>\alpha\}$. Then μ is a Carleson measure if for every real τ and every $h>0$

$$\mu(\{z | \tau - h \leqq \mathrm{Im}\,(z) \leqq \tau + h,\ \alpha < \mathrm{Re}\,(z) \leqq \alpha + h\}) \leqq Ah \tag{2.25}$$

for some positive A depending only on μ (not on h).

For a Carleson measure we have

THEOREM 2.4. *If μ is a Carleson measure on $\{z|\mathrm{Re}\,(z)>\alpha\}$ with A as in* (2.25), *if $\phi \in H_\alpha^2$, and ϕ_α is given by* (2.24), *then*

$$\int_{\{z|\mathrm{Re}\,(z)>\alpha\}} |\phi(z)|^2\, d\mu(z) \leqq \frac{1{,}000A}{\pi^2} \int_{-\infty}^{\infty} |\phi_\alpha(\eta)|^2\, d\eta. \tag{2.26}$$

A proof of this theorem is offered, for the sake of completeness, in § 4 of this paper. The relevance of this theorem for our present studies is exhibited in the selection of a particular measure μ. For $b \in Y'$ and a given discrete spectrum $\{\lambda_k\}$ for A, let

$$\mu = \mu_{b,\{\lambda_k\}}$$

be defined by

$$\mu(-\lambda_k) = |b_k|^2, \qquad k \in K, \tag{2.27}$$

$$\mu(\{z \mid \mathrm{Re}\,(z) > \alpha\} - \{\lambda_k \mid k \in K\}) = 0. \tag{2.28}$$

In this case the left-hand side of (2.26) becomes (cf. (2.21))

$$\sum_K |b_k \phi(-\lambda_k)|^2.$$

The Plancherel theorem, on the other hand, gives

$$\int_{-\infty}^{\infty} |\phi_\alpha(\eta)|^2 \, d\eta = 2\pi \int_0^{\infty} |e^{-\alpha t} f(t)|^2 \, dt \leqq 2\, e^{2|\alpha|T} \pi \int_0^T |f(t)|^2 \, dt$$

when the support of f is restricted to $[0, T]$. Thus

$$\sum_K |b_k \phi(-\lambda_k)|^2 \leqq 2{,}000\, e^{2|\alpha|T} \frac{A}{\pi} \int_0^T |f(t)|^2 \, dt$$

and, in view of our earlier discussion, we have

COROLLARY 2.5. *A sufficient condition in order that $b \in Y'$ should be an admissible input element for the system* (2.1), *wherein $\sigma(A) = \{\lambda_k \mid k \in K\}$ and the corresponding eigenvectors ϕ_k, $k \in K$, form a strongly independent, uniformly l^2-convergent Schauder basis for X, is that the measure $\mu_{b,\{\lambda_k\}}$ defined by* (2.27), (2.28) *should be a Carleson measure in $\{z \mid \mathrm{Re}\,(z) > \alpha\}$ for some real α.*

We remark that the assumption (i) above together with the Hille–Yoshida theorem [4], [11] implies that the complex numbers $-\lambda_k$, $k \in K$, are, indeed, confined to some right half plane $\mathrm{Re}\,(z) > \alpha$. The fact that the support of f is restricted to $[0, T]$ implies that the corresponding Laplace transform ϕ is entire and satisfies an inequality (2.23) for every real α ($M_\phi = M_{\phi,\alpha}$ here).

3. Identification of admissible and inadmissible input elements; examples. Our first task in this section will be to develop a method whereby input elements b not in the state space X may be identified as particular elements of a larger space Y'. The assumptions made will be somewhat more restrictive than those introduced in § 2. They are by no means necessary conditions.

Let us suppose that the operator A, generating a strongly continuous semigroup $S(t)$ on the Hilbert space X, has (dense) domain $\mathcal{D}(A)$ and that A possesses discrete eigenvalues λ_k, $k \in K$, with

$$\lim_{\rho(k) \to \infty} |\lambda_k| = \infty.$$

Here $\rho(k)$ denotes the number of elements $l \in K$ such that $l < k$ with respect to the assumed order relation on K. The corresponding normalized eigenvectors ϕ_k are assumed to form a uniform basis for X. We denote the dual operator by A'. It has the same eigenvalues λ_k, and the corresponding eigenvectors ψ_k, $k \in K$, will be assumed normalized so that

$$\langle \psi_k, \phi_l \rangle = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases}$$

The ψ_k also form a uniform basis for X, as is well known. Then it is easy to see that

$$\mathcal{D}(A) = \left\{ y = \sum_K x_k \phi_k \,\middle|\, \sum_K |\lambda_k x_k|^2 < \infty \right\}$$

and that

$$\mathcal{D}(A') = \left\{ y = \sum_K y_k \psi_k \,\middle|\, \sum_K |\lambda_k y_k|^2 < \infty \right\}.$$

For the work of this section we take $Y = \mathcal{D}(A')$ with the graph norm

$$\|y\|_Y^2 = \sum_K (1 + |\lambda_k|^2)|y_k|^2,$$

where

$$y = \sum_{k=1}^{\infty} y_k \psi_k$$

in X. Then $Y \subset X$ and the injection mapping is continuous. It will often be possible to identify a Hilbert space $Z \subset X$ with continuous injection map such that $\|\cdot\|_Z$ is a familiar (e.g. Sobolev) norm and Y is a closed subspace of Z on which the norms $\|\cdot\|_Z$ and $\|\cdot\|_Y$ are equivalent.

We will be concerned with two different extensions of the operator A. We suppose first of all that there is an element $\hat{x} \in X$ not in $\mathcal{D}(A)$ and that L is an operator on X such that

$$\mathcal{D}(L) = \{\xi + u\hat{x} \,|\, \xi \in \mathcal{D}(A),\ u \text{ scalar}\}, \qquad Lx = Ax, \quad x \in \mathcal{D}(A).$$

We will refer to L as an "operational extension" of A. Its significance arises from the fact that many of the inhomogeneous boundary value problems arising in applications can be expressed in the form

$$\frac{dx}{dt} = Lx, \tag{3.1}$$

with the restriction

$$x = \xi + u\hat{x} \in \mathcal{D}(L). \tag{3.2}$$

The second extension of A, which is a map

$$\hat{A} : X \to Y',$$

is a standard one, often used, e.g. in [14]. If $y, \eta \in \mathcal{D}(A), \mathcal{D}(A')$, respectively, we have

$$\langle Ay, \eta \rangle = \langle y, A'\eta \rangle.$$

Since $A' : Y = \mathcal{D}(A') \to X$ is continuous, the form $\langle y, A'\eta \rangle$ extends to $\langle x, A'\eta \rangle$, $x \in X$, by continuity and density of $\mathcal{D}(A)$ in X and, so extended, $\langle x, A'\eta \rangle$ defines, for each fixed $x \in X$, a continuous linear functional on Y, i.e., an element of Y'. We define

$$\hat{A} : X \to Y' = (\mathcal{D}(A'))'$$

by

$$\langle \hat{A}x, \eta \rangle = \langle x, A'\eta \rangle, \qquad x \in X, \quad \eta \in Y = \mathcal{D}(A').$$

Our first goal, with reference to the system (3.1), (3.2), is to replace it by an infinite set of scalar ordinary differential equations

$$\frac{dx_k}{dt} = \lambda_k x_k + b_k u, \qquad k \in K, \tag{3.3}$$

where

$$x(t) = \sum_{k \in K} x_k(t)\phi_k,$$

convergent in X. In order to do this we recognize first of all that

$$z = \sum_{K} \lambda_k x_k \phi_k$$

represents not Lx, but rather $\hat{A}x$, since

$$\langle \hat{A}x, \psi_k \rangle = \langle x, A'\psi_k \rangle = \langle x, \lambda_k, \psi_k \rangle = \lambda_k x_k.$$

We rewrite (3.1) in the form

$$\frac{dx}{dt} = \hat{A}x + Lx - \hat{A}x, \tag{3.4}$$

an equation in Y'. Then, since x is to have the form (3.2) with $\xi \in \mathcal{D}(A)$, and since

$$L\xi = \hat{A}\xi = A\xi, \qquad \xi \in \mathcal{D}(A),$$

(3.4) becomes

$$\frac{dx}{dt} = \hat{A}x + (L\hat{x} - \hat{A}\hat{x})u.$$

We define $b \in Y'$, a continuous linear functional on $Y = \mathcal{D}(A')$, by

$$\langle b, \eta \rangle = \langle L\hat{x} - \hat{A}\hat{x}, \eta \rangle = \langle L\hat{x}, \eta \rangle - \langle \hat{x}, A'\eta \rangle \tag{3.5}$$

for $\eta \in \mathcal{D}(A') \equiv Y$. We then have

$$b = \sum_{K} b_k \phi_k$$

where the control input coefficients, b_k, are given by

$$b_k = \langle b, \psi_k \rangle = \langle L\hat{x}, \psi_k \rangle - \langle \hat{x}, A'\psi_k \rangle = \langle L\hat{x}, \psi_k \rangle - \lambda_k \langle \hat{x}, \psi_k \rangle. \tag{3.6}$$

In most examples we shall have $L\hat{x} = 0$. Then, if

$$\hat{x} = \sum_{K} \hat{x}_k \phi_k,$$

convergent in X, we obtain, in place of (3.6),

$$b_k = -\lambda_k \hat{x}_k, \qquad k \in K. \tag{3.7}$$

Also, in this case, the equation (3.5) becomes

$$\langle b, \eta \rangle = -\langle x, A'\eta \rangle. \tag{3.8}$$

The equation (3.5) (or (3.8)) will generally be used to identify the functional form of b while (3.6) (or (3.7)) will be used to identify its expansion coefficients in terms of the eigenvectors ϕ_k of the operator A. While not all admissible input elements can be treated this way, the class is large enough, we believe, to warrant the detailed description we have given here.

Example 1. *Heat equation.* Let $x(s, t)$ satisfy

$$\frac{\partial x}{\partial t} = \frac{\partial^2 x}{\partial s^2}, \qquad 0 < s < 1, \quad t > 0, \tag{3.9}$$

with boundary conditions

$$x(0,t)=0, \qquad \alpha x(1,t)+\beta\frac{\partial x}{\partial s}(1,t)=u(t), \tag{3.10}$$

where α, β are real numbers, not both equal to zero. In this case we take

$$X=L^2[0,1],$$

$$Ax=\frac{\partial^2 x}{\partial s^2}, \qquad x\in\mathcal{D}(A)=\{x\in H^2[0,1]\,|\,x(0)=0,\ \alpha x(1)+\beta x'(1)=0\},$$

$$Lx=\frac{\partial^2 x}{\partial s^2}, \qquad x\in\mathcal{D}(L)=\{x\in H^2[0,1]\,|\,x(0)=0\},$$

$$\hat{x}(s)=\begin{cases}\dfrac{s}{\alpha+\beta}, & \alpha+\beta\neq 0,\\[2ex] \dfrac{s(2-s)}{\alpha}, & \alpha+\beta=0.\end{cases} \tag{3.11}$$

With

$$\langle x, y\rangle=\int_0^1 x(s)y(s)\,ds,$$

we see that if $x, y\in\mathcal{D}(A)$,

$$\begin{aligned}\langle Ax, y\rangle-\langle x, Ay\rangle&=\int_0^1 (x''(s)y(s)-x(s)y''(s))\,ds\\ &=\int_0^1 \frac{d}{ds}(x'(s)y(s)-x(s)y'(s))\,ds\\ &=x'(1)y(a)-x(1)y'(a) \qquad (\text{since } x(0)=y(0)=0)\\ &=\begin{cases}\left(x'(1)+\dfrac{\alpha}{\beta}x(1)\right)y(1)-x(1)\left(\dfrac{\alpha}{\beta}y(1)+y'(1)\right), & \beta\neq 0,\\[2ex] x'(1)\left(y(1)+\dfrac{\beta}{\alpha}y'(1)\right)-\left(x(1)+\dfrac{\beta}{\alpha}x'(1)\right)y'(1), & \alpha\neq 0\end{cases}\\ &=0,\end{aligned}$$

and we conclude $A=A'$. In the first case of (3.11), $\alpha+\beta\neq 0$, $L\hat{x}=0$ and we have, for $\eta\in\mathcal{D}(A')=\mathcal{D}(A)$,

$$\begin{aligned}\langle b, \eta\rangle&=-\langle\hat{x}, A'\eta\rangle=-\frac{1}{\alpha+\beta}\int_0^1 s\eta''(s)\,ds\\ &=\frac{1}{\alpha+\beta}(-s\eta'(s)|_0^1+\eta(s)|_0^1)=\frac{-\eta'(1)+\eta(1)}{\alpha+\beta}\\ &=\begin{cases}\dfrac{1}{\beta}\eta(1), & \beta\neq 0,\\[2ex] -\dfrac{1}{\alpha}\eta'(1), & \alpha\neq 0.\end{cases}\end{aligned}$$

Thus we have

$$b = \begin{cases} \frac{1}{\beta}\delta_{(1)}, & \beta \neq 0, \\ \frac{1}{\alpha}\delta'_{(1)}, & \alpha \neq 0. \end{cases} \tag{3.12}$$

The two agree if neither α nor β is zero because the linear functional $(1/\beta)\delta_{(1)} + (1/\alpha)\delta'_{(1)}$ is zero in $(\mathcal{D}(A))' = Y'$ in this case.

The eigenvalues of A are $\lambda_k = -\omega_k^2$ where, for $k = 1, 2, 3, \cdots$,

$$\alpha \sin(\omega_k) + \beta\omega_k \cos(\omega_k) = 0. \tag{3.13}$$

Let

$$\frac{\alpha}{\sqrt{\alpha^2 + \beta^2\omega_k^2}} = \sin\theta_k, \qquad \frac{\beta\omega_k}{\sqrt{\alpha^2 + \beta^2\omega_k^2}} = \cos\theta_k,$$

and (3.13) becomes

$$\cos(\omega_k - \theta_k) = 0,$$

so that

$$\omega_k - \theta_k = \left(\frac{2k-1}{2}\right)\pi, \qquad k = 1, 2, 3, \cdots,$$

giving

$$\omega_k = \left(\frac{2k-1}{2}\right)\pi + \sin^{-1}\left(\frac{\alpha}{\sqrt{\alpha^2 + \beta^2\omega_k^2}}\right).$$

It is easy to see that $1/\omega_k = O(1/k)$ as $k \to \infty$ so

$$\omega_k = \left(\frac{2k-1}{2}\right)\pi + O\left(\frac{1}{k}\right), \qquad \beta \neq 0, \tag{3.14}$$

$$\omega_k = \left(\frac{2k-1}{2}\right)\pi + \frac{\pi}{2} = k\pi, \qquad \beta = 0. \tag{3.15}$$

Defining

$$\nu_k^2 = \int_0^1 \sin(\omega_k s)^2\, ds,$$

it is easily seen that in all cases the ν_k are nonzero and

$$\lim_{k\to\infty} \nu_k = \frac{1}{\sqrt{2}}.$$

Then the eigenfunctions

$$\phi_k(s) = \frac{1}{\nu_k}\sin(\omega_k s)$$

form an orthonormal basis for $L^2[0, 1]$. It follows that the coefficients of the input distribution elements (3.12) are given by

$$b_k = \begin{cases} \dfrac{1}{\beta\nu_k}\sin(\omega_k), & \beta \neq 0, \\ \dfrac{\omega_k \cos(\omega_k)}{\alpha\nu_k}, & \alpha \neq 0. \end{cases} \tag{3.16}$$

We consider here the case $\beta \neq 0$, saving the analysis for $\beta = 0$ until later in this section. If $\beta \neq 0$, formula (3.16) shows the b_k to be uniformly bounded. The complex numbers $-\lambda_k = \omega_k^2$ have the property (from (3.14))

$$-\lambda_k = \left(\frac{2k-1}{2}\right)^2 \pi^2 + O(1). \tag{3.17}$$

Thus the number of such $-\lambda_k$ in any set $|\mathrm{Im}\,(z) - \tau| \leqq h$, $\alpha \leqq \mathrm{Re}\,(z) \leqq \alpha + h$, is $O(h^{1/2})$, and it follows that the measure μ with $\mu(-\lambda_k) = |b_k|^2$, $\mu(\{\mathrm{Re}\,(z) \geqq \alpha\} - \cup_{k=1}^{\infty}\{-\lambda_k\}) = 0$, is a Carleson measure. Hence if $\beta \neq 0$, the boundary input (3.10) is admissible.

In this case the result is easily obtained without the Carleson measure theorem, for, if the coefficients c_k are square summable and $T > 0$,

$$\begin{aligned} \left\| \sum_{k=1}^{\infty} b_k c_k e^{\lambda_k t} \right\|_{L^2[0,T]} &\leqq \sup_k \{|b_k|\} \sqrt{\sum_{k=1}^{\infty} |c_k|^2} \sqrt{\sum_{k=1}^{\infty} \int_0^T e^{2\lambda_k t}\, dt} \\ &\leqq \sup_k \{|b_k|\} \sqrt{\sum_{k=1}^{\infty} |c_k|^2} \sqrt{\sum_{k=1}^{\infty} \int_0^{\infty} e^{2\lambda_k t}\, dt} \\ &= \sup_k \{|b_k|\} \sqrt{\sum_{k=1}^{\infty} |c_k|^2} \sqrt{\sum_{k=1}^{\infty} \frac{-1}{2\lambda_k}} < \infty \end{aligned} \tag{3.18}$$

since $\sup_k \{|b_k|\} < \infty$, and we conclude that the function sequence $\{b_k e^{-\lambda_k t}\}$ is l^2-convergent in $L^2[0, T]$. Our next example is chosen in such a way that a simple argument of this types does not apply, and the Carleson theorem is actually needed.

Example 2. *Another heat conduction system.* As a further example we ask the reader to consider the system shown in Fig. 3.1.

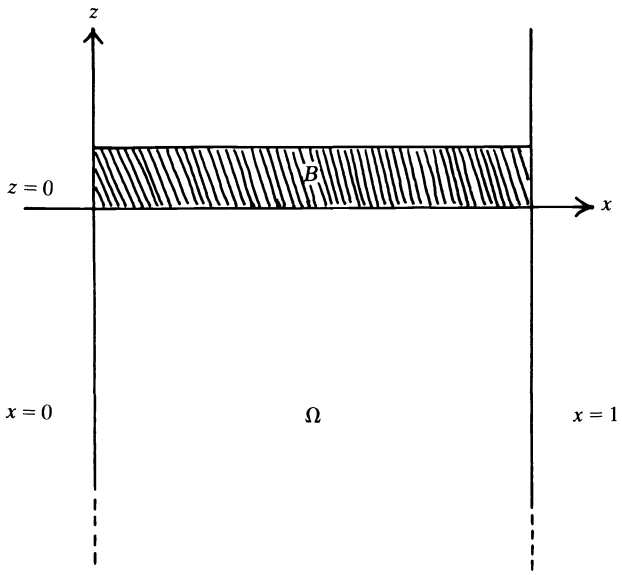


FIG. 3.1

The shaded horizontal bar, B, represents a layer of material, whose depth will be assumed negligible, and whose heat conductivity, k, is small in comparison to its specific heat R, while the region Ω consisting of the half strip

$$\Omega: 0 \leqq x \leqq 1, \qquad z \leqq 0,$$

is assumed filled with a material whose specific heat, r, is small by comparison with its conductivity, K. The heat flow equations are thus

$$R\frac{\partial T}{\partial t} = k\frac{\partial^2 T}{\partial x^2} - K\frac{\partial \tau}{\partial z}, \tag{3.19}$$

$$r\frac{\partial \tau}{\partial t} = K\left(\frac{\partial^2 \tau}{\partial x^2} + \frac{\partial^2 \tau}{\partial z^2}\right), \tag{3.20}$$

together with boundary conditions

$$\frac{\partial T}{\partial x}(0, t) = 0, \qquad \frac{\partial T}{\partial x}(1, t) = 0, \tag{3.21}$$

$$\frac{\partial \tau}{\partial x}(0, z, t) = 0, \qquad \frac{\partial \tau}{\partial x}(1, z, t) = g(z)u(t), \tag{3.22}$$

$$\lim_{z \to -\infty} \frac{\partial \tau}{\partial z}(x, z, t) = \lim_{z \to -\infty} \tau(x, z, t) = 0, \tag{3.23}$$

$$\tau(x, 0, t) = T(x, t), \qquad 0 \leqq x \leqq 1. \tag{3.24}$$

The inhomogeneous boundary condition along $x = 1$, $z \leqq 0$ represents the input heat flux. In (3.19), (3.21), $T(x, t)$ is the temperature in the bar, $\tau(x, z, t)$ the temperature in Ω.

If we assume k, r very small by comparison with R, K, we may, as an idealization, replace (3.19) and (3.20) by

$$R\frac{\partial T}{\partial t} = -K\frac{\partial \tau}{\partial z}, \tag{3.25}$$

$$\frac{\partial^2 \tau}{\partial x^2} + \frac{\partial^2 \tau}{\partial z^2} = 0, \tag{3.26}$$

retaining the boundary conditions (3.21)–(3.24). We take as our basic state space

$$\mathcal{T} = \{T = T(x) | T \in L^2[0, 1]\}.$$

We define an operator A on $\mathcal{T}$ with domain

$$\mathcal{D}(A) = H^1[0, 1]$$

as follows. Given $T \in \mathcal{D}(A)$, we let $\tau = \tau(x, z)$ satisfy (3.26) in Ω together with

$$\tau(x, 0) = T(x), \qquad 0 \leqq x \leqq 1, \tag{3.27}$$

and

$$\frac{\partial \tau}{\partial x}(0, z) = 0, \qquad \frac{\partial \tau}{\partial x}(1, z) = 0, \tag{3.28}$$

$$\lim_{z \to -\infty} \frac{\partial \tau}{\partial z}(\cdot, z) = 0 \quad \text{in } L^2[0, 1], \tag{3.29}$$

$$\lim_{z \to -\infty} \tau(\cdot, z) = 0 \quad \text{in } H^1[0, 1]. \tag{3.30}$$

From [14], for $T \in H^1[0, 1]$ we have $\tau \in H^{3/2}(\Omega)$. The trace theorem [1], [14] then gives

$$\frac{\partial \tau}{\partial z}(\cdot, 0) \in L^2[0, 1],$$

and we define

$$AT = -\frac{K}{R}\frac{\partial \tau}{\partial z}(\cdot, 0). \tag{3.31}$$

So doing, (3.25) becomes

$$\dot{T} = AT \tag{3.32}$$

and (3.31) is subsumed in the definition of A.

LEMMA 3.1. *The operator $-A$ is the positive square root of the Stürm–Liouville operator*

$$ST = -\frac{K^2}{R^2}\frac{d^2T}{dx^2}$$

with

$$\mathcal{D}(T) = \left\{ T \in H^2[0, 1] \middle| \frac{dT}{dx}(0) = \frac{dT}{dx}(1) = 0 \right\}.$$

Proof. We compute $(-A)^2T$ for $T \in \mathcal{D}(T)$. For such T the solution of (3.27)–(3.30) $\in H^{5/2}(\Omega)$. If we let

$$\hat{\tau}(x, z) = \frac{K}{R}\frac{\partial \tau}{\partial z}(x, z),$$

then

$$\hat{\tau}(\cdot, 0) = -AT$$

and

$$(-A)^2T = -A\hat{\tau}(\cdot, 0) = \frac{K^2}{R^2}\frac{\partial^2 \tau}{\partial z^2}(\cdot, 0) = -\frac{K^2}{R^2}\frac{\partial^2 \tau}{\partial x^2}(\cdot, 0) = -\frac{K^2}{R^2}\frac{d^2T}{dx^2},$$

since $\tau \in H^{5/2}(\Omega)$ together with (3.26) implies that

$$\frac{\partial^2 \tau}{\partial x^2}(\cdot, 0) + \frac{\partial^2 \tau}{\partial z^2}(\cdot, 0) = 0 \quad \text{in } L^2[0, 1]$$

and $T = \tau(\cdot, 0)$.

The positivity of $-A$ follows from the divergence theorem. If $T \in \mathscr{D}(A)$ and if $\tau = \tau(x, z)$ is constructed as above, we have

$$\int_\Omega \int \left[\left(\frac{\partial \tau}{\partial x}(x, z) \right)^2 + \left(\frac{\partial \tau}{\partial z}(x, z) \right)^2 \right] dx\, dz$$

$$= \int_\Omega \int \|\nabla \tau(x, z)\|^2 \, dx\, dz \qquad (\nabla = \text{gradient})$$

$$= \int_\Omega \int [\operatorname{div}(\tau(x, z) \nabla \tau(x, z)) - \tau(x, z) \Delta^2 \tau(x, y)] \, dx\, dz \qquad (\Delta^2 = \text{Laplacian})$$

$$= \int_\Omega \int \operatorname{div}(\tau(x, z) \nabla \tau(x, z)) \, dx\, dz \qquad (\text{from (3.26)})$$

$$= \int_0^1 \tau(x, 0) \frac{\partial \tau}{\partial z}(x, 0) \, dx = (T, -AT)_{L^2[0,1]} \quad (\text{using (3.27)–(3.30)}).$$

This completes the proof.

Accordingly, A is selfadjoint with eigenfunctions

$$\phi_0(x) \equiv 1, \qquad \phi_k(x) = \sqrt{2} \cos(k\pi x), \quad k = 1, 2, 3, \cdots, \tag{3.33}$$

and eigenvalues

$$\lambda_0 = 0, \qquad \lambda_k = -\frac{K}{R} k\pi, \quad k = 1, 2, 3, \cdots. \tag{3.34}$$

Let $w(x, z)$ be the solution of the following inhomogeneous boundary value problem:

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial z^2} = 0 \quad \text{in } \Omega,$$

$$\frac{\partial w}{\partial x}(0, z) = 0, \qquad \frac{\partial w}{\partial x}(1, z) = g(z),$$

$$\lim_{z \to -\infty} \frac{\partial w}{\partial z}(x, z) = \lim_{z \to -\infty} w(x, z) = 0,$$

$$w(x, 0) \equiv 0, \qquad 0 \leqq x \leqq 1.$$

We will assume that $g(z)$ is such that the resulting $w(x, z) \in H^2(\Omega)$.

In this case the inhomogeneous equation can be interpreted as

$$\dot{T} = AT + bu$$

where $b = b(x)$ is given by

$$b(x) = -\frac{K}{R} \frac{\partial w}{\partial z}(x, 0).$$

To compute the coefficients of the expansion

$$b(x) = \sum_{k=0}^{\infty} b_k \phi_k(x),$$

we note that since A is selfadjoint, $\psi_k(x) = \phi_k(x)$, and

$$b_k = \int_0^1 \phi_k(x) b(x)\, dx. \tag{3.35}$$

Let $\Phi_k(x, z)$ be the solution of

$$\frac{\partial^2 \Phi_k}{\partial x^2} + \frac{\partial^2 \Phi_k}{\partial z^2} = 0 \quad \text{in } \Omega$$

with

$$\Phi_k(x, 0) = \Phi_k(x)$$

and homogeneous boundary conditions of the type (3.27)–(3.30) otherwise. Then, with $\Delta^2 = \partial^2/\partial x^2 + \partial^2/\partial z^2$,

$$\begin{aligned} 0 &= \int_\Omega [\Phi_k(x, z)\Delta^2 w(x, z) - w(x, z)\Delta^2 \Phi_k(x, z)]\, dx\, dz \\ &= \int_\Omega \operatorname{div} [\Phi_k(x, z) \operatorname{grad} w(x, z) - w(x, z) \operatorname{grad} \Phi_k(x, z)]\, dx\, dz \\ &= \int_{\partial\Omega} [\Phi_k(x, z) \operatorname{grad} w(x, z) - w(x, z) \operatorname{grad} \Phi_k(x, z)] \cdot \nu(x, z)\, ds \\ &= -\frac{R}{K} \int_0^1 \phi_k(x) b(x)\, dx + \int_{-\infty}^0 \Phi_k(1, z) g(z)\, dz, \end{aligned}$$

giving (cf. (3.35))

$$b_k = \frac{K}{R} \int_{-\infty}^0 \Phi_k(1, z) g(z)\, dz.$$

Now it is easily checked that for $k = 1, 2, 3, \cdots$

$$\Phi_k(x, z) = (\sqrt{2} \cos k\pi x)(\exp (k\pi z)),$$

so that

$$\Phi_k(1, z) = (-1)^k \sqrt{2} \exp (k\pi z)$$

and thus

$$b_k = \frac{(-1)^k \sqrt{2} K}{R} \int_{-\infty}^0 \exp (k\pi z).$$

The Carleson measure theorem can be used in a slightly different way from that set forth in Corollary 2.5 to show that if $g \in L^2(-\infty, 0]$ then the b_k are square summable and b is, consequently, an element of $L^2[0, 1]$. Writing $\zeta = -z$, $g(-\zeta) = \tilde{g}(\zeta)$, we see that

$$b_k = \frac{(-1)^k \sqrt{2} K}{R} \int_0^\infty \exp (-k\pi\zeta) \tilde{g}(\zeta)\, d\zeta.$$

Since the measure μ assigning the value 1 to each of the points $k\pi$, $k = 0, 1, 2, \cdots$, is clearly a Carleson measure, and since $(-1)^k \sqrt{2}\, K/R$ changes only in sign, $\{b_k\} \in l^2$.

If $g(z)$ is just bounded and measurable on $-\infty < z \leqq 0$, we can almost trivially obtain

$$b_k = O\left(\frac{1}{k}\right)$$

and the b_k will be square summable.

It is obviously possible to replace $g(z)$ by distributions of various types. Taking $g(z) = \delta_{(0)}$ corresponds to a point heat source at the corner $x = 1$, $z = 0$ and leads to

$$b_k = \frac{(-1)^k \sqrt{2}\, K}{R}. \tag{3.36}$$

In our present example $X = L^2[0, 1]$, $Y = \mathscr{D}(A) = H^1[0, 1]$ and $Y' = H^{-1}[0, 1]$. The coefficients (3.36) may be recognized as those corresponding to $\delta_{(1)}$ (referring now to distributions along the x-axis).

Any measure μ assigning to the points $-\lambda_k = Kk\pi/R$ values $|b_k|^2$ which are bounded evidently yields a Carleson measure, and we conclude that all of the above cases correspond to admissible input elements. In this case the argument represented by the inequalities (3.18) will not work because the series $\sum_{k=1}^{\infty}(-1/2\lambda_k)$ is not summable in this example.

Example 3. *Hyperbolic and neutral systems.* A wide variety of systems involving linear hyperbolic partial differential equations in two independent variables x, t, or neutral functional equations, lead to systems of the form described at the beginning of this section, the eigenvectors, ϕ_k, of A forming a uniform l^2-basis for the state space X and the eigenvalues λ_k confined to a vertical strip $\alpha < \mathrm{Re}\,(\lambda) < \beta$ in the complex plane. It also usually turns out in these cases that the number of λ_k in any rectangle

$$\alpha < \mathrm{Re}\,(\lambda) < \beta, \qquad \gamma < \mathrm{Im}\,(\lambda) < \delta$$

is less than or equal to $M(\delta - \gamma)$, where M is a fixed positive number. It is evident that the measure (2.27), (2.28) is a Carleson measure in these cases whenever the control input coefficients b_k constitute a bounded set.

Example 4. *Linear surface waves.* If the operator A is defined as in (3.31) but, instead of the first order system (3.32) we consider the second order counterpart

$$\ddot{\zeta} + A\zeta = 0, \tag{3.37}$$

we obtain the linearized equations for small amplitude waves on the surface of an incompressible fluid. The theory is more fully developed in [16], [17], [19]. With $\eta = \dot{\zeta}$, (3.37) is equivalent to the first order system

$$\begin{pmatrix} \dot{\zeta} \\ \eta \end{pmatrix} = \begin{pmatrix} 0 & I \\ -A & 0 \end{pmatrix} \begin{pmatrix} \zeta \\ \eta \end{pmatrix} \equiv \mathfrak{A} \begin{pmatrix} \zeta \\ \eta \end{pmatrix}. \tag{3.38}$$

To obtain a topology corresponding to the energy of the system, one defines

$$\left\| \begin{pmatrix} \zeta \\ \eta \end{pmatrix} \right\|_e^2 = \|\zeta\|^2_{L^2_0[0,1]} + (\eta, A^{-1}\eta)_{L^2_0[0,1]}, \tag{3.39}$$

where

$$L^2_0[0, 1] = \left\{ \zeta \in L^2[0, 1] \,\middle|\, \int_0^1 \zeta(x)\, dx = 0 \right\}. \tag{3.40}$$

The restriction to $L_0^2[0, 1]$ corresponds to conservation of fluid volume. On the domain

$$\mathscr{D}_0(A) = \left\{ \zeta \in H^1[0, 1] \middle| \int_0^1 \zeta(x)\, dx = 0 \right\}$$

the operator A is invertible. Its eigenvalues are (cf. (3.34))

$$\lambda_k = -\frac{K}{R} k\pi, \qquad k = 1, 2, 3, \cdots, \tag{3.41}$$

with the same eigenfunctions $\phi_k(x)$, $k = 1, 2, 3, \cdots$, as shown in (3.33). Correspondingly, the operator $\mathfrak{A}$ has eigenvalues

$$i\omega_k, -i\omega_k, \omega_k = \left(\frac{K\pi}{R}\right)^{1/2} k^{1/2} \equiv \gamma k^{1/2}, \qquad k = 1, 2, 3, \cdots, \tag{3.42}$$

and the eigenvectors, orthonormalized with respect to $\|\cdot\|_e$ and the corresponding inner product, are

$$\psi_k = \begin{pmatrix} \phi_k \\ i\omega_k \phi_k \end{pmatrix}, \quad \psi_{-k} = \begin{pmatrix} \phi_k \\ -i\omega_k \phi_k \end{pmatrix}, \quad k = 1, 2, 3, \cdots. \tag{3.43}$$

To discuss admissible input elements in this case we let β_k, β_{-k} be nonnegative numbers, $k = 1, 2, 3, \cdots$, and define

$$\mu\{i\omega_k\} = \beta_k, \quad \mu\{-i\omega_k\} = \beta_{-k}, \quad k = 1, 2, 3, \cdots,$$

$$\mu\left(\{\mathrm{Re}\,(z) \geqq \alpha\} - \bigcup_{k=1}^{\infty} (\{i\omega_k\} \cup \{-i\omega_k\})\right) = 0.$$

Let $\beta(\omega)$, $-\infty < \omega < \infty$, be defined as the piecewise linear function such that in the interval $[i\omega_k, i\omega_{k+1}]$

$$\beta(\omega) = \frac{\beta_k(\omega_{k+1} - \omega) + \beta_{k+1}(\omega - \omega_k)}{\omega_{k+1} - \omega_k}. \tag{3.44}$$

Since

$$\int_{\omega_k}^{\omega_{k+1}} \omega^{1/2} \beta(\omega)\, d\omega = \frac{\beta_k \omega_k^{1/2} + \beta_{k+1} \omega_{k+1}^{1/2}}{2} [\omega_{k+1} - \omega_k]$$

$$\to \tfrac{1}{4}(\beta_k + \beta_{k+1}), \qquad k \to \infty,$$

we conclude that μ is a Carleson measure just in case there is a constant C such that

$$\int_\sigma^\tau \omega^{1/2} \beta(\omega)\, d\omega \leqq C|\tau - \sigma| \tag{3.45}$$

whenever $0 < \sigma < \tau$, together with a comparable condition involving the β_{-k} and negative values of ω. But (3.45) is true just in case

$$\omega_k^{1/2} \beta_k \leqq C, \qquad k = 1, 2, 3, \cdots,$$

and the comparable condition for negative k is

$$\omega_k^{1/2} \beta_{-k} \leqq C, \qquad k = 1, 2, 3, \cdots.$$

Thus for the inhomogeneous system

$$\begin{pmatrix}\dot{\zeta}\\ \dot{\eta}\end{pmatrix} = \mathfrak{A}\begin{pmatrix}\zeta\\ \eta\end{pmatrix} + \begin{pmatrix}b_1\\ b_2\end{pmatrix} u$$

the input element $\binom{b_1}{b_2}$ with

$$\left\langle \begin{pmatrix}\phi_k\\ i\omega_k\phi_k\end{pmatrix}, \begin{pmatrix}b_1\\ b_2\end{pmatrix} \right\rangle = \beta_k, \qquad \left\langle \begin{pmatrix}\phi_k\\ -i\omega_k\phi_k\end{pmatrix}, \begin{pmatrix}b_1\\ b_2\end{pmatrix} \right\rangle = \beta_{-k}$$

is admissible, from this criterion, if

$$k^{1/2}(|\beta_k|^2 + |\beta_{-k}|^2) < C \tag{3.46}$$

for some fixed positive number C. It will be noted that this is (slightly) less restrictive than the requirement

$$\left\| \begin{pmatrix}b_1\\ b_2\end{pmatrix} \right\|_e < \infty.$$

Example 5. *Negative results.* For any system similar to the one in Example 4 but with $|\omega_{k+1} - \omega_k| = O(1/k^{1/2} + \varepsilon)$ the Carleson measure condition will be stronger than requiring $b \in X$. Hence failure of the Carleson measure condition cannot be used to show that an element b is not admissible, for any $b \in X$ is admissible.

To illustrate what can be done in a negative direction, we return to Example 1 with $\beta = 0$. This situation has been studied, using a different approach, in [13]. We present here an argument more in the spirit of the present work. As shown in (3.15),

$$\lambda_k = -\omega_k^2 = -k^2\pi^2, \tag{3.47}$$

and (cf. (3.16) and without loss of generality, taking $\alpha = 1$)

$$b_k = \sqrt{2}\, k\pi \cos(k\pi) = (-1)^k \sqrt{2}\, k\pi. \tag{3.48}$$

Since $\beta_k = |b_k|^2 = 2k^2\pi^2$ while $(k+1)^2\pi^2 - k^2\pi^2 = 2k\pi^2 + \pi^2$, it is not hard to see that the measure $\mu = \mu_{b,\{\lambda_k\}}$, $\mu(-\lambda_k) = |b_k|^2$ is not a Carleson measure in this case. As we have remarked, this by itself is not enough to show that the input element with coefficients (3.48) is not admissible. To show this, we ask the reader to consider the function

$$\psi_r(z) = (z+1)^{-r},$$

analytic in the complex plane minus the cut consisting of $\{z \mid z \text{ real}, z \leqq -1\}$. If $r > \frac{1}{2}$, ψ_r is square integrable on any vertical line $\{z \mid \mathrm{Re}\,(z) = \xi, \xi \geqq 0\}$ with uniformly bounded L^2 norm and $\psi_r(z)$ is bounded for $\mathrm{Re}\,(z) \geqq 0$. It follows that $\psi_r(z)$ is the Laplace transform of a function $f_r = f_r(t)$ with $f_r \in L^2[0, \infty)$. Then

$$\begin{aligned} b_k \int_0^\infty e^{-k^2\pi^2 t} f_r(t)\, dt &= (-1)^k \sqrt{2}\, k\pi \psi_r(k^2\pi^2) \\ &= \frac{(-1)^k \sqrt{2}\, k\pi}{(k^2\pi^2 + 1)^r} - O(|k|^{1-2r}), \qquad k \to \infty. \end{aligned} \tag{3.49}$$

This expression is not square summable if r satisfies the inequalities

$$1 - 2r \geqq -\tfrac{1}{2},$$

so we require

$$\tfrac{1}{2} < r \leqq \tfrac{3}{4}.$$

Let E be the closed subspace spanned by the functions $e^{-k^2\pi^2 t}$ in $L^2[0,\infty)$ and let E_T, $T>0$, be the subspace of $L^2[0,T]$ consisting of restrictions to $[0,T]$ of functions in E. If $\hat{f}_r$ is the orthogonal projection of f_r onto E we clearly have

$$\int_0^\infty e^{-k^2\pi^2 t}\hat{f}_r(t)\,dt = \int_0^\infty e^{-k^2\pi^2 t}f_r(t)\,dt.$$

It is shown in [7], [18] that the natural restriction map $R: E \to E_T$ is onto, (obviously) bounded and (not so obviously) boundedly invertible with respect to the induced $L^2[0,\infty)$, $L^2[0,T]$ topologies of E, E_T, respectively. Thus, with $p_k(t) = e^{-k^2\pi^2 t}$,

$$\int_0^\infty e^{-k^2\pi^2 t}\hat{f}_r(t)\,dt = (\hat{f}_r, p_k)_{L^2[0,\infty)}$$

$$= (\hat{f}_r, R^{-1}Rp_k)_{L^2[0,\infty)} = ((R^{-1})^*\hat{f}_r, Rp_k)_{L^2[0,T]} = \int_0^T e^{-k^2\pi^2 t}\varphi_r(t)\,dt$$

where

$$\varphi_r = (R^{-1})^*\hat{f}_r \in E_T \subset L^2[0,T].$$

It follows that φ_r is an element of $L^2[0,T]$ such that the numbers

$$(-1)^k\sqrt{2}\,k\pi\int_0^T e^{-k^2\pi^2 t}\varphi_r(t)\,dt, \qquad k = 1, 2, 3, \cdots,$$

are not square summable. From earlier developments, the input element b with coefficients (3.12) corresponding to the boundary condition (3.10), with $\beta = 0$, $\alpha = 1$:

$$x(1,t) = u(t),$$

is not an admissible input element.

4. A proof of Theorem 2.4. It is clear that the Carleson measure theorem in H^2_α, Theorem 2.4, is central to our work in this paper. This result, in one form or another, has been known for somewhat more than a decade. A proof for $H^2(D)$, where D is the unit disc in the complex plane, appears in Duren [5]. A proof for functions in H^1_α is given by Koosis in his recent book [12]. The reader is also referred to the recent book [8] by J. Garnett. Because the result is not particularly well known outside the circle of mathematicians working in H^p theory and because the results are rather scattered and not readily available in precisely the form we require, we offer here a proof of Theorem 2.4 which is a direct adaptation to the half plane of the result for the unit disc appearing in Duren's book [5]. The proof given here originally formed part of the first author's doctoral dissertation [9]. As in Duren's work, the proof makes use of a relatively simple case of the Marcinkiewicz interpolation theorem [23, Chap. XI] and, again following Duren, we do not quote the general Marcinkiewicz theorem but, rather, give a direct proof for the simple special case required here.

We begin with a covering lemma of Vitali type.

LEMMA 4.1. *Let $\{I_\lambda | \lambda \in \Lambda\} \equiv \mathcal{I}$ be a family of intervals in R^1. Suppose there is a positive number K such that for any finite collection $\{I_{\lambda_1}, I_{\lambda_2}, \cdots, I_{\lambda_n}\}$ of disjoint intervals in $\mathcal{I}$*

$$\sum_{k=1}^n |I_{\lambda_k}| < K. \tag{4.1}$$

Then we can choose a sequence $\{I_{\lambda_k} | k = 1, 2, 3, \cdots\}$ *of disjoint intervals from* $\mathcal{I}$ *with the property: for every* $\lambda \in \Lambda$ *there exists* $k \in \{1, 2, 3, \cdots\}$ *such that*

$$I_\lambda \subset J_k$$

where J_k *is the interval having the same center as* I_{λ_k} *but five times the length of* I_{λ_k}.

Proof. From (4.1) it follows, in particular, that the length, $|I_\lambda|$, of I_λ is uniformly bounded (take $n = 1$, $\lambda_1 = \lambda$). Define the sequence $\{I_{\lambda_k}\}$ inductively as follows. Let I_{λ_1} be such that

$$|I_{\lambda_1}| \geqq \tfrac{1}{2} \sup_{\lambda \in \Lambda} |I_\lambda|.$$

For $k = 2, 3, 4, \cdots$ let I_{λ_k} be disjoint from I_{λ_i}, $i = 1, 2, \cdots, n-1$, and such that

$$|I_{\lambda_k}| \geqq \tfrac{1}{2} \sup \left\{ |I_\lambda| \,\middle|\, \lambda \in \Lambda, I_\lambda \cap I_{\lambda_i} = \phi, i = 1, 2, \cdots, k-1 \right\}. \tag{4.2}$$

Since the I_{λ_k} are disjoint it follows from (4.1) that

$$\lim_{k \to \infty} |I_{\lambda_k}| = 0. \tag{4.3}$$

Let $I_\lambda \in \mathcal{I}$. Then there exists k such that

$$I_\lambda \cap I_{\lambda_k} \neq \phi. \tag{4.4}$$

Otherwise (4.2) and (4.3) could not both be true. Let k_0 be the smallest integer such that (4.4) is true. Then

$$|I_\lambda| \leqq 2|I_{\lambda_{k_0}}|$$

and, together with the fact that $I_\lambda \cap I_{\lambda_{k_0}} \neq \phi$, this implies that $I_\lambda \subset J_k$, completing the proof.

We subdivide the rest of the proof of Theorem 2.4 into several propositions for clarity. The proof is given for the half plane $\operatorname{Re}(z) > 0$, without loss of generality, and we designate H_0^2 simply by H^2.

PROPOSITION 4.2. *Let* $\phi \in H^2$ *and let* $\phi_0(i\,\cdot)$ *be the corresponding boundary function in* $L^2(-\infty, \infty)$. *For* $z = \sigma + i\tau$, $\sigma > 0$, *let* I_z *be the interval*

$$I_z = [\tau - \sigma, \tau + \sigma] \tag{4.5}$$

and let

$$\tilde{\phi}(z) = \sup_{I \in \mathcal{I}_z} \frac{1}{|I|} \int_I |\phi_0(it)|\, dt, \tag{4.6}$$

where $\mathcal{I}_z$ *is the set of all finite intervals containing* I_z. *Then*

$$|\phi(z)| \leqq \frac{10}{\pi} \tilde{\phi}(z). \tag{4.7}$$

Proof. From the Poisson integral formula in the half plane we have

$$\phi(z) = -\frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\sigma \phi_0(it)\, dt}{\sigma^2 + (\tau - t)^2}$$

so that

$$|\phi(z)| \leqq \frac{1}{\pi}\left[\sum_{N=0}^{\infty}\int_{2^N\sigma\leqq|t-\tau|\leqq 2^{N+1}\sigma}\frac{\sigma|\phi_0(it)|\,dt}{\sigma^2+(\tau-t)^2}+\int_{|t-\tau|<\sigma}\frac{\sigma|\phi_0(it)|\,dt}{\sigma^2+(\tau-t)^2}\right]$$

$$\leqq \frac{1}{\pi}\left[\sum_{N=0}^{\infty}\int_{|t-\tau|\leqq 2^{N+1}\sigma}\frac{|\phi_0(it)|\,dt}{4^N\sigma}+\int_{|t-\tau|\leqq\sigma}\frac{|\phi_0(it)|\,dt}{\sigma}\right]$$

$$\leqq \frac{1}{\pi}\left[\sum_{N=0}^{\infty}\frac{1}{2^{N-2}}\tilde{\phi}(z)+2\tilde{\phi}(z)\right]$$

$$= \frac{10}{\pi}\tilde{\phi}(z).$$

PROPOSITION 4.3. *Let* $\psi \in L^1(-\infty,\infty)$ *and, for* $z=\sigma+i\tau$, $\sigma>0$, *let* I_z *be given by* (4.5) *while* (*cf.* (4.6))

$$\tilde{\psi}(z) = \sup_{I\in\mathcal{I}_z}\frac{1}{|I|}\int_I|\psi(t)|\,dt. \tag{4.8}$$

Let μ *be a Carleson measure and, for* $s\geqq 0$, *let* E_s *be the Borel measurable subset of* $\{z\,|\,\mathrm{Re}\,(z)>0\}$ *given by*

$$E_s=\{z\,|\,\mathrm{Re}\,(z)>0,\ \tilde{\psi}(z)>s\}.$$

Then, with A as in Definition 2.3,

$$\mu(E_s)\leqq\frac{5A}{2s}\|\psi\|_{L^1(-\infty,\infty)}. \tag{4.9}$$

Proof. Let $\mathcal{I}$ be the family of all finite intervals in R^1 such that

$$\frac{1}{|I|}\int_I|\psi(t)|\,dt>s. \tag{4.10}$$

If $I_1, I_2, \cdots, I_n\in\mathcal{I}$ are disjoint, then (4.10) gives, for every n,

$$\sum_{k=1}^{n}|I_k|\leqq\frac{1}{s}\sum_{k=1}^{n}\int_{I_k}|\psi(t)|\,dt\leqq\frac{1}{s}\|\psi\|_{L^1(-\infty,\infty)}. \tag{4.11}$$

Thus $\mathcal{I}$ satisfies the hypotheses of Lemma 4.1 and we can find a disjoint sequence $\{I_n\,|\,n=1,2,3,\cdots\}\subset\mathcal{I}$ such that, J_n having the same center as I_n but five times the length, each $I\in\mathcal{I}$ is contained in some J_n.

If $z\in E_s$, then $I_z\subset I$ for some $I\in\mathcal{I}$ and we have, for some n,

$$I_z=[\tau-\sigma,\tau+\sigma]\subset J_n.$$

Then clearly,

$$z\in S_n=\left\{\sigma+i\tau\,|\,0<\sigma\leqq\frac{|J_n|}{2},\ \tau\in J_n\right\}.$$

This being true for all $z\in E_s$,

$$E_s\subset\bigcup_{n=1}^{\infty}S_n.$$

The positivity of $-A$ follows from the divergence theorem. If $T \in \mathscr{D}(A)$ and if $\tau = \tau(x, z)$ is constructed as above, we have

$$\int_\Omega \int \left[\left(\frac{\partial \tau}{\partial x}(x, z) \right)^2 + \left(\frac{\partial \tau}{\partial z}(x, z) \right)^2 \right] dx\, dz$$

$$= \int_\Omega \int \|\nabla \tau(x, z)\|^2 \, dx\, dz \qquad (\nabla = \text{gradient})$$

$$= \int_\Omega \int [\operatorname{div}(\tau(x, z) \nabla \tau(x, z)) - \tau(x, z) \Delta^2 \tau(x, y)] \, dx\, dz \qquad (\Delta^2 = \text{Laplacian})$$

$$= \int_\Omega \int \operatorname{div}(\tau(x, z) \nabla \tau(x, z)) \, dx\, dz \qquad (\text{from (3.26)})$$

$$= \int_0^1 \tau(x, 0) \frac{\partial \tau}{\partial z}(x, 0) \, dx = (T, -AT)_{L^2[0,1]} \quad (\text{using (3.27)–(3.30)}).$$

This completes the proof.

Accordingly, A is selfadjoint with eigenfunctions

$$\phi_0(x) \equiv 1, \qquad \phi_k(x) = \sqrt{2} \cos(k\pi x), \quad k = 1, 2, 3, \cdots, \tag{3.33}$$

and eigenvalues

$$\lambda_0 = 0, \qquad \lambda_k = -\frac{K}{R} k\pi, \quad k = 1, 2, 3, \cdots. \tag{3.34}$$

Let $w(x, z)$ be the solution of the following inhomogeneous boundary value problem:

$$\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial z^2} = 0 \quad \text{in } \Omega,$$

$$\frac{\partial w}{\partial x}(0, z) = 0, \qquad \frac{\partial w}{\partial x}(1, z) = g(z),$$

$$\lim_{z \to -\infty} \frac{\partial w}{\partial z}(x, z) = \lim_{z \to -\infty} w(x, z) = 0,$$

$$w(x, 0) \equiv 0, \qquad 0 \leqq x \leqq 1.$$

We will assume that $g(z)$ is such that the resulting $w(x, z) \in H^2(\Omega)$.

In this case the inhomogeneous equation can be interpreted as

$$\dot{T} = AT + bu$$

where $b = b(x)$ is given by

$$b(x) = -\frac{K}{R} \frac{\partial w}{\partial z}(x, 0).$$

To compute the coefficients of the expansion

$$b(x) = \sum_{k=0}^{\infty} b_k \phi_k(x),$$

Hence, from (4.9) of Proposition 4.3,

$$\mu(E_r) \leqq \mu(F_{r/2}) \leqq \frac{5A}{r} \|\psi_r\|_{L^1(-\infty,\infty)}$$

so that

$$\begin{aligned}\int_0^\infty r\alpha(r)\,dr &= \int_0^\infty r\mu(E_r)\,dr \\ &\leqq 5A \int_0^\infty \|\psi_r\|_{L^1(-\infty,\infty)}\,dr \\ &\leqq 5A\|\phi_0(i\cdot)\|^2_{L^2(-\infty,\infty)} \qquad \text{(using (4.13))}.\end{aligned}$$

Then (4.14) gives the inequality (4.12).

The proof of Theorem 2.4 is completed by combining (4.7) of Proposition 4.2 with (4.12) above to give

$$\int_{\mathrm{Re}\,(z)>0} |\phi(z)|^2\,d\mu(z) \leqq \frac{100}{\pi^2} \int_{\mathrm{Re}\,(z)>0} (\tilde{\phi}(z))^2\,d\mu(z) \leqq \frac{1{,}000A}{\pi^2} \int_{-\infty}^{\infty} |\phi_0(it)|^2\,dt$$

as claimed in (2.26), except for the trivial detail of replacing $\phi_0(i\cdot)$ by $\phi_\alpha = \phi(\alpha + i\cdot)$.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] LENNART CARLESON, *On infinite differential equations with constant coefficients*, I, Math. Scand., 1 (1953), pp. 31–38.

[3] R. CURTAIN AND A. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes on Control and Information Sciences, 8, Springer-Verlag, New York, 1978.

[4] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* I: *General Theory*, Vol. I, Interscience, New York, 1957.

[5] P. L. DUREN, *Theory of H^p Spaces*, Academic Press, New York, 1970.

[6] H. O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 349–385.

[7] H. O. FATTORINI AND D. L. RUSSELL, *Uniform bounds on biorthogonal functions for real exponentials with an application to the control theory of parabolic equations*, Quart. Appl. Math., 32 (1974), pp. 45–69.

[8] J. GARNETT, *Bounded Analytic Functions*, Academic Press, New York, 1981.

[9] L. F. HO, *Controllability and spectral assignability of a class of hyperbolic control systems with retarded control canonical forms*, Thesis, Univ. Wisconsin, Madison, WI, August 1981.

[10] K. HOFFMANN, *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[11] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[12] P. KOOSIS, *Introduction to H Spaces*, Cambridge Univ. Press, New York, 1980.

[13] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, S. K. Mitter, trans., Springer-Verlag, New York, 1971.

[14] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems*, Vols. I, II, Springer-Verlag, New York, 1971.

[15] P. H. RABINOWITZ, *Periodic solutions of nonlinear hyperbolic partial differential equations*, Comm. Pur. Appl. Math., 20 (1967), pp. 145–204.

[16] R. M. REID, *Control of linear surface waves on water*, Thesis, University of Wisconsin, Madison, WI, August 1979.

[17] R. M. REID AND D. L. RUSSELL, *Water waves and problems of infinite time control*, in Proc. IRIA Symp. on Analysis and Control of Systems, Rocquencourt, Dec. 1978.

[18] L. SCHWARTZ, *Etude des sommes d'exponentielles*, 2nd ed., Hermann, Paris, 1959.

[19] J. J. STOKER, *Water Waves*, Interscience, New York, 1957.

[20] D. C. WASHBURN, *A semigroup theoretic approach to modelling of boundary input problems*, in Proc. IFIP Conf. on Modelling and Identification of Distributed Parameter Systems, Rome, 1976.

[21] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, San Francisco, 1980.
[22] J. ZABCZYK, *A semigroup approach to boundary value control*, in Proc. 2nd IFAC Symposium on Distributed Parameter Systems, Warwick, June 1977.
[23] A. ZYGMUND, *Trigonometric Series*, Cambridge Univ. Press, Cambridge, 1968.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

SINGLE-VALUED REPRESENTATION OF SET-VALUED MAPPINGS II; APPLICATION TO DIFFERENTIAL INCLUSIONS*

A. D. IOFFE†

Abstract. We consider a set-valued mapping $Q(t, x)$, the first argument ranging through a measurable space and the second through a space with a different structure (topological, metric, differentiable, etc.). We are interested in representing Q by a single-valued mapping with properties compatible with the structures, i.e., in the existence of a (topological) space Z and a mapping $f(t, x, z)$ measurable in t, continuous in x, z and depending on x as dictated by the structure of the space of x's (say, continuous, Lipschitz or differentiable in x, etc.) and such that $f(t, x, Z) = Q(t, x)$ for all t, x. Approximate representations (i.e., those for which the latter equality holds approximately) of convex-valued mappings are also considered and certain applications, mainly to differential inclusions, are discussed.

Key words. representation of set-valued mappings, differential equations equivalent to differential inclusions

AMS subject classification. 34A60, 49E10, 49A50.

1. Introduction. To a large extent, this paper was motivated by the following question. Given a differential inclusion

$$\dot{x} \in Q(t, x), \tag{1}$$

under what conditions is there a differential equation with control

$$\dot{x} = f(t, x, u), \qquad u \in U \tag{2}$$

which is equivalent to (1)?

Of course, only such an f may be of an interest which has certain analytical properties, say, measurable in t, continuous in (x, u) and satisfying some additional requirements as a function of x so that the differential equation be reasonably good.

A natural way to define equivalence of (1) and (2) is to require that any solution to (1) should be a solution to (2) and vice versa. If for any (t, x) the set of possible velocities at t of those solutions $x(\cdot)$ to (1) which satisfy $x(t) = x$ coincides with $Q(t, x)$, then (under natural measurability assumptions) this is the same as

$$Q(t, x) = f(t, x, U) \quad \forall t, x. \tag{3}$$

According to the terminology used in [6], this equality means that *f represents Q*. The principal results to be proved here are just representation theorems for Q or certain portions of Q. The main theorem of [6] is crucial for proving them, but there is an important difference in the situations considered. In [6] we studied multifunctions $M(t)$ depending on one argument ranging through a measurable space. Here the multifunction depends on two variables t, x, the second taken from a space with a different structure (say, topological, metric or differentiable) and we seek a representation depending accordingly on each variable.

To briefly explain the nature of the main new assumption that appears, let us suppose for a while that we are interested in such an f which is C^1 in x. If (3) holds, then, whenever t and u are fixed, the mapping $x \to f(t, x, u)$ is a C^1-selection of $Q(t, \cdot)$. Thus, a necessary condition for (3) to be valid with f continuously differentiable in x is that for any (t, x_0) and $y_0 \in Q(t, x_0)$, the set-valued mapping $x \to Q(t, x)$ has a

* Received by the editors March 2, 1981, and in revised form February 20, 1982.

† Profosyuznaya 85-1-203, Moscow 117279, U.S.S.R.

C^1-selection $y(x)$ such that $y(x_0)=y_0$. Our principal result is that this condition (together with natural measurability assumptions on Q) is also sufficient.

This fact is contained in Theorem 1. The second theorem deals with convex-valued Q in which case an f exists which, in addition, is linear in u (cf. [6, Cor. 1.4]). Finally, in Theorem 3 we are interested in the existence of an "approximate representation" such that $f(t, x, U)$ is close to $Q(t, x)$ in a certain sense. The point is that "exact representations" like those in (3) are usually possible with U having rather a complicated structure (say, a closed set in a zero-dimensional, uncountable Polish space or a convex subset of an infinite dimensional Fréchet space) and only an approximate representation can be obtained with U being a "simple" set in a finite dimensional linear space.

The paper is concluded by a brief discussion of some possible applications. We show in particular that differential inclusions with local sections introduced by Boltyanskii [1] can be always equivalently represented by differential equations with control. It follows that the maximum principle established by Boltyanskii for optimal control problems involving such inclusions is a direct corollary of the standard Pontryagin maximum principle.

The results presented in the paper were essentially obtained in 1977 and partly announced in [7].

2. Main theorems. To begin with, we shall briefly introduce necessary notation and definitions to be used throughout the paper.

We shall be dealing with a measurable space $(T, \mathcal{M})$, a locally compact metrizable space X and a complete metric space (E, ρ). For any Polish space Z, we denote by $\mathcal{B}(Z)$ the collection of Borel subsets of Z; $\mathcal{M}\otimes\mathcal{B}(Z)$ will denote the product σ-algebra generated by all rectangles $G\times B$, where $G\in\mathcal{M}$, $B\in\mathcal{B}(Z)$. We shall say that $\mathcal{M}$ is a Souslin algebra if it is stable under the A-operation of Souslin or, equivalently, if $\mathcal{M}$ contains projections on T of $\mathcal{M}\otimes\mathcal{B}(Z)$-measurable sets, where Z is Polish. A set-valued mapping M from T into a topological space Z is called $\mathcal{M}$-measurable if

$$M^{-}(U)=\{t|M(t)\cap U\neq\varnothing\}\in\mathcal{M}$$

for any open $U\subset Z$. If (Z, d) is a metric space, this is equivalent to

$$d(z, M(t)) \text{ is } \mathcal{M}\text{-measurable in } t \text{ for any } z\in Z.$$

Usually we write simply "measurable" instead of "$\mathcal{M}$-measurable"

By $C(X, E)$, or simply by C, we denote the space of all continuous mappings from X into E with the topology of uniform convergence on compact subsets of X. Inasmuch as X is locally compact metrizable and E is Polish, $C(X, E)$ is itself a Polish space.

We shall also consider a closed-valued mapping Q from $T\times X$ into E and a collection L of continuous mappings from X into E. An *L-selection* of a set-valued mapping M from X into E is a $\xi(\cdot)\in L$ such that $\xi(x)\in M(x)$ all x. We set

$$Q_L(t, x)=\{\xi\in E|\xi=\xi(x) \text{ for an } L\text{-selection } \xi(\cdot) \text{ of } Q(t, \cdot)\}.$$

The following assumptions on L and Q will be adopted:

(H_1) for any $t\in T$, there is at least one L-selection of $Q(t, \cdot)$;

(H_2) L is a Polish space and the imbedding $\pi: L\to C(X, E)$ is continuous;

(H_3) $\mathcal{M}$ is a Souslin algebra;

(H_4) Q is $\mathcal{M}\otimes\mathcal{B}(X)$-measurable;

(H_5) for any $x \in X$, the set-valued mapping $t \to Q(t, x)$ is measurable, and for any $t \in T$, $\xi \in E$, the function $x \to \rho(Q(t, x), \xi)$ is lower semicontinuous.

THEOREM 1. *Assume* (H_1), (H_2) *and either* (H_5) *or* (H_3) *and* (H_4). *Then there are a Polish space* Z *and a mapping* $f: T \times X \times Z \to E$ *such that*

(a) f *is measurable in* t *for any* (x, z);

(b) $f(t, \cdot, z)$ *belongs to* L *for any* (t, z) *and the family* $\{\varphi_t | t \in T\}$ *of mappings from* Z *into* L *defined by* $\varphi_t: z \to f(t, \cdot, z)$ *is uniformly equicontinuous*;

(c) $f(t, x, Z) = Q_L(t, x)$ *for all* (t, x).

Proof. Fix a compact set $K \subset X$ and set

$$p_K(t, \xi(\cdot)) = \sup_{x \in K} \rho(Q(t, x), \xi(x)).$$

This function is obviously continuous in $\xi(\cdot)$ on $C(X, E)$ (and hence on L thanks to (H_2)).

On the other hand, p_K is measurable in t for any $\xi(\cdot) \in C$. Indeed, if (H_5) is valid, then $q(t, x) = \rho(Q(t, x), \xi(x))$ is measurable in t and l.s.c. in x whenever $\xi(\cdot) \in C$, the latter because

$$\begin{aligned} &|\rho(Q(t, x'), \xi(x')) - \rho(Q(t, x), \xi(x))| \\ &\leqq |\rho(Q(t, x'), \xi(x)) - \rho(Q(t, x), \xi(x))| + \rho(\xi(x'), \xi(x)). \end{aligned}$$

Therefore,

$$\sup_{x \in K} q(t, x) = \sup_{x \in K'} q(t, x)$$

whenever K' is a dense countable subset of K so that $p_K(t, \xi(\cdot))$ is the upper bound of a countable family $\{q(t, x) | x \in K'\}$ of measurable functions.

If, on the other hand, (H_3) and (H_4) are satisfied, then $q(t, x)$ is $\mathcal{M} \otimes \mathcal{B}(X)$-measurable as the composition of a continuous and an $\mathcal{M} \otimes \mathcal{B}(X)$-measurable mapping. Then $\{t | p_K(t, \xi(\cdot)) > \varepsilon\}$ is the projection on T of the $\mathcal{M} \otimes \mathcal{B}(X)$-measurable set $\{t, x | q(t, x) > \varepsilon$, hence belonging to $\mathcal{M}$ since $\mathcal{M}$ is a Souslin algebra.

We note further that p_K is nonnegative and $p_K(t, \xi(\cdot)) = 0$ if and only if $\xi(x) \in Q(t, x)$ for all $x \in K$.

Choose a countable family $\{K_i\}$ of compact subsets of X such that $X = \cup K_i$, and let

$$p(t, \xi(\cdot)) = \sum_{i=1}^{\infty} 2^{-i} p_{K_i}(t, \xi(\cdot))(1 + p_{K_i}(t, \xi(\cdot)))^{-1}.$$

Then p is nonnegative, measurable in t, continuous in $\xi(\cdot)$ on $C(X, E)$ and $p(t, \xi(\cdot)) = 0$ if and only if $\xi(x)$ belongs to $Q(t, x)$ for all x. Thus

$$\psi(t) = \{\xi(\cdot) \in L | p(t, \xi(\cdot)) = 0\}$$

is just the collection of all L-selections of $Q(t, \cdot)$. According to (H_1), $\psi(t) \neq \varnothing$ for every t and, clearly, each $\psi(t)$ is a closed set (which follows from (H_2) since Q is closed-valued). We shall show that ψ is a measurable set-valued mapping.

Let U be an open subset of L and $\{\xi_1(\cdot), \xi_2(\cdot), \cdots\}$ a dense countable subset of U. Since $p(t, \xi(\cdot))$ is continuous in $\xi(\cdot)$ on L, we have for any $t \in T$

$$\begin{aligned} \{t | \psi(t) \cap U \neq \varnothing\} &= \{t | p(t, \xi(\cdot)) = 0 \text{ for a } \xi(\cdot) \in U\} \\ &= \bigcap_{\delta > 0} \bigcup_{i=1}^{\infty} \{t | p(t, \xi_i(\cdot)) < \delta\} \in \mathcal{M}. \end{aligned}$$

Thus ψ meets all the requirements of [6, Thm. 1], and we can find a Polish space Z and a mapping $F\colon T\times Z\to L$ such that

(i) F is measurable in t and continuous in z;

(ii) the family $\{\Phi_t|t\in T\}$ of mappings from Z into L defined by $\Phi_t\colon z\to F(t,z)$ is uniformly equicontinuous;

(iii) $\psi(t)=F(t,z)$ for every t.

It remains to set

$$f(t,x,z)=F(t,z)(x).$$

COROLLARY. *In addition to the assumptions of the theorem, let E and L be Banach spaces and Q a convex-valued mapping. Then the conclusion of Theorem* 1 *holds with Z being a convex closed set in a separable Banach space W and*

$$f(t,x,w)=(A(t)w)(x)+a(t)(x), \tag{4}$$

where $A(\cdot)$ is a mapping from T into the space of bounded linear operators from W into L such that $\|A(t)\|\leqq 1$ for all t and the mappings $t\to A(t)w$ and $t\to a(t)$ from T into L are measurable, the first for any w.

Proof. In this case ψ is a convex-valued mapping into a Banach space and, instead of [6, Thm. 1], we can apply [6, Cor. 1.4].

This result was announced in [7] (with certain additional assumptions which are in fact needless). The assumption that L is a Banach space seems to be natural only if X is compact. To extend the result to the case of noncompact X, additional (though not very restrictive) requirements have to be imposed:

(H_6) E is a separable Banach space, L is a separable Fréchet space and for any compact $K\subset X$ the factor space $L_K=L/N_K$, where

$$N_K=\{\xi(\cdot)\in L|\xi(x)=0 \text{ for } x\in K\}$$

is normable;

(H_7) for any locally finite covering $\{V_i|i\in I\}$ of X by open sets with compact closures, there is a partition of unity $\{m_i(\cdot)|i\in I\}$ subordinate to this covering and such that

$$\sum_{i\in I} m_i(\cdot)\xi_i(\cdot)\in L$$

whenever $\xi_i(\cdot)\in L$ for every $i\in I$;

(H_8) for any partition of unity $\{m_i(\cdot)\}$ chosen in accordance with (H_7), the mapping

$$\{\xi_i(\cdot)|i\in I\}\to\sum_{i\in I} m_i(\cdot)\xi_i(\cdot)$$

from L^I into L is continuous (L^I being considered with the product topology).

Remarks. (a) Subspaces N_K are closed because L is continuously imbedded into $C(X,E)$. Therefore any factor-space L_K is complete metrizable [2, Ch. 9, § 3]. Thus (H_6) actually implies that every L_K is a Banach space.

(b) Elements of L_K can be naturally thought of as restrictions of elements of L to K. Thus, the imbedding $\pi_K\colon L_K\to C(K,E)$ is well defined. If (H_2) is valid, this mapping is continuous.

(c) Since X is locally compact metrizable, any open locally finite covering of X is countable.

THEOREM 2. *We posit the hypotheses of Theorem* 1 *and* (H_6)–(H_8). *If in addition Q is convex-valued, then the conclusion of Theorem* 1 *holds with Z being a closed*

convex subset of a separable Fréchet space W and f being defined as in (4), *where the collection* $\{A(t)|t \in T\}$ *of continuous linear operators from W into L is equicontinuous and the mappings* $t \to A(t)w$ *and* $t \to a(t)$ *from T into L are measurable, the first for any* $w \in W$.

Proof. Fix a compact set $K \subset X$, and let $\psi_K(t)$ be the collection of L_K-selections of the restriction of $Q(t, \cdot)$ to K (which, in view of Remark (b) above, is the collection of the restrictions to K of those elements of L which satisfy $\xi(x) \in Q(t, x)$ for all $x \in K$).

The assumptions of the corollary will be obviously fulfilled if we take K and L_K instead of X and L. Let a Banach space W_K, a closed convex set $Z_K \subset W_K$ and mappings $A_K(\cdot)$ from T into the space of bounded linear operators from W_K into L_K and $a_K(t)$ from T into L_K be chosen in accordance with the conclusion of the corollary so as to have

$$Q_L(t, x) = Q_{L_K}(t, x) = (A_K(t)Z_K)(x) + a_K(t)(x) \quad \forall t \in T, \quad \forall x \in K.$$

Now let $\{V_i|1 \leqq i < \infty\}$ be a locally finite covering of X by open sets with compact closures K_i, and let $\{m_i(\cdot)\}$ be a partition of unity subordinate to this covering and chosen in accordance with (H_7). We set

$$W = \prod_{i=1} W_i, \qquad Z = \prod_{i=1} Z_i$$

(W being considered with the product topology),

$$a(t) = \sum_{i=1}^{\infty} m_i(\cdot)a_i(t),$$

$$A(t)w = \sum_{i=1}^{\infty} m_i(\cdot)(A_i(t)w_i) \qquad (w = (w_1, w_2, \cdots))$$

(for simplicity, we have denoted $W_i = W_{K_i}$, $Z_i = Z_{K_i}$, etc.).

It is easy to see that W, Z, $A(\cdot)$ and $a(\cdot)$ have all the properties listed in the statement. Indeed, W is a separable Fréchet space as a countable product of Banach spaces, Z is closed and convex, $a(t)$ and $A(t)w$ belong to L thanks to (H_7) and are measurable in t. The family $\{A(t)|t \in T\}$ is an equicontinuous family of continuous linear operators in view of (H_8) and because $\|A_i(t)\| \leqq 1$ for all t and i.

Finally, if $\xi \in Q_L(t, x)$, then, whenever i is such that $x \in V_i$, there is $z_i \in Z_i$ such that $\xi = (A_i(t)z_i)(x) + a_i(t)(x)$. Taking arbitrary $z_i \in Z_i$ for other i and setting $z = (z_1, z_2, \cdots)$, we have

$$\xi = \sum_{i=1}^{\infty} m_i(x)\xi = \sum_{i=1}^{\infty} m_i(x)((A_i(t)z_i)(x) + a_i(t)(x))$$

$$= (A(t)z) + a(t)(x).$$

This completes the proof.

Remarks. (d) Both theorems are trivial if Q does not depend on t; in which case one can take Z to be the collection of all L-selections of Q and define f by

$$f(x, z) = z(x).$$

(e) The space W may be infinite dimensional even if $\dim E < \infty$; likewise Z may be noncompact even if Q is compact-valued. It is not difficult to make up an example where such a Z does not exist at all. For instance, let $X = [-1, 1]$, $E = R$, $L = C^1$, the space of continuously differentiable functions and $Q(x) = \{y \in R|0 \leqq y \leqq |x|\}$. Then the assumptions of both theorems are satisfied and $Q_L(x) = Q(x)$ for all x. However, no

compact Z can exist such that there is a mapping $f: [-1, 1]\times Z \to R$ with both f and $\partial f/\partial x$ jointly continuous and satisfying $f(x, Z) = [0, |x|]$ for all x. (Indeed, in this case we would have $(\partial f/\partial x)(0, z) \equiv 0$ since $f(x, z) \geqq 0$ and $f(0, z) = 0$. By virtue of the compactness of Z, it follows that $\max f(x, z) = o(x) \neq |x|$.)

However approximate representations with finite dimensional W (if E is also) and compact Z (if Q is compact-valued) do exist. To prove a corresponding result, we shall introduce the following notation:

B^n is the closed unit ball in R^n;

S^{n-1} is the unit sphere in R^n;

$\Sigma^k = \{a = (\alpha_1, \cdots, \alpha_k) | \alpha_i \geqq 0, \Sigma\alpha_i = 1\}$ is the standard $(k-1)$-simplex;

$(\cdot, \cdot)$ is the inner product in R^n;

$|\cdot|$ is the Euclidean norm in R^n;

$h(P, S)$ is the Hausdorff distance between the sets P and S; diam P is the diameter of P;

$s(P, \eta) = \sup\{(\xi, \eta) | \xi \in P\}$ is the support function of P.

THEOREM 3. *We adopt the following assumptions*:

(i) (H_2), (H_7);

(ii) $E = R^n$;

(iii) *Q is a convex and compact-valued multifunction measurable in t and Hausdorff continuous in x;*

(iv) *$Q_L(t, x)$ is dense in $Q(t, x)$ for any (t, x).*

Then for any $\varepsilon > 0$, $\delta > 0$ there are an integer $k > 0$ (depending only on n and ε) and a mapping $g(t, x, a): T \times X \times \Sigma^k \to R^n$ which is measurable in t, continuous in (x, a), affine in a, belongs to L as a function of x and satisfies

$$g(t, x, \Sigma^k) \subset Q(t, x) \subset g(t, x, \Sigma^k) + (\varepsilon \text{ diam } Q(t, x) + \delta)B^n$$

for any (t, x).

The three lemmas to follow will be proved first.

LEMMA 1. *Let $P \subseteq R^n$ be a bounded convex set, and let $(\eta_1, \cdots, \eta_k)$ be an ε-net in S^{n-1}. Choose $\xi_i \in P$ in such a way that*

$$(\xi_i, \eta_i) \geqq s(P, \eta_i) - \delta, \qquad i = 1, \cdots, k,$$

and let P' be the convex hull of $\{\xi_1, \cdots, \xi_k\}$. Then

$$h(P, P') \leqq \varepsilon \text{ diam } P + \delta.$$

Proof. Let $\xi \in P \backslash P'$. Since P' is closed and convex, there is a unique $\xi' \in P'$ closest to ξ. We have setting $\eta = (\xi - \xi')/|\xi - \xi'|$:

$$\eta \in S^{n-1}, \quad (\eta, \xi') = s(P', \eta), \quad (\eta, \xi - \xi') = \rho(P', \xi).$$

In particular,

$$\rho(P', \xi) = (\eta, \xi - \xi_i) + (\eta, \xi_i - \xi') \quad \forall i = 1, \cdots, k.$$

Chose such an i that $|\eta - \eta_i| < \varepsilon$. Since both ξ_i and ξ' belong to P', we have

$$(\eta, \xi_i - \xi') = (\eta, \xi_i) - s(P', \eta) \leqq 0.$$

On the other hand, since both ξ and ξ_i belong to P,

$$\begin{aligned}(\eta, \xi - \xi_i) &\leqq (\eta_i, \xi - \xi_i) + \varepsilon \text{ diam } P \\ &\leqq (\eta_i, \xi) - s(P, \eta_i) + \delta + \varepsilon \text{ diam } P \\ &\leqq \varepsilon \text{ diam } P + \delta.\end{aligned}$$

Thus

$$\rho(P', \xi) \leqq \varepsilon \operatorname{diam} P + \delta$$

for any $\xi \in P \backslash P'$ which yields the desired inequality because $P' \subset P$.

LEMMA 2. *Let $M(x)$ be a Hausdorff continuous multifunction with convex and compact values from X into R^n. Assume that $M_L(x)$ is dense in $M(x)$ for any x and* (H_7) *is valid. Then for any $\eta \in S^{n-1}$ and $\delta > 0$ there is a $\xi(\cdot) \in L$ such that $\xi(x) \in M(x)$ for all x and*

$$s(M(x), \eta) \leqq (\xi(x), \eta) + \delta \quad \forall x \in X.$$

Proof. Since $M_L(x)$ are dense in $M(x)$, for any $u \in X$ there is an L-selection $\xi_u(\cdot)$ of M such that

$$(\xi_u(u), \eta) \geqq s(M(u), \eta) - \frac{\delta}{2}.$$

The function $x \to s(M(x), \eta)$ is continuous because M is Hausdorff continuous. Therefore, we can find a family $\{V_u | u \in X\}$ of neighbourhoods with compact closures such that

$$(\xi_u(x), \eta) \geqq s(M(x), \eta) - \delta \quad \forall u \in X, \quad \forall x \in V_u.$$

This family obviously covers X, and it remains to take a locally finite subcovering $\{V_{u_i}\}$, choose a partition of unity $\{m_i(\cdot)\}$ in accordance with (H_7) and set

$$\xi(x) = \sum_i m_i(x) \xi_{u_i}(x).$$

LEMMA 3. *Under the assumptions of Theorem* 3, *for any $\delta > 0$, $\eta \in S^{n-1}$ there is a mapping $\xi(t, x)$: $T \times X \to R^n$ which is measurable in t, belongs to L as a function of x and satisfies*

$$s(Q(t, x), \eta) \leqq (\xi(t, x), \eta) + \delta. \tag{5}$$

Proof. Fix a compact set $K \subset X$ and consider the function

$$r_K(t, \xi(\cdot)) = \sup_{x \in K} (s(Q(t, x), \eta) - (\xi(x), \eta)).$$

This function is obviously continuous in $\xi(\cdot)$ on $C(X, R^n)$ and hence on L. On the other hand, since Q is measurable in t and Hausdorff continuous in x, the function $s(Q(t, x), \eta) - (\xi(x), \eta)$ (for a fixed $\xi(\cdot)$) is measurable in t and continuous in x so that r_K is measurable in t (see the proof of Theorem 1).

Consider also the function

$$q_K(t, \xi(\cdot)) = \max \{0, r_K(t, \xi(\cdot)) - \delta\}.$$

It is also measurable in t and continuous in $\xi(\cdot)$. Moreover, q_K is nonnegative and $q_K(t, \xi(\cdot)) = 0$ if and only if $r_K(t, \xi(\cdot)) \leqq \delta$, which is the same as

$$s(Q(t, x), \eta) \leqq (\xi(x), \eta) + \delta \quad \forall x \in K.$$

Let now $\{K_i | i = 1, 2, \cdots\}$ be a countable collection of compact subset of X which covers X. We set

$$q(t, \xi(\cdot)) = \sum_{i=1}^{\infty} 2^{-i} q_{K_i}(t, \xi(\cdot))(1 + q_{K_i}(t, \xi(\cdot)))^{-1}.$$

This function is measurable in t, continuous in $\xi(\cdot)$ and nonnegative, too, and $q(t, \xi(\cdot)) = 0$ if and only if $q_{K_i}(t, \xi(\cdot)) = 0$ for all i, or, equivalently, if and only if

$$s(Q(t, x), \eta) \leqq (\xi(x), \eta) + \delta \quad \forall x \in X. \tag{6}$$

Let $p(t, (\cdot))$ be the same as in the proof of Theorem 1. Since Q is measurable in t and Hausdorff continuous in x, (H_5) holds and therefore p is measurable in t, continuous in x, nonnegative and $p(t, \xi(\cdot)) = 0$ if and only if

$$\xi(x) \in Q(t, x) \quad \forall x \in X. \tag{7}$$

The set-valued mapping

$$\Phi(t) = \{\xi(\cdot) \in L | p(t, \xi(\cdot)) + q(t, \xi(\cdot)) = 0\}$$

is therefore closed-valued and measurable (see the proof of Theorem 1). Since p and q are nonnegative, $\Phi(t)$ contains precisely those $\xi(\cdot) \in L$ which satisfy (6), (7), hence $\Phi(t) \neq \varnothing$ by Lemma 2. It remains to take a measurable selection of Φ.

Proof of Theorem 3. Let $\eta_1, \cdots, \eta_k$ be an ε-net in S^{n-1}. By Lemma 3 there are $\xi_i(t, x)$: $T \times T \to R^n$ which are measurable in t, belong to L as functions of x and satisfy

$$\xi_i(t, x) \in Q(t, x), \qquad s(Q(t, x), \eta_i) \leqq (\xi_i(t, x), \eta_i) + \delta$$

for all (t, x). It remains to set

$$g(t, x, a) = \sum_{i=1}^{k} \alpha_i \xi_i(t, x)$$

and apply Lemma 1.

THEOREM 3 allows us to obtain various "unconditional" approximation results. Here is one of them.

COROLLARY. *Let X be an open domain in a finite dimensional Euclidean space, let $E = R^n$ and let $Q(t, x)$ satisfy condition* (iii) *of Theorem* 3. *Assume that for any $t \in T$, $x \in X$ and for any y of a dense subset of $Q(t, x)$ there are $\lambda > 0$ and a C^r-mapping ξ: $X \to E$ such that $\xi(x) = y$ and $\xi(u) \in Q(t, u)$ for all $u \in X$, $\|u - x\| < \lambda$.*

Then, whenever X' is a compact subset of X, for any $\varepsilon > 0$, $\delta > 0$ there are an integer $k = k(\varepsilon)$ and a mapping $g(t, x, a)$: $T \times T' \times \Sigma^k \to E$ which is measurable in t, continuous in (x, a), affine in a, C^r in x and satisfies the conclusion of Theorem 3 *for all $t \in T$, $x \in X'$.*

Remarks. (f) It is to be noted in connection with the previous corollary that in Theorems 2 and 3, (H_1) can be replaced by its local version: for any $t \in T$, $x_0 \in X$, there is a $\xi(\cdot) \in L$ such that $\xi(x) \in Q(t, x)$ for x close to x_0 and $\xi(x_0)$ is a given element of $Q(t, x_0)$.

(g) Except for (H_1), all of the hypotheses that have been used are rather natural and general. Anyway, if we consider the differential inclusion(1), they are automatically satisfied.

Indeed, in this case, T is a real segment with Lebesgue measure, hence (H_3) holds, X is a finite dimensional domain and L is the space of Lipschitz or C^1-mappings so that (H_6), (H_7) and (H_8) also follow; (H_4) of course, imposes no practical restrictions and (H_5) is needless since (H_3) and (H_4) are valid. The Hausdorff continuity assumption in Theorem 3 is somewhat restrictive but even this assumption is weaker than those usually imposed to characterize the dependence of the set-valued mapping in the inclusion on the state variable.

Thus (H_1) is the crucial hypothesis. As was pointed out in the introduction, it is also necessary for the results to hold. The weak point, however, is that there are no

general criteria to verify the hypothesis though in a particular case it may be an easy task. The only general result of such type known to the author is the selection theorem of Michael [8]. It can be applied only in case $L = C$, and we usually need more when dealing with differential inclusions.

3. Applications.

3.1. Selections. The following proposition is an immediate consequence of Theorem 1.

PROPOSITION 1. *Under the assumptions of Theorem* 1 *there is a selection of* Q *which is measurable in* t *and belongs to* L *as a function of* x.

As a particular case, we have

PROPOSITION 2. *Assume that* E *is a separable Banach space,* Q *is convex-valued, lower semicontinuous in* x *and either* (H_3), (H_4) *or* (H_5) *are satisfied. Then there is a Carathéodory selection of* Q, *that is to say, a mapping* $\xi(t, x)$: $T \times X \to E$ *measurable in* t, *continuous in* x *and satisfying* $\xi(t, x) \in Q(t, x)$ *for all* t, x.

Proof. We set $L = C(X, E)$ which ensures (H_2). The lower semicontinuity and convexity assumptions provide for (H_1), thanks to the Michael selection theorem. It remains to apply Proposition 1.

Earlier results of such sort were proved by Castaing [3], [4] and Cellina [5] (the latter was not availablle to me). The first of Castaing's results follows from Proposition 2. It is actually required there that (H_3), (H_4), (H_5) be satisfied simultaneously. The second result of Castaing is not contained in Proposition 2 for X is allowed to be Polish there, not necessarily locally compact. Instead, T is assumed Polish with $\mathcal{M}$ being a σ-algebra of measurable sets connected with a positive finite Radon measure on T. (No assumption similar to (H_5) is used.)

It would be interesting to know if a locally compact X may be replaced by a Polish space both in Propositions 1 and 2. Our proof obviously does not permit such an extension (because of the separability requirement on L). Likewise, Castaing's proof heavily depends on the topological properties of T and $\mathcal{M}$ and cannot be extended to a broader class of measurable spaces. Of course, in the situation of Proposition 2, Theorem 1 gives much more information than the proposition.

3.2. Equivalence theorem for differential inclusions. From now on, T is a real interval, $\mathcal{M}$ is the σ-algebra of Lebesgue measurable subsets of T and X is a domain in R^n. We consider again the differential inclusion

$$\dot{x} \subset Q(t, x). \tag{1}$$

Let L be the collection of all locally Lipschitz (or, say, C^1-) mappings from X into R^n.

PROPOSITION 3. *Assume either* (H_4) *or* (H_5). *If* $Q_L(t, x) = Q(t, x)$ *for all* $t \in T$, $x \in X$, *then the inclusion* (1) *is equivalent to a control differential equation*

$$\dot{x} = f(t, x, u), \qquad u \in U, \tag{2}$$

where U *is a Polish space,* f: $T \times X \times U \to R^n$ *is measurable in* t, *continuous in* (x, u) *and locally Lipschitz* (*or* C^1-) *in* x.

Proof. As was pointed out in Remark (g), all of the assumptions of Theorem 1 are satisfied. Let U, f be a Polish space and a mapping $T \times X \times U \to R^n$ satisfying the conclusion of Theorem 1. Then any solution of (2) is, clearly, a solution of (1) (because $f(t, x, u) \in Q(t, x)$). Conversely, let $x(t)$ be a solution of (1) defined on a subinterval $\Delta \subset T$. The mapping $h(t, u) = f(t, x(t), u)$ is measurable in t and continuous in u so

that the set-valued mapping

$$M(t)=\{u\in U|\dot{x}(t)=h(t,u)\}$$

from Δ into U is closed-valued and measurable. If $u(t)$ is a measurable selection of M, then

$$\dot{x}(t)=f(t,x(t),u(t)) \quad \text{a.e. on } \Delta$$

which completes the proof.

Assume now that X is compact and the solution $x(t)$ of (1) is such that for almost every $t\in T$ there is a C^1-selection $\xi(\cdot)$ of $Q(t,\cdot)$ such that $\xi(x(t))=\dot{x}(t)$ and $\|\xi\cdot)\|_{C^1}\leqq k(t)$, where $k(\cdot)\geqq 0$ is a summable function. As follows from Theorem 1, the mapping $u\to f(t,\cdot,u)$ from U into C^1 is continuous so that $(t,u)\to f(t,\cdot,u)$ is a Carathéodory mapping.

On the other hand, as follows from the proof of Theorem 1, any C^1-selection of $Q(t,\cdot)$ can be obtained as $f(t,\cdot,u)$ if $u\in U$ is suitably chosen. Therefore

$$\inf\{\|f(t,\cdot,u)\|_{C^1}|u\in M(t)\}\leqq k(t),$$

and (see [9]) for any $\varepsilon>0$, there is a measurable selection $u(t)$ of $M(t)$ such that

$$\|f(t,\cdot,u(t)\|_{C^1}\leqq k(t)+\varepsilon \quad \text{a.e.}$$

If we set $\varphi(t,x)=f(t,x,u(t))$, then φ satisfies all the conditions which ensure continuous dependence of the solution $x(t)$ of the equation

$$\dot{x}=\varphi(t,x)$$

on initial conditions and parameters etc. In particular, one can embed $x(t)$ into a smooth family of solutions of the differential inclusion (1). We refer to [10] for earlier results of this type.

Thus any differential inclusion satisfying the assumption of the proposition can be equivalently described by a differential equation with control having good analytical properties if the inclusion is sufficiently good.

This class of inclusions contains, in particular, differential inclusions with "local sections" considered by Boltyanskii and we conclude that the maximum principle established by Boltyanskii follows immediately from the standard Pontryagin maximum principle by way of the equivalent reduction justified by the proposition and subsequent remark and applied in a neighbourhood of the optimal trajectory.

Moreover, it follows that the assumptions imposed in [1] are redundant. Boltyanskii assumes the existence of two kinds of C^1-selections of Q, the first like here and the other similar to the above-mentioned $\varphi(t,x)$ and connected with the given solution $x(t)$. Such a selection, however, necessarily exists, as we have just seen, if the inclusion is reasonably good.

We remark finally that in fact things are much simpler in the situation considered in [1] because only set-valued mappings $Q(x)$, not depending on t, are considered there. In this case one can take U to be the collection of all C^1-selections of Q and define f by

$$f(x,u)=u(x)$$

so that no reference to Theorem 1 is actually needed.

3.3. Concluding remarks. It would be interesting to find an internal characterization for set-valued maps that admit (local) Lipschitz selections through any point

of the graph. Simple examples show that such a map need not be Lipschitz or even u.s.c. (it is l.s.c. if compact-valued). On the other hand, it seems to be unknown whether any Lipschitz (or even convex-valued Lipschitz) set-valued map has a Lipschitz selection.

REFERENCES

[1] V. G. BOLTYANSKII, *Mathematical Methods in Optimal Control*, Nauka, Moscow, 1969 (in Russian).

[2] N. BOURBAKI, *Topologie générale*, Hermann, Paris.

[3] CH. CASTAING, *Sur l'existence des sections séparément mesurables et séparément continues d'une multi-application*, Travaux du Séminaire d'Analyse Convexe du Languedoc 5, 1975.

[4] ———, *A propos de l'existence des sections séparément mesurables et séparement continus d'une multi-application séparément mesurable et séparément semi-continue inférieurement*, Travaux du Séminaire d'Analyse Convexe, Univ. du Languedoc 6, 1976.

[5] A. CELLINA, *A selection theorem*, Sem. Math. Univ. Padova, 55 (1976), pp. 143–149.

[6] A. D. IOFFE, *Single-valued representation of set-valued mappings*, Trans. Amer. Math. Soc., 252 (1979), pp. 133–145.

[7] ———, *Representation theorems for multifunctions and analytic sets*, Bull. Amer. Math. Soc., 84 (1978), pp. 142–144.

[8] E. MICHAEL, *Continuous selections* 1, Ann. Math., 63 (1956), pp. 361–382.

[9] R. T. ROCKAFELLAR, *Measurable dependence of convex sets and functions on parameters*, J. Math. Anal. Appl., 28 (1969), pp. 4–25.

[10] G. I. STASSINOPOULOS AND R. B. VINTER, *Continuous dependence of solutions of differential inclusions on the right hand side with applications to stability of optimal control problems*, this Journal, 17 (1979), pp. 432–450.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 4, July 1983

0363-0129/83/2104-0009 $01.25/0

CONTROLLED MARKOV CHAINS AND STOCHASTIC NETWORKS*

VIVEK S. BORKAR†

Abstract. Controlled Markov chains with average cost criterion and with special cost and transition structures are studied. Existence of optimal stationary strategies is established for the average cost criterion. Corresponding dynamic programming equations are derived. A stochastic network problem that includes interconnected queues as a special case is described and studied within this framework.

Key words. stable strategies, optimal control, average cost, dynamic programming, stochastic networks, interconnected queues

Introduction. Existence of stationary optimal controls for countable state Markov chains with average cost and the corresponding dynamic programming conditions are by now classical problems and have been treated in many texts, such as [2], [8]. However, these results have been usually derived under certain assumptions on the transition matrix [2] or certain recurrence conditions [8]. For many practical problems such as those arising in controlled queues, these assumptions fail. Naturally enough, the usual argument of treating the average cost as a limiting case of the discounted cost as the discount factor approaches unity fails to work for these problems. An example of such complications is provided in the recent work of Rosberg, Varaiya and Walrand [7]. They take a rather unconventional approach of treating average cost as the limiting case of finite time-horizon problems, but the details of their argument depend heavily on the specifics of the problem they consider, viz., the control of two queues in tandem with a special cost structure. However, this and many other problems arising in applications do have the following two aspects in common: transitions are possible only between "neighboring" states and the cost of being in "far away" states is high. (This will be made more precise later.) The purpose of this paper is to use such special cost and transition structures to settle the existence-dynamic programming question for a large class of such problems.

The approach taken here goes from the general to the particular. Results are first established in an abstract discrete-time Markov chain setting §§ 1–4 with a simple extension to continuous-time § 5 and then they are applied to the specific problem of controlling a stochastic network § 7. The latter has provided the motivation for this work and includes as a special case interconnected queues. Among the potential applications of such a set-up, one has:

(1) Optimal sequencing of traffic signals in a transportation network based on sensor data on the traffic level in each link.
(2) Optimal time-sharing of common information channels in a multi-agent management decision system or a computer communication network based on backlog information.

In §§ 1–5, many standard terms and facts from the theory of Markov chains and of weak convergence of probability measures are implicitly used. Readers unfamiliar with these should consult [5] and [1], respectively.

1. The discrete-time control problem. Consider a controlled Markov chain $\{X_n, n = 1, 2, \cdots\}$ on a countable infinite state space S. Without any loss of generality, let $S = \{0, 1, 2, \cdots\}$. Let P_u denote the transition probability matrix, indexed by the control vector $u = [u_1, u_2, \cdots]$ such that for each $i, j \in S$, u_i is in some compact Polish

* Received by the editors February 12, 1982, and in revised form July 16, 1982.
† Tata Institute of Fundamental Research, P.O. Box 1234, Bangalore 560012, India.

space $D(i)$ and the (i, j)th element of P_u is $p(i, j, u_i) \in [0, 1]$ with $\sum_{j \in S} p(i, j, u_i) = 1$. The functions $p(i, j, \cdot)$ are assumed to be continuous. Let $L = \prod_{i \in S} D(i)$ with the product topology. Standard topological arguments show that L has a suitable metrization under which it is compact and Polish. A control strategy (CS) is a sequence $\{\xi_n\}$ of L-valued random variables such that for each n, $\xi_n = [\xi_n(1), \xi_n(2), \cdots]$ satisfies:

$$P(X_{n+1} = i / f_n \vee \sigma(\xi_n)) = p(X_n, i, \xi_n(X_n)) \quad \text{for } i \in S$$

where $f_n = \sigma(X_i, i \leqq n; \xi_j, j < n)$.

If the $\{\xi_n\}$ above is a sequence of i.i.d. random variables with a common distribution Φ, we call it a stationary randomized strategy (SRS) and denote it by $\gamma[\Phi]$. If Φ is a point mass concentrated at $\xi_0 \in L$, we call it a stationary strategy (SS), to be denoted by $\gamma\{\xi_0\}$. Note that under either an SRS or an SS, $\{X_n\}$ is a Markov chain on S with stationary transition probabilities. The transition probability for a transition, say, from i to j, under an SRS $\gamma[\Phi]$ or an SS $\gamma\{\xi_0\}$ will simply be $E_\Phi[p(i, j, \xi(i))]$, $p(i, j, \xi_0(i))$ respectively. (Here, $E_\Phi(\cdot)$ denotes the expectation with respect to Φ and $\xi = [\xi(1), \cdots]$ a dummy variable of integration.) Let $P[\Phi]$, $P\{\xi_0\}$ respectively, denote the corresponding transition probability matrices. If this Markov chain is positive recurrent as well, call the corresponding CS a stable SRS (write SSRS) or a stable SS (write SSS), as the case may be.

Let $k: S \to [0, \infty)$ be given. The objective is to find a CS that a.s. minimizes the cost

$$\limsup_{n \to \infty} \frac{1}{n} \sum_{m=1}^{n} k(X_m).$$

If such a CS exists, call it an optimal CS. Our primary aim is to show the existence of an optimal SSS. We make the following assumptions:

A1. For any $M \in [0, \infty)$, there is a finite integer N such that $k(i) \geqq M$ whenever $i \geqq N$.

A2. For each $i \in S$, there exists a subset R_i of S with finite cardinality such that for all $l \notin R_i$, we have

$$p(i, l, \cdot) \equiv 0.$$

A3. For any finite subset F of S and any integer M, there is a finite integer N such that whenever $i \geqq N$, the minimum path length from i to any state in F exceeds M.

A4. For each $u = [u_1, u_2, \cdots]$, S forms a single communicating class under P_u.

A5. There exists at least one SS which gives a.s. bounded asymptotic cost.

An example of a Markov chain satisfying A1–A3 is a chain on the integer lattice in $\mathcal{R}^n$ with no transition possible between two states unless they form the corners of a unit cube in $\mathcal{R}^n$ and the cost being simply the $\mathcal{R}^n$-norm of the state.

Remarks. (1) Clearly, A1 ensures that the SS in A5 must be an SSS. Since each SS(SSS) is also an SRS(SSRS), the statement of A5 continues to hold with SRS(SSRS) replacing SS.

(2) A4, needed here for technical reasons, seems partly dispensable. A3, though convenient, is unnecessary. More on this later.

(3) A2 has the following important consequence: Starting from any state in S, at most finitely many other states can be reached in a finite length of time. Thus for any function $f: S \to (-\infty, \infty)$, the conditional expectations $E(f(X_{n+m})/X_n)$ for n, m finite can be written as finite sums and hence make perfect sense even when $Ef(X_{n+m})$ is either undefined or unbounded. This fact will be of crucial use in § 4.

Let $\psi_n = (1/n)\sum_{m=1}^{n} k(X_m)$. A4 ensures that for an SSRS $\gamma[\Phi]$ or an SSS $\gamma\{\xi_0\}$, the Markov chain has unique invariant probability distribution. Call it $\pi[\Phi]$, $\pi\{\xi_0\}$ respectively. Then $\lim_{n\to\infty} \psi_n$ exists a.s. in $[0, \infty]$ and is a.s. equal to the expectation of $k(\cdot)$ under $\pi[\Phi]$ or $\pi\{\xi_0\}$ as the case may be. Call this limit $C[\Phi]$, $C\{\xi_0\}$ resp. (Note that we have allowed $+\infty$ as a possible value for these limits.) By A5, we can pick $H \in [0, \infty)$ such that the set of SSRS (SSS) for which $C[\Phi]$ ($C\{\xi_0\}$) does not exceed H is nonempty. Let $\beta(\alpha)$ denote the infimum over all SSRS(SSS) of $C[\Phi]$ ($C\{\xi_0\}$). Then $\beta \leqq \alpha \leqq H$. Let $\mathcal{O}_1 = \{\pi[\Phi] | C[\Phi] \leqq H\}$, $\mathcal{O}_2 = \{\pi\{\xi_0\} | C\{\xi_0\} \leqq H\}$. The task of proving the existence of an optimal SSS will be achieved in two steps:

(i) There exists an SSS $\gamma\{\xi_0\}$ such that $C\{\xi_0\} = \alpha$.

(ii) For any CS, $\liminf_{n\to\infty} \psi_n \geqq \beta \geqq \alpha$ a.s.

Note that the usual conditions under which such existence results are established imply uniform bounds on the mean hitting times of state 0 from any other state [4]. These fail precisely because of A3.

2. Preliminary results. For any event A, let I_A denote its indicator function. For $i \in S$, B a Borel set in $D(i)$, $\{\xi_n\}$ a CS, define, for $n = 1, 2, \cdots$,

$$\nu_n(i) = \frac{1}{n}\sum_{m=1}^{n} I_{\{X_m = i\}}, \qquad \mu_{ni}(B) = \sum_{m=1}^{n} I_{\{\xi_m(i) \in B, X_m = i\}} \Big/ \left(\sum_{m=1}^{n} I_{\{X_m = i\}}\right),$$

whenever the denominator is nonzero and $\tilde{\mu}_i(B)$ otherwise, where $\tilde{\mu}_i(\cdot)$ is an arbitrary probability measure on $D(i)$.

Clearly, for each sample path and each n, $\nu_n(\cdot)$, $\mu_{ni}(\cdot)$ are probability measures on S and $D(i)$ respectively.

The following result is well known [6, p. 53]:

LEMMA 2.1. *Let M_n, $n = 1, 2, \cdots$ be a zero-mean martingale with respect to some increasing family of σ-fields, such that $E(|M_{n+1} - M_n|^2)$ is bounded uniformly in n. Then $\lim_{n\to\infty} (M_n/n) = 0$ a.s.*

COROLLARY 2.1. *For any* CS $\{\xi_n\}$, *the following holds outside a set of probability zero: For all $i \in S$,*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=2}^{n} \left[I_{\{X_m = i\}} - \sum_{j \in S} I_{\{X_{m-1} = j\}} p(j, i, \xi_{m-1}(j)) \right] = 0.$$

Proof. Note that the summand in square brackets is a martingale difference sequence with respect to the progressive σ-fields generated by (X_n, ξ_n), $n = 1, 2, \cdots$, and apply the preceding lemma. □

LEMMA 2.2. *$\mathcal{O}_1$, $\mathcal{O}_2$ are sequentially compact in the topology of weak convergence of probability measures on S.*

Proof. A1 and the definition of $\mathcal{O}_1$ clearly imply that $\mathcal{O}_1$ is tight. Hence we only need to show that it is weakly closed. Let $\pi_n \triangleq \pi[\Phi_n]$, $n = 1, 2, \cdots$, be a sequence in $\mathcal{O}_1$ weakly converging to some probability measure π_∞ on S. Then it also converges pointwise (i.e., $\pi_n(i) \to \pi_\infty(i)$ for each $i \in S$) and by Scheffe's theorem (see, e.g., [1]), in total variation. Since Φ_n, $n = 1, 2, \cdots$, is a sequence of probability measures on the compact Polish space L, there exists a probability measure Φ_∞ on L such that $\Phi_n \to \Phi_\infty$ weakly along a subsequence (also called $\{n\}$ by abuse of notation). We have $\pi_n = \pi_n P[\Phi_n]$, $n = 1, 2, \cdots$, where π_n is written as an infinite row vector. We only need to show that $\pi_\infty = \pi_\infty P[\Phi_\infty]$ under the same notation. But

$$\pi_\infty - \pi_\infty P[\Phi_\infty] = (\pi_\infty - \pi_n) + (\pi_n - \pi_\infty)P[\Phi_n] + \pi_\infty(P[\Phi_n] - P[\Phi_\infty]).$$

Since $\pi_n \to \pi_\infty$ pointwise and in total variation, the first two terms on the right go to zero termwise as $n \to \infty$. Since $\Phi_n \to \Phi_\infty$ weakly and $P(i, j, \cdot)$ are continuous for all

(i, j); $P[\Phi_n] \to P[\Phi_\infty]$ termwise. Hence the last term on the right goes to zero termwise by the dominated convergence theorem. Sequential compactness of $\mathcal{O}_1$ follows. That of $\mathcal{O}_2$ follows similarly on noting that the weak limit of probability measures each concentrated at a point in L must also be a probability measure concentrated at a point in L. □

LEMMA 2.3. *There exist an* SSRS $\gamma[\Phi]$ *and an* SSS $\gamma\{\xi_0\}$ *such that*

$$C[\Phi] = \beta,\ C\{\xi_0\} = \alpha.$$

Proof. Let $\pi[\Phi_n]$, $n = 1, 2, \cdots$ be a sequence in $\mathcal{O}_1$ such that $C[\Phi_n]$ converges to β. By Lemma 2.2, we may assume that $\pi[\Phi_n] \to \pi[\Phi_\infty]$ weakly, and hence pointwise, for some probability measure Φ_∞ on L. Let $k_M(\cdot) = k(\cdot) \wedge M$ for $M \in [0, \infty)$. Write $\pi[\Phi_n] = [\pi_n(1), \pi_n(2), \cdots]$ for $n = 1, 2, \cdots, \infty$. Consider $k_M(\cdot)\pi_n(\cdot)$ as a bounded map from S to $[0, \infty)$. Fatou's lemma yields

$$\sum_{i \in S} k_M(i)\pi_\infty(i) \leqq \liminf_{n \to \infty} \sum_{i \in S} k_M(i)\pi_n(i) \leqq \liminf_{n \to \infty} \sum_{i \in S} k(i)\pi_n(i) = \beta.$$

Letting $M \to \infty$ in the leftmost term, we get

$$C[\Phi_\infty] \leqq \beta.$$

But $C[\Phi_\infty] \geqq \beta$. Hence the claim follows. The second claim follows similarly. □

LEMMA 2.4. *For any* CS, $\liminf_{n \to \infty} \psi_n \geqq \beta$ *a.s.*

Proof. Let N be the null set where the statement of Corollary 2.1 fails. Fix a sample path outside N. Let $\{n_m\}$ be a subsequence of $\{n\}$ along which

$$\psi_{n_m} \to \liminf_{n \to \infty} \psi_n.$$

The limit on the right-hand side can be assumed to be finite, otherwise there is nothing to prove. Note that for each n, ψ_n is the expectation of $k(\cdot)$ under the probability measure $\nu_n(\cdot)$. Suppose $\{\nu_{n_m}\}$ is not a tight family. Then there exists an $\varepsilon > 0$ such that for any finite integer M, there exists an m such that $\sum_{i=1}^{M} \nu_{n_m}(i) \leqq 1 - \varepsilon$. From this and A1, it is easy to deduce that $\psi_{n_m} \to +\infty$ along a further subsequence, leading to a contradiction. Hence $\{\nu_{n_m}\}$ is a tight family and has a weak sequential limit ν_∞ which is a probability measure on S. For each $i \in S$, $\{\mu_{ni}\}$ is a sequence of probability measures on a compact Polish space and hence is tight and therefore weakly sequentially compact. By a diagonal argument, we can pick a subsequence of $\{n_m\}$ such that along this subsequence,

$$\nu_{n_m} \to \nu_\infty, \tag{2.1}$$

$$\mu_{n_m i} \to \mu_{\infty i}, \qquad i \in S, \tag{2.2}$$

weakly, where $\mu_{\infty i}$ is some probability measure on $D(i)$ for each i. Let ψ_∞ denote the expectation of $k(\cdot)$ under ν_∞. An application of Fatou's lemma as in the proof of Lemma 2.3 yields

$$\psi_\infty \leqq \liminf_{m \to \infty} \psi_{n_m}.$$

Restrict attention to the subsequence of $\{n_m\}$ along which (2.1), (2.2) hold and call it $\{n_m\}$, again by abuse of notation. For $n = 1, 2, \cdots, \infty$, let Φ_n denote the product probability measure on L whose restriction to the ith projection is μ_{ni}. From (2.2), $\Phi_{n_m} \to \Phi_\infty$ weakly. Hence for each $i, j \in S$,

$$E_{\Phi_{n_m}}(p(i, j, \xi(i))) \to E_{\Phi_\infty}(p(i, j, \xi(i))),$$

where $E_{\Phi_n}(\cdot)$ denotes the expectation with respect to Φ_n, ξ being a dummy variable of integration. Thus $P[\Phi_n] \to P[\Phi_\infty]$, termwise. Write ν_n, $n = 1, 2, \cdots, \infty$, as infinite row vectors. After some rearrangement, Corollary 2.1 can be restated as

$$\lim_{n\to\infty} (\nu_{n+1} - \nu_n P[\Phi_n]) = 0 \tag{2.3}$$

termwise. But

$$\nu_n P[\Phi_n] - \nu_\infty P[\Phi_\infty] = (\nu_n - \nu_\infty)P[\Phi_n] + \nu_\infty(P[\Phi_n] - P[\Phi_\infty]).$$

By arguments identical to those in Lemma 2.2, both terms on the right go to zero termwise along $\{n_m\}$. From this, the fact that $\nu_{n_m} \to \nu_\infty$ termwise (clearly, $\nu_{n_m+1} \to \nu_\infty$ termwise as well) and (2.3), it then follows that

$$\nu_\infty = \nu_\infty P[\Phi_\infty].$$

Hence

$$\nu_\infty = \pi[\Phi_\infty] \quad \text{and} \quad \psi_\infty = C[\Phi_\infty] \geqq \beta. \qquad \square$$

It is clear that for the desired existence result, we only need to show $\beta = \alpha$, or, equivalently, $\beta \geqq \alpha$.

3. Main results. At the outset, we would like to remind the reader of Remark (3) of § 1, which will be implicitly, but critically used often in this section.

Fix an SSRS $\gamma[\Phi]$ in $\mathcal{C}_1$ and consider the corresponding positive recurrent Markov chain. For each $i \in S$ such that $i \neq 0$, let $a(i)$, $b(i)$ denote, respectively, the mean hitting time for state 0 starting from i and the mean hitting time for state i starting from 0. Let $a(0)$, $b(0)$ each equal the mean return (recurrence) time of state 0. Let $g(i) = E[\sum_{m=1}^{\tau} k(X_m) | X_1 = i]$ for $i \in S$, where τ is the first $m > 1$ such that $X_m = 0$. Then $g(i)$ is the mean cost incurred before hitting 0 if $i \neq 0$ and the mean cost incurred between consecutive returns to 0 if $i = 0$. From [5, Prop. 79, pp. 28–29], it follows that $a(i)$, $b(i)$ are finite for $i \in S$.

LEMMA 3.1. *For each $i \in S$, $g(i)$ is finite.*

Proof. Fix $i \in S$. Let τ_1 be the first m such that $X_m = i$. For $n = 1, 2, \cdots$, let τ_{2n} be the first m strictly exceeding τ_{2n-1} for which $X_m = 0$ and τ_{2n+1} the first m strictly exceeding τ_{2n} for which $X_m = i$. Then

$$\frac{(1/n)\sum_{j=1}^{n}(\sum_{m=\tau_{2j-1}}^{\tau_{2j}} k(X_m))}{(1/n)(\tau_{2n})} \leqq \frac{\sum_{m=1}^{\tau_{2n}} k(X_m)}{\tau_{2n}}.$$

Since $\tau_{2n} \to \infty$ a.s., taking limits as $n \to \infty$ on both sides of the above inequality, we get

$$0 \leqq \frac{g(i)}{a(i)+b(i)} \leqq C[\Phi].$$

Hence

$$g(i) \leqq C[\Phi](a(i)+b(i)). \qquad \square$$

The proof of the next corollary is similar.

COROLLARY 3.1. $C[\Phi] = g(0)/a(0)$.

LEMMA 3.2. *Let $v(i) = g(i) - C[\Phi]a(i)$, $i \in S$, and $V = [v(0), v(i), \cdots]^T$. Then V satisfies*

$$C[\Phi]1_c = (P[\Phi] - U)V + Q, \tag{3.1}$$

where 1_c *is the infinite column vector of all* 1's, U *is the infinite dimensional identity matrix and* Q *is the column vector whose ith element is* $k(i)$ *for* $i \in S$. *Moreover, any solution to* (3.1) *differs from* V *only by a scalar multiple of* 1_c.

Remark. It is clear that the addition to V of a scalar multiple of 1_c leaves (4.1) unaltered. Also, note that V depends on the SSRS being used and that, by Corollary 4.1, $v(0) \equiv 0$.

Proof. Let τ be the first $m > 1$ such that $X_m = 0$. Then

$$\begin{aligned} v(i) &= E\Big(\sum_{m=1}^{\tau} (k(X_m) - C[\Phi])/X_1 = i\Big) \\ &= k(i) - C[\Phi] + E\Big(\sum_{m=2}^{\tau} (k(X_m) - C[\Phi])/X_1 = i\Big) \\ &= k(i) - C[\Phi] + (P[\Phi])_i V, \end{aligned}$$

where $(P[\Phi])_i$ denotes the ith row of $P[\Phi]$. The first claim follows. Suppose $W = [w(0), w(1), \cdots]^T$ also satisfies (4.1). Then for $m = 1, 2, \cdots$,

$$C[\Phi] = E(w(X_{m+1})/X_m) - w(X_m) + k(X_m).$$

Summing over $m = 1, 2, \cdots, \tau$, and taking expectation conditioned on $X_1 = i$,

$$\begin{aligned} C[\Phi]a(i) &= E\Big(\sum_{m=1}^{\tau} (E(w(X_{m+1})/X_m) - w(X_m))/X_1 = i\Big) + E\Big(\sum_{m=1}^{\tau} k(X_m)/X_1 = i\Big) \\ &= E(w(X_2)/X_1 = 0) - w(i) + g(i). \end{aligned}$$

Hence $w(i) = v(i) + E(w(X_2)/X_1 = 0)$, the last term being a constant independent of i. The second claim now follows. □

LEMMA 3.3. *There exists a finite integer* M *such that* $v(i) > 0$ *for* $i \geqq M$.

Proof. Let $\varepsilon > 0$ and $B = \{0\} \cup \{i \in S | k(i) \leqq C[\Phi] + \varepsilon\}$. By A1, B has finite cardinality. Let $X_1 = i$ for some $i \notin B$. Let τ_i be the first $m > 1$ such that $X_m = 0$. Then

$$v(i) = E\Big(\sum_{m=1}^{\tau_i} (k(X_m) - C[\Phi])/X_1 = i\Big).$$

Let

$$\varepsilon_1 = \min_{i \in B} v(i).$$

The strong Markov property implies

$$v(i) = E\Big(\sum_{m=1}^{\tau_i' - 1} (k(X_m) - C[\Phi])\Big) + \sum_{j \in B} v(j) P(X_{\tau_i'} = j),$$

where τ_i' is the first m such that $X_m \in B$. Hence,

$$v(i) \geqq \varepsilon E(\tau_i' - 1) + \varepsilon_1.$$

By A3,

$$E\tau_i' \to \infty \quad \text{as } i \to \infty.$$

The claim follows. □

THEOREM 3.1. *An optimal* SSS *exists.*

Proof. Choose $\gamma[\Phi]$ in the above such that $C[\Phi] = \beta$. Let $u(i)$ for each $i \in S$ be an element of $D(i)$ that minimizes $\sum_{j \in S} p(i, j, \cdot) v(j)$. (Note that this is a finite sum

and the existence of at least one such $u(i)$ is guaranteed by the usual compactness-continuity arguments.) Let $l(i)$ be the ith element of $(P[\Phi]-U)V$. For each $n = 1, 2, \cdots$, and any CS,

$$\beta = l(X_n)+k(X_n) \geqq \sum_{j\in S} P(X_n, j, u(X_n))v(j)-v(X_n)+k(X_n). \tag{3.2}$$

Let $\xi_0=[u(0), u(1), \cdots]$ and consider the chain governed by the SS $\gamma\{\xi_0\}$. From (3.2),

$$\beta \geqq \frac{1}{n}E(v(X_{n+1})/X_n)-\frac{1}{n}\sum_{m=2}^{n}(v(X_m)-E(v(X_m)/X_{m-1}))-\frac{1}{n}v(X_1)+\psi_n.$$

Hence,

$$\begin{aligned}\beta &\geqq \frac{1}{n}E(v(X_{n+1})/X_1)-\frac{1}{n}\sum_{m=2}^{n}(E(v(X_m)/X_1)-E(v(X_m)/X_1))-\frac{1}{n}v(X_1)+E(\psi_n/X_1)\\ &=\frac{1}{n}E(v(X_{n+1})/X_1)-\frac{1}{n}v(X_1)+E(\psi_n/X_1)\\ &\geqq \frac{1}{n}(E(v(X_{n+1})I_{\{X_{n+1}\leqq M\}}/X_1)-v(X_1))+E(\psi_n/X_1),\end{aligned}$$

where M is as in the preceding lemma. Since $v(X_{n+1})I_{\{X_{n+1}\leqq M\}}$ is bounded, taking limits as $n\to\infty$ gives

$$\beta \geqq \limsup_{n\to\infty} E(\psi_n/X_1) \geqq E\left(\liminf_{n\to\infty}\psi_n/X_1\right),$$

the last inequality being easily deducible from Fatou's lemma.

Suppose $\gamma\{\xi_0\}$ is not stable. Then $\{X_n\}$ is not positive recurrent and the following holds: For any $\varepsilon>0$ and any finite integer N, $\sum_{i\leqq N}\nu_n(i)<\varepsilon$ for sufficiently large n, a.s. From A1, this implies that $\liminf_{n\to\infty}\psi_n=\infty$ a.s. This is clearly not possible. Hence $\gamma\{\xi_0\}$ is an SSS. Therefore for each $N\in\{1, 2, \cdots\}$, $\liminf_{n\to\infty}(1/n)\sum_{m=1}^{n}k(X_m)\wedge N = E(k(\cdot)\wedge N)$ a.s., where the expectation is with respect to $\pi\{\xi_0\}$. By the dominated convergence theorem,

$$\beta \geqq E\left(\liminf_{n\to\infty}\psi_n/X_1\right) \geqq E\left(\liminf_{n\to\infty}\frac{1}{n}\sum_{m=1}^{n}k(X_m)\wedge N/X_1\right) = E(k(\cdot)\wedge N).$$

Letting $N\to\infty$ on the right-hand side,

$$\beta \leqq E\left(\liminf_{n\to\infty}\psi_n/X_1\right) \geqq E(k(\cdot)) = C\{\xi_0\} \geqq \alpha.$$

In view of the comment at the end of the preceding section, the result follows. □

THEOREM 3.2. *Let $\gamma\{\xi_0\}$ be an optimal* SSS. *Let V be as in Lemma* 3.3 *with Φ = the probability measure concentrated at ξ_0. Then the following "dynamic programming" conditions hold*:

$$\alpha 1_c = \min_{u\in L}(P_u - U)V + Q = (P\{\xi_0\}-U)V+Q,$$

where the minimum is termwise.

Proof. Suppose that the first equality is false. Then there exists $\Gamma=[\Gamma_1, \Gamma_2, \cdots]^T \in L$, distinct from ξ_0, such that

$$\alpha 1_c = (P_\Gamma - U)V + Q + \Delta,$$

where $\Delta=[\Delta(0), \Delta(1), \cdots]^T$ is a vector of nonnegative elements at least one of which

is strictly positive. Proceeding as in the proof of Theorem 4.1, one can show that $\gamma\{\Gamma\}$ is an SSS and $\alpha \geqq C\{\Gamma\}+\delta$, where δ is the expectation of $\Delta(\cdot)$ under $\pi\{\Gamma\}$. It is easily seen that $\delta > 0$. Hence we have a contradiction. So the first equality holds. The second is immediate. □

Remarks. In the arguments so far, A4 was needed mainly for technical reasons. In the previous section, it was needed to ensure the uniqueness of $\pi[\Phi]$ ($\pi\{\xi_0\}$) corresponding to a given $\gamma[\Phi]$ ($\gamma\{\xi_0\}$). Without A4, the proofs still go through modulo this uniqueness. Thus we can show that there exists an SRS $\gamma[\Phi]$ (an SS $\gamma\{\xi_0\}$) and some probability measure $\pi[\Phi]$ ($\pi\{\xi_0\}$) that is invariant under it so that the expectation of k with respect to this probability measure is $\beta(\alpha)$, $\beta(\alpha)$ being the infimum of such expectations over all SRS (SS). The invariant probabilities must be concentrated on a union of disjoint communicating classes, each of which is positive recurrent. Also, the invariant probability measure is a weighted average of the unique invariant probability measure for each of these classes, these weights satisfying the obvious requirements that they are positive and add up to 1. The cost itself will be a weighted average (with the same weights) of the costs incurred if the chain started in any one of these classes. These costs must be identical, because otherwise the invariant probability measure concentrated on one of these classes would give a lower cost than what $\pi[\Phi]$ ($\pi\{\xi_0\}$) does. The arguments of this section, which lead to Theorem 3.1, 3.2 can be worked out for each individual class, yielding an optimal SSS if the chain started with probability 1 in one of these classes. Consider the following alternative to A4:

A4′. For at least one SRS, the chain is positive recurrent with only one communicating class.

The SRS guaranteed by A4′ can be used until the chain enters one of the desired classes under the optimal SSS. This is a stopping time with bounded expectation. After this instant, the strategy can be switched to the optimal SSS. The cost still equals the optimal cost under this scheme, as can be easily verified.

A3 holds in most cases arising in practice, but is quite unnecessary. Note that $v(i)$, $i \in S$, in Theorem 3.1, is bounded from below. Let $B' = \{i \in S, v(i) \leqq 0\}$. Then $v(i)$ is bounded in B'. Replace the indicator $I_{\{X_{n+1} \leqq m\}}$ by $I_{\{X_{n+1} \in B'\}}$ whenever the former occurs in the proof of Theorem 3.1. The same arguments continue to hold.

4. An extension to continuous time. Consider a continuous-time Markov chain X_t, $t \in [0, \infty)$, on S, described as follows: Associated with this chain is a transition probability matrix P_u as in § 1 and a function $r: S \to [d_1, d_2]$, $0 < d_1 \leqq d_2 < \infty$. Transitions occur at random times $\tau_1 < \tau_2 < \tau_3 \cdots$. These are easily seen to be stopping times with respect to the progressive σ-fields generated by X_t, $t \in [0, \infty)$. Let $\tau_0 = 0$. The random intervals $[\tau_n, \tau_{n+1})$, $n = 0, 1, 2, \cdots$ are called transition epochs. A control process $\xi(t)$, $t \in [0, \infty)$, is a process satisfying:

(i) for each t, $\xi(t) \in L$;

(ii) $\xi(t)$ is constant on each transition epoch;

(iii) $P(X_{\tau_{n+1}} = j / X_t, \xi(t); t < \tau_{n+1}) = p(X_{\tau_n}, j, \xi(\tau_n)(X_{\tau_n}))$, where $\xi(t)(i)$ is the ith component of $\xi(t)$ for $i \in S$.

Conditioned on X_{τ_n}, the random variable $\tau_{n+1} - \tau_n$ is assumed to be:

(i) independent of $\{X_t, t < \tau_n$ and $t \geqq \tau_{n+1}$; $\xi_t, t \in [0, \infty)\}$;

(ii) exponentially distributed with mean $1/r(X_{\tau_n})$.

Let $k(\cdot)$ be a function as in § 1 satisfying A1. The objective is to choose a control process so as to a.s. minimize

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t k(X_s)\, ds. \tag{4.1}$$

Let $\hat{X}_n = X_{\tau_n}$, $\xi_n = \xi(\tau_n)$, $n = 1, 2, \cdots$ and $\Lambda(\cdot) = 1/r(\cdot)$.

LEMMA 4.1. *Outside a set of probability zero,*

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t k(X_s)\,ds = \limsup_{n\to\infty} \frac{(1/n)\sum_{m=1}^n k(\hat{X}_m)\Lambda(\hat{X}_m)}{(1/n)\sum_{m=1}^n \Lambda(\hat{X}_m)}.$$

Proof. For $M \in [0, \infty)$, let $k_M(\cdot) = k(\cdot) \wedge M$. Define $n(t) = n$ when $t \in [\tau_n, \tau_{n+1})$. Then

$$\begin{aligned}\frac{1}{\tau_{n(t)+1}}\int_0^{\tau_{n(t)}} k_M(X_s)\,ds &= \frac{1}{\tau_{n(t)+1}}\sum_{m=0}^{n(t)-1} k_M(\hat{X}_m)(\tau_{m+1}-\tau_m)\\ &\leqq \frac{1}{t}\int_0^t k_M(X_s)\,ds \leqq \frac{1}{\tau_{n(t)}}\int_0^{\tau_{n(t)+1}} k_M(X_s)\,ds \qquad (4.2)\\ &= \frac{1}{\tau_{n(t)}}\sum_{m=0}^{n(t)} k_M(\hat{X}_m)(\tau_{m+1}-\tau_m).\end{aligned}$$

But

$$\begin{aligned}\limsup_{t\to\infty} \frac{1}{\tau_{n(t)+1}}\sum_{m=0}^{n(t)-1} k_M(\hat{X}_m)(\tau_{m+1}-\tau_m) &= \limsup_{n\to\infty} \frac{1}{\tau_{n+1}}\sum_{m=0}^{n-1} k_M(\hat{X}_m)(\tau_{m+1}-\tau_m)\\ &= \limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^{n-1} k_M(\hat{X}_m)(\tau_{m+1}-\tau_m)}{(1/n)\sum_{m=0}^{n}(\tau_{m+1}-\tau_m)}.\end{aligned}$$

By Lemma 2.1,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} [k_M(\hat{X}_m)(\tau_{m+1}-\tau_m) - k_M(\hat{X}_m)\Lambda(\hat{X}_m)] = 0 \quad \text{a.s.},$$

$$\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} [(\tau_{m+1}-\tau_m) - \Lambda(\hat{X}_m)] = 0 \quad \text{a.s.}$$

Since $\Lambda(\cdot)$ and hence $(1/n)\sum_{m=1}^n \Lambda(\hat{X}_m)$ is bounded away from zero from below and bounded from above, we have

$$\begin{aligned}\limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^{n-1} k_M(\hat{X}_m)(\tau_{m+1}-\tau_m)}{(1/n)\sum_{m=0}^{n}(\tau_{m+1}-\tau_m)} &= \limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^{n-1} k_M(\hat{X}_m)\Lambda(\hat{X}_m)}{(1/n)\sum_{m=0}^{n}\Lambda(\hat{X}_m)} \quad \text{a.s.}\\ &= \limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^{n} k_M(\hat{X}_m)\Lambda(\hat{X}_m)}{(1/n)\sum_{m=0}^{n}\Lambda(\hat{X}_m)}. \qquad (4.3)\end{aligned}$$

Similarly the right-hand side of (4.2) can be shown to equal (4.3). Hence, outside a set of zero probability,

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t k_M(X_s)\,ds = \limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^{n} k_M(\hat{X}_m)\Lambda(\hat{X}_m)}{(1/n)\sum_{m=0}^{n}\Lambda(\hat{X}_m)},$$

implying

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t k(X_s)\,ds \geqq \limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^{n} k_M(\hat{X}_m)\Lambda(\hat{X}_m)}{(1/n)\sum_{m=0}^{n}\Lambda(\hat{X}_m)}.$$

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t k_M(X_s)\,ds \leqq \limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^{n} k(\hat{X}_m)\Lambda(\hat{X}_m)}{(1/n)\sum_{m=0}^{n}\Lambda(\hat{X}_m)}.$$

Letting $M \to \infty$ along a countable subset of $[0, \infty)$ in the above, it is easily deduced that

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t k(X_s)\,ds = \limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^n k(\hat{X}_m)\Lambda(\hat{X}_m)}{(1/n)\sum_{m=0}^n \Lambda(\hat{X}_m)}$$

outside a set of zero probability. □

The concepts SS, SRS, SSRS, SSS introduced in § 2 have obvious analogues for the continuous-time problem above. In view of Lemma 4.1, this problem is equivalent to choosing a CS $\{\xi_n\}$ for a discrete time Markov chain $\hat{X}_n$, $n = 0, 1, 2, \cdots$ on S so as to a.s. minimize

$$\limsup_{n\to\infty} \frac{(1/n)\sum_{m=0}^n k(\hat{X}_m)\Lambda(\hat{X}_m)}{(1/n)\sum_{m=0}^n \Lambda(\hat{X}_m)}. \tag{4.4}$$

The following claims can be easily verified from the above lemma and the fact that $\Lambda(\cdot)$ is bounded away from zero from below and bounded from above:

(i) A2–A5 hold for the chain $\{\hat{X}_n\}$ if and only if they hold for the original chain.

(ii) An SS (resp. SRS, SSS, SSRS) for the chain $\{\hat{X}_n\}$ corresponds to an SS (resp. SRS, SSS, SSRS) for the original chain and vice versa.

(iii) A1 holds for $k(\cdot)\Lambda(\cdot)$ in place of $k(\cdot)$.

(iv) (4.4) is finite if and only if $\limsup_{n\to\infty}(1/n)\sum_{m=0}^n k(\hat{X}_m)$ is.

Assume that A2–A5 hold for the discrete time problem introduced above, and hence for the original continuous-time problem. Consider the discrete-time problem for the time being. Let $\phi_n = (1/n)\sum_{m=0}^n k(\hat{X}_m)\Lambda(\hat{X}_m)$, $\delta_n = (1/n)\sum_{m=0}^n \Lambda(\hat{X}_m)$. Under either an SSRS $\gamma[\Phi]$ or an SSS $\gamma\{\xi_0\}$, ϕ_n and δ_n are seen to a.s. converge to the expectations of $k(\cdot)\Lambda(\cdot)$, $\Lambda(\cdot)$ under $\pi[\Phi]$, $\pi\{\xi_0\}$ respectively. (We allow $+\infty$ as a possible value of the former.) The asymptotic cost, denoted as before by $C[\Phi]$, $C\{\xi_0\}$, respectively, will a.s. equal to the ratio of these two expectations. The statement and proof of Lemma 2.2 carry over to the present problem in view of (iii), (iv) above. The following analogue of Lemma 2.3 holds:

LEMMA 4.2. *There exist an* SSRS $\gamma[\Phi]$ *and an* SSS $\gamma\{\xi_0\}$ *such that* $C[\Phi] = \beta$, $C\{\xi_0\} = \alpha$.

Proof. Let $\pi[\Phi_n]$, $n = 1, 2, \cdots$ be a sequence in $\mathcal{O}_1$ such that $C[\Phi_n]$ monotonically decreases to β. As in the proof of Lemma 2.3, we let $\pi[\Phi_n] \to \pi[\Phi_\infty]$ weakly for some probability measure Φ_∞ on L. Let $E_n(\cdot)$, $n = 1, 2, \cdots, \infty$, denote the expectation under $\pi[\Phi_n]$. Then $E_n(\Lambda) \to E_\infty(\Lambda)$. By arguments similar to those used in the proof of Lemma 2.3, we can show that $E_\infty(k\Lambda) \leqq \liminf_{n\to\infty} E_n(k\Lambda)$. Since $C[\Phi_n] = E_n(k\Lambda)/E_n(\Lambda)$, we have

$$\beta \leqq C[\Phi_\infty] \leqq \liminf_{n\to\infty} C[\Phi_n] = \beta.$$

The first claim follows. The second follows similarly. □

A similar modification of the proof of Lemma 3.4 yields the following result.

LEMMA 4.3. *For any* CS, $\liminf_{n\to\infty}(\phi_n/\delta_n) \geqq \beta$ a.s.

Coming back to the continuous-time chain X_t, $t \in [0, \infty)$, we can define for any SSRS $\gamma[\Phi]$, the quantities $a(\cdot)$, $g(\cdot)$, $v(\cdot)$ as follows: Let τ be the first $t > 0$ such that $x_t = 0$ if $i \neq 0$; and the first $t > 0$ such that $X_s \neq 0$ for some $t > s > 0$ and $X_t = 0$, if $i = 0$. For $i \in S$, let $a(i) = E(\tau/X_0 = i)$, $g(i) = E(\int_0^\tau k(X_s)\,ds/X_0 = i)$. Clearly, for each sample path, τ coincides with some τ_m, $m = 1, 2, \cdots$. Let η be the $m > 0$ such that $\tau_\eta = \tau$. Then η is a random variable taking values in $\{1, 2, 3, \cdots\}$ such that $g(i) = E(\sum_{i=0}^{\eta-1} k(\hat{X}_i)(\tau_{i+1} - \tau_i)/X_0 = i)$. Let $v(i) = g(i) - C[\Phi]a(i)$ and $V = [v(0), v(1), \cdots]$. Because of the observations following Lemma 4.1, the claims of Lemma 4.2 hold for

the original continuous-time chain as well. Let W denote the infinite column vector whose ith element is $k(i)\Lambda(i)$ and Y the infinite column vector whose ith element is $\Lambda(i)$ for $i \in S$. We then have the following analogues of Lemma 3.2 and Theorems 3.1–3.2.

LEMMA 4.4. *Let $Y[\Phi]$ be an* SSRS *and $V = [v(0), v(1), \cdots]^T$ the vector defined as above. Then V satisfies*

$$C[\Phi]Y = (P[\Phi] - U)V + W \tag{4.5}$$

where $P[\Phi]$, U are as in § 3. *Moreover, V is the unique solution to* (4.5) *modulo the addition of a constant multiple of 1_c.*

Proof. The arguments here are a simple modification of those used to prove Lemma 3.2. We have, for η as above

$$\begin{aligned} v(i) &= E\left(\sum_{m=0}^{\eta-1} (k(\hat{X}_m) - C[\Phi])(\tau_{m+1} - \tau_m) / X_0 = i \right) \\ &= k(i)\Lambda(i) - C[\Phi]\Lambda(i) + E\left(\sum_{m=1}^{\eta-1} (k(\hat{X}_m) - C[\Phi])(\tau_{m+1} - \tau_m) / X_0 = i \right) \\ &= k(i)\Lambda(i) - C[\Phi]\Lambda(i) + (P[\Phi])_i V, \end{aligned}$$

with $(P[\Phi])_i$ denoting the ith row of $P[\Phi]$. The first claim follows. A similar modification of the second half of the proof of Lemma 3.2 establishes the second claim. □

THEOREM 4.1. *An optimal* SSS *exists. Suppose $\gamma\{\xi_0\}$ is an optimal* SSS *and V is as above with Φ = the probability measure concentrated at ξ_0. Then*

$$\alpha Y = \min_u (P_u - U)V + W = (P\{\xi_0\} - U)V + W,$$

where the minimum is termwise.

Proof. Again, the arguments are quite similar to those used to prove Theorems 4.1–4.2. Only the key steps are indicated. The claims of Lemma 3.3 can be easily shown to hold for the present problem. Let $u(i)$ for each $i \in S$ be an element of $D(i)$ that minimizes $\sum_{j \in S} p(i, j, \cdot)v(j)$. Consider the SS that uses $u(X_{\tau_n})$ over the transition epoch $[\tau_n, \tau_{n+1})$, $n = 1, 2, \cdots$. Along the lines of Theorem 3.1, one can show

$$\beta E(\delta_n / \hat{X}_0) \geqq \frac{1}{n} E(v(\hat{X}_{n+1}) / \hat{X}_0) - \frac{1}{n} v(\hat{X}_0) + E(\psi_n / \hat{X}_0),$$

and therefore,

$$\beta E\left(\limsup_{n \to \infty} \delta_n / \hat{X}_0 \right) \geqq E\left(\liminf_{n \to \infty} \psi_n / \hat{X}_0 \right).$$

Using the fact that $\{\delta_n\}$ is a bounded sequence and arguments analogous to those in Theorem 3.1, we can show that the SS $\xi_0 = [u(0), u(1), \cdots]$, for the chain $\{\hat{X}_n\}$, is actually an SSS with cost $E(k\Lambda)/E(\Lambda)$, $E(\cdot)$ being the expectation with respect to $\pi\{\xi_0\}$. Also, $\psi_n \to E(k\Lambda)$ a.s. and $\delta_n \to E(\Lambda)$ a.s. Then it is easy to deduce that

$$\beta \geqq E(k\Lambda)/E(\Lambda).$$

The first claim follows. The second claim follows on a similar modification of the proof of Theorem 3.2. □

Remarks. For many problems, the possibility of continuously updating the control process $u(t)$ exists, but $u(t)$ affects the probabilistic behaviour of the Markov chain only through its values at the transition times, e.g., the problem discussed in the next section. Then nothing is lost by considering the control to be constant over the transition epochs. See also [7] for a discussion of this issue.

5. Stochastic networks. A stochastic network consists of a collection of nodes $J_1, J_2, \cdots, J_{N+1}$. Nodes $J_1, J_2, \cdots, J_M$ for some $M \leqq N$ are identified as being input nodes. The node J_{N+1} is identified as "environment". Each input node J_m has a stream of customers entering it according to a Poisson process with rate $\lambda_m > 0$. These Poisson processes are independent of each other. Each entering customer undergoes random transitions from node to node, spending a random amount of time in each node he visits, till he gets eventually absorbed in J_{N+1}, all according to the mechanism described below.

Let $F = \{[a_1, a_2, \cdots, a_N] | a_i \text{ is an integer} \geqq 0 \text{ for } 1 \leqq i \leqq N\}$. A function $f_i : F \to [0, \infty)$ satisfying $f_i(a_1, a_2, \cdots, a_N) = 0$ if $a_i = 0$ is assigned to every node J_i. Let $f_{N+1}(\cdot) \equiv 0$. To each node J_i assign a subset R_i of $\{J_1, J_2, \cdots, J_{N+1}\}$ such that R_i does not contain J_i. (No confusion need arise between these and the R_i's in the statement of A2, since the latter will not be explicitly referred to in the remainder.) R_{N+1} is assumed to be the empty set. For each pair (i, j) such that $J_j \in R_i$, assign a function $p_{ij} : F \to [0, 1]$ and a process $u_{ij}(t)$, $t \in [0, \infty)$, taking values in the set $\{0, 1\}$. We assume that $\sum_{J_j \in R_i} p_{ij}(s) u_{ij}(t) \leqq 1$ for all $\{u_{ij}(t)\}$ satisfying the constraints to be specified later and all $s \in F$.

Let $x_i(t)$ denote the number of customers at node J_i at time t. Let $X_t = \{x_1(t), x_2(t), \cdots, x_N(t)\}$. Let $A_i(t)$ denote the event that a customer leaves the node J_i in the interval $[t, t+dt]$. $A_i(t)$, $i = 1, 2, \cdots, N+1$ are assumed to be conditionally independent of each other and of the arrival processes given $\{X_s, s \leqq t\}$, the conditional probability of $A_i(t)$ conditioned on $\{X_s, s \leqq t\}$ being $f_i(X_t)\, dt + o(dt)$. Heuristically, this ensures that conditioned on the past up to t, the processes $I_{\{A_i(t)\}}$ and the arrival processes act like independent Poisson processes on the infinitesimal interval $[t, t+dt]$. The probability of occurrence of $A_{N+1}(t)$ is $o(dt)$ and hence $A_{N+1}(t)$ can be ignored. The probability of more than one of $A_i(t)$, $i = 1, 2, \cdots, N$, taking place at any t is also $o(dt)$. Thus X_{t-}, defined as the value of X_s just prior to t, makes sense. The customer leaving node J_i goes to $J_j \in R_i$ with probability $p_{ij}(X_{t-}) u_{ij}(t)$ or returns to J_i with probability $1 - \sum_{J_j \in R_i} p_{ij}(X_{t-}) u_{ij}(t)$. Let $u(t)$ denote the enumeration of $\{u_{ij}(t)\}$ (according to a fixed ordering of (i, j), $i \leqq N$, $J_j \in R_i$) written as a vector. Then for each t, $u(t)$ is a vector of finite dimension equal to $\sum_{i=1}^{N} |R_i|$ where $|R_i|$ is the cardinality of R_i. Also, each element of this vector is either 0 or 1. Let $\|X_t\| = \sum_{i=1}^{N} x_i(t)$. Our objective is to choose a nonanticipative control process $u(t)$ so as to a.s. minimize

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t \|X_s\|\, ds. \tag{5.1}$$

Let e_{ij} denote an N-dimensional vector whose ith coordinate is -1, jth coordinate is 1 and the rest are 0. Let q_i, c_i denote the N-dimensional vectors whose ith coordinate is 1, -1 respectively, the rest being 0. F_i will denote the subset of F such that $[a_1, a_2, \cdots, a_N] \in F_i$ if and only if $a_i \geqq 1$. It is easily verified that X_t, $t \in [0, \infty)$ is a continuous-time controlled Markov chain on state space F (which, on suitable relabelling, corresponds to S) and with the control process $u(t)$. The following correspon-

dences with § 4 hold:

$$r(\cdot)=\sum_{i=1}^{N} f_i(\cdot)+\sum_{i=1}^{M}\lambda_i, \qquad \Lambda(\cdot)=\frac{1}{r(\cdot)}, \tag{5.2}$$

$$k(\cdot)=\|\cdot\|, \tag{5.3}$$

and, for $a \in F_i$, b, $a' \in F$,

$$p(a,b,u(t))=\begin{cases}\dfrac{1}{r(a)}f_i(a)p_{ij}(a)u_{ij}(t) & \text{if } b=a+e_{ij},\\ & J_j\in R_i,\quad j\neq N+1,\\ \dfrac{1}{r(a)}f_i(a)p_{i(N+1)}(a)u_{i(N+1)}(t) & \text{if } b=a+c_i,\\ & J_{N+1}\in R_i,\end{cases}$$

$$p(a',b,u(t))=\begin{cases}\dfrac{1}{r(a')}\lambda_j & \text{if } b=a'+q_j,\\ & j\leqq M,\\ 0 & \text{if not covered by any of the cases}\\ & \text{above and } b\neq a',\\ 1-\sum\limits_{\substack{b'\neq a'\\ b'\in F}} p(a',b',u(t)) & \text{if } b=a'.\end{cases}$$

Remarks. Part of the above observation is based on the following simple fact: If $E_1, E_2, \cdots, E_n$ are independent events, the probability that E_i occurred conditioned on the fact that at least one of $E_1, E_2, \cdots, E_n$ occurred is simply $P(E_i)/(\sum_{j=1}^{n} P(E_j))$.

In physical situations like traffic signals, the controls may be required to satisfy some constraints of one or more of the following types. (Here the subscripts of u refer to a "typical" subscript in the appropriate sense according to the context and not to any particular value.)

(i) $u_{ij}(t)=u_{lm}(t)$ for all $t\in[0,\infty)$.

(ii) At most one of $u_{i_m j_m}(t)$, $m=1,2,\cdots,n$ can be 1 at any t.

(iii) $u_{ij}(t)=1$ (or 0, for that matter) for all t.

Consider the combinations of 0's and 1's that can be assigned to $\{u_{ij}(t)\}$ without violating the above constraints. The family of such combinations is independent of t. We assume that this family is nonempty and enumerate these combinations as $C_1, C_2, \cdots, C_K$. To each combination C_i, assign a $\{0,1\}$-valued process $z_i(t)$, $t\in[0,\infty)$, *with the convention that* $z_i(t)=1$ if and only if the $\{u_{ij}(t)\}$ used correspond to the combination C_i. Let E_{ij} be the subset of $\{1,2,\cdots,K\}$ such that if C_n is the combination used at time t, then $u_{ij}(t)=1$ if and only if $n\in E_{ij}$. Note that this definition does not depend upon t. A little thought shows that exactly one of $z_1(t), z_2(t), \cdots, z_K(t)$ will be 1 at any $t\in[0,\infty)$ and that $u_{ij}(t)=\sum_{m\in E_{ij}} z_m(t)$.

We treat $Z(t)=[z_1(t), z_2(t), \cdots, z_K(t)]$ as the new control process taking values in the finite set $G=$ the set of all K-dimensional vectors with only one element equal to 1 and the rest all 0. The advantage of this new formulation, which is completely equivalent to the original one, is that the problem now does not have any additional constraints on the controls. Write $\tilde{p}(a,b,Z)=p(a,b,u)$ when $a,b\in F$ and $u_{ij}=\sum_{m\in E_{ij}} z_m$ for $u=\{u_{ij}\}$, $Z=[z_1,z_2,\cdots,z_K]$. $\tilde{P}_Z$ will denote the infinite dimensional transition probability matrix whose (i,j)th element is $\tilde{p}(b_i,b_j,Z)$, where $[b_1,b_2,\cdots]$ is some fixed enumeration of F.

In view of the remarks at the end of the previous section, we assume that the control process is constant over transition epochs. We then have the complete set-up of § 4. A1–A3 are easily verified. A4, A5 will be assumed to hold. For each $\xi \in \prod_{i \in S} G_i$, where G_i's are replicas of G, let $\tilde{P}\{\xi\}$ denote the transition probability matrix $\tilde{P}_Z$ when the (new) control process is the SS $\gamma\{\xi\}$. We then have the following result, which is merely a restatement of Theorem 4.1 in the present context.

THEOREM 5.1. *The conclusions of Theorem* 4.1 *hold for this problem with the correspondences given by* (5.2), (5.3) *and with* Z, $\tilde{P}_Z$, $\tilde{P}\{\xi_0\}$ *replacing* u, P_u, $P\{\xi_0\}$ *respectively.*

Note that the termwise minimization in the dynamic programming equations is particularly simple in this case. In each term, the expression to be minimized with respect to Z is linear in the components of Z. Thus the minimizing choice of Z, not necessarily unique, is simply that element of G which has a 1 for the component which has the smallest coefficient in the expression to be minimized.

Remarks. (1) We have allowed transitions from a node to itself. These are dummy transitions, since the state does not change. However, an optimal SSS for a problem without dummy transitions can be constructed from the optimal SSS for a problem with dummy transitions in the following manner, as is easily verified: Let the strategies coincide over the transition epochs that are not initiated by dummy transitions and for those that are, simply retain the control from the previous epoch. The advantage of using dummy transitions is to remove the explicit dependence of $r(\cdot)$ on the control.

(2) Our formulation implies that a customer leaving a node J_i returns to that node if he finds the desired transition blocked. This may be unrealistic in practice, since he may simply wait till the desired transition gets unblocked and then leave immediately. This can be partially remedied by modifying the above formulation as follows: To each node J_i, $1 \leqq i \leqq N$, assign a companion node J_{i_0} with $R_i = R_{i_0}$, $p_{ij}(\cdot) = p_{i_0 j}(\cdot)$, $u_{ij}(\cdot) = u_{i_0 j}(\cdot)$ for each $J_j \in R_i$. (Note that the constraint $u_{ij}(\cdot) = u_{i_0 j}(\cdot)$ fits into the constraint structure introduced earlier.) Change the mechanism of motion to the following: A customer leaving J_i at t either goes to J_j in R_i with probability $p_{ij}(X_{t-})u_{ij}(t)$ or goes to J_{i_0} with probability $1 - \sum_{J_j \in R_i} p_{ij}(X_{t-})u_{ij}(t)$. A customer leaving J_{i_0} goes either to some J_j in R_{i_0} with probability $p_{i_0 j}(X_{t-})u_{i_0 j}(t)$ or returns to J_{i_0} with probability $1 - \sum_{J_j \in R_{i_0}} p_{i_0 j}(X_{t-})u_{i_0 j}(t)$. The rate function $f_{i_0}(\cdot)$ assigned to J_{i_0} can be chosen to have very high values for all values of its arguments whose i_0th component is nonzero, thus making transitions away from J_{i_0} very rapid. Note, however, that we have now doubled the dimension of the state vector, i.e., at each node, we observe both the total number of customers and the number of customers ready to leave, but waiting due to blocked transitions.

Examples. (1) Let each node J_i, $i = 1, 2, \cdots, N$, represent a finite collection of service stations $H_{i1}, H_{i2}, \cdots, H_{iN_i}$, with exponential service rates $h_{i1}, h_{i2}, \cdots, h_{iN_i}$, respectively. A customer entering J_i goes to the H_{ij} with the smallest j which is free. If all H_{ij}'s are occupied, he waits till one is free. The customers are served on a first come—first served basis. It is easily verified that this corresponds to the case

$$f_i(X_t) = \sum_{m=1}^{x_i(t) \wedge N_i} h_{im}.$$

Suppose with each J_i, $1 \leqq i \leqq N$, is associated a number $M_i \in [1, 2, \cdots, \infty]$, such that $M_i = \infty$ if $i \leqq M$ (i.e., J_i is the input node) and no customer can enter J_i at time t if $x_i(t) = M_i$. For each $i \in \{1, 2, \cdots, N\}$, $J_j \in R_i$, let $\{a_{ij}\}$ be fixed positive numbers such that $\sum_{J_j \in R_i} a_{ij} \leqq 1$. Suppose that each customer leaving node J_i goes to $J_j \in R_i$ with probability $a_{ij}u_{ij}(t)$ unless $x_j(t) = M_j$ and returns to J_i otherwise. This situation can be

modelled by letting

$$p_{ij}(X_{t-}) = a_{ij} I_{\{x_j(t-)<M_j\}},$$

with $x_j(t-)$ having the obvious meaning. Now the state space is a subset of F. However, the basic set-up remains unchanged. Thus we can accommodate queueing networks with finite buffers between two nodes within this framework.

(2) Suppose J_1 is the only input node and $f_i(X_t) = u_i I_{\{x_i(t)\geq 1\}}$ for some $u_i > 0$, $i = 1, 2, \cdots, N$. Also suppose that $p_{1j}(\cdot) \equiv 1$ for $j = 2, 3, \cdots, N$, $p_{i(N+1)}(\cdot) \equiv 1$ for $i = 2, \cdots, N$ and $p_{ij}(\cdot) \equiv 0$ otherwise. The control $u_{i(N+1)}(t) \equiv 1$ for all t and $i = 2, 3, \cdots, N$. At most one of the controls $u_{1j}(t)$, $j = 2, 3, \cdots, N$, can be 1 at any $t \in [0, \infty)$. This is recognized as the "routing" problem [3]. Clearly, any optimal SSS will not choose $u_{1j}(t) = 0$, $j = 2, 3, \cdots, N$, for any t. Thus the possibility of return to J_1 on leaving J_1 is a mere technicality. The node J_1 is itself a dummy node, in the sense that we may simply consider a Poisson stream of customers arriving and being routed to $J_2, J_3, \cdots, J_N$.

6. Conclusions. We have established the existence of stable stationary optimal strategies for a class of Markov chains and have given the corresponding dynamic programming equations. As pointed out in the introduction, these problems cannot be easily tackled by conventional approaches.

Many possible extensions of the results presented here suggest themselves. Some of them are obvious, e.g., some simple explicit dependence of $k(\cdot)$ and $r(\cdot)$ on the controls can be allowed. Among the more interesting possibilities are the problem of characterizing the networks that satisfy A5 and the control problem with partial observations.

Acknowledgments. The author is grateful to the Department of Applied Mathematics, Twente University of Technology, the Netherlands, for its financial support and hospitality during the course of this work. He would also like to thank Rien Moraal for helpful discussions.

REFERENCES

[1] P. BILLINGSLEY, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[2] E. B. DYNKIN AND A. A. YUSHKEVICH, *Controlled Markov Processes*, Springer-Verlag, Berlin-Heidelberg, 1979.

[3] A. EPHREMIDES, P. P. VARAIYA AND J. WALRAND, *A simple dynamic routing problem*, IEEE Trans. Automatic Control, AC-25 (1980), pp. 690–693.

[4] A. FEDERGRÜN, A. HORDIJK AND H. C. TIJMS, *A note on simultaneous recurrence conditions on a set of denumerable stochastic matrices*, J. Appl. Prob., 15 (1978), pp. 842–847.

[5] D. FREEDMAN, *Markov Chains*, Holden-Day, San Francisco, 1971.

[6] M. LOEVE, *Probability Theory* II, 4th ed., Springer-Verlag, New York, 1978.

[7] Z. ROSBERG, P. P. VARAIYA AND J. WALRAND, *Optimal control of service in tandem queues*, Memo. No. UCB/ERL M80/42, Electronics Research Lab., Univ. California, Berkeley, Sept. 1980.

[8] S. ROSS, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0001 $01.25/0

FEEDBACK STABILIZATION OF LINEAR DIFFUSION SYSTEMS*

YOSHIYUKI SAKAWA†

Abstract. This paper treats the feedback stabilization of linear diffusion systems by using a finite dimensional feedback dynamic controller. We construct a finite dimensional observer using the output functions from sensors, and the control inputs to the system are given by the feedback of the observer output. Assuming, for some fixed finite number L, that the first L modes are controllable and observable, we prove that it is possible to construct a finite dimensional feedback dynamic controller such that the diffusion system has an arbitrarily large damping constant.

Key words. feedback stabilization, diffusion system, finite dimensional dynamic compensator

1. Introduction. In our previous paper [15], we discussed feedback stabilization of linear diffusion systems by designing actuator influence functions or sensor influence functions properly. In this paper, given arbitrary actuator and sensor influence functions, we construct a finite dimensional feedback dynamic controller using an observer. By using the pole assignment theory for finite dimensional linear systems [16], it is possible to stabilize the distributed systems so that it has an arbitrarily large damping constant.

Balas [3] discussed the same problem under the assumption that no observation "spillover" [2] is present. He also discussed the feedback stabilization problem for dissipative hyperbolic systems [4]. We do not neglect the observation spillover in this paper, and we obtain a sharper estimate for the influence of the control and observation spillovers on the stability of the system. It will be proved that the influence of the spillover on the stability of the system can be made arbitrarily small, if we increase the number of state variables of the dynamic controller.

2. Diffusion systems. Let Ω be a *bounded* domain in a finite-dimensional Euclidean space, and let $L^2(\Omega)$ denote the Hilbert space of all square integrable real-valued functions with the inner product

$$(u_1, u_2) = \int_\Omega u_1(x)u_2(x)\,dx.$$

We consider diffusion processes in Ω described by the linear differential equation

$$\frac{du(t)}{dt} + Au(t) = Bf(t) = \sum_{k=1}^{r} b^k f^k(t), \qquad t > 0, \tag{1}$$

where $u(t) \in L^2(\Omega)$, $b^k \in L^2(\Omega)$, $f^k(t)$ are scalar functions Hölder-continuous on $[0, \infty)$, and

$$B = (b^1, \cdots, b^r), \qquad f(t) = \begin{bmatrix} f^1(t) \\ \vdots \\ f^r(t) \end{bmatrix}.$$

We assume that A is a *selfadjoint* operator with the domain $D(A)$ which is dense in $L^2(\Omega)$, that the resolvent $(A-\lambda)^{-1}$ of A exists and is *compact* for some λ, and that A is bounded from below.

* Received by the editors January 25, 1982, and in final revised form June 14, 1982.

† Department of Control Engineering, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka 560, Japan.

From the assumption we see that A is closed [9, p. 16], that there is a constant γ such that [8, p. 278]

$$(Au, u) \geqq \gamma(u, u), \qquad u \in D(A), \tag{2}$$

and that the resolvent $(A-\lambda)^{-1}$ exists and is compact for any real λ satisfying $\lambda < \gamma$ [8, p. 187].

From the Hilbert–Schmidt theory [11, p. 159] for the compact selfadjoint operators, it is well known that there exist eigenvalues λ_i and corresponding eigenfunctions $\phi_{ij}(x)$ of the operator A satisfying the following conditions [6], [11, p. 167]:

(i) $\gamma \leqq \lambda_1 < \lambda_2 < \cdots < \lambda_i < \cdots$, $\lim \lambda_i = \infty$.

(ii) $A\phi_{ij} = \lambda_i \phi_{ij}$, $j = 1, \cdots, m_i$, $i = 1, 2, \cdots$, where $m_i < \infty$ for each i.

(iii) The set $\{\phi_{ij}(\cdot)\}$ of the eigenfunctions forms a *complete orthonormal system* in $L^2(\Omega)$.

Since $u \in L^2(\Omega)$ has a unique expression

$$u(\cdot) = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} (u, \phi_{ij}) \phi_{ij}(\cdot),$$

$D(A)$ consists of all elements $u \in L^2(\Omega)$ such that

$$\sum_{i=1}^{\infty} \sum_{j=1}^{m_i} \lambda_i^2 (u, \phi_{ij})^2 < \infty. \tag{3}$$

The semigroup e^{-tA} generated by $-A$ is analytic in $t > 0$, and is expressed as [11, p. 309]

$$e^{-tA} u = \sum_{i=1}^{\infty} \sum_{j=1}^{m_i} e^{-\lambda_i t} (u, \phi_{ij}) \phi_{ij}, \qquad t \geqq 0, \tag{4}$$

where $u \in L^2(\Omega)$.

From (4) we see that

$$\| e^{-tA} \| \leqq e^{-\lambda_1 t}, \qquad t \geqq 0. \tag{5}$$

If $\lambda_1 < 0$, the diffusion system

$$\frac{du(t)}{dt} + Au(t) = 0 \tag{6}$$

is clearly unstable. We consider this case, and we synthesize the input functions $f^k(t)$ by using a feedback dynamic controller so that (1) is stabilized.

We assume that there are p sensors, whose outputs are given by

$$y^k(t) = (c^k, u(t)), \qquad k = 1, \cdots, p, \tag{7}$$

where $c^k(x)$ are sensor influence functions in $L^2(\Omega)$. Let us define the output vector function

$$y(t) = \begin{bmatrix} y^1(t) \\ \vdots \\ y^p(t) \end{bmatrix}.$$

Let $\sigma > 0$ be a given damping constant. We take an integer l such that

$$\lambda_{l+1} > \sigma. \tag{8}$$

We take another integer n such that $n \geqq l$, and we define the orthogonal projection operators P_n and Q_n by

$$P_n u = \sum_{i=1}^{n} \sum_{j=1}^{m_i} (u, \phi_{ij})\phi_{ij},$$

$$Q_n u = (I - P_n)u = \sum_{i=n+1}^{\infty} \sum_{j=1}^{m_i} (u, \phi_{ij})\phi_{ij}.$$

Let $u(t)$ be the solution of (1) satisfying the initial condition

$$\lim_{t \to +0} u(t) = u_0 \in L^2(\Omega).$$

Since $u(t) \in D(A)$ for $t > 0$, from (1) we obtain

$$P_n \dot{u}(t) + A P_n u(t) = P_n B f(t), \tag{9}$$

$$Q_n \dot{u}(t) + A Q_n u(t) = Q_n B f(t). \tag{10}$$

The solution of (9) with the initial condition $P_n u(0) = P_n u_0$ can be expressed as

$$P_n u(t) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} u_{ij}(t)\phi_{ij}(\cdot), \tag{11}$$

where $u_{ij}(t)$ is the unique solution of

$$\dot{u}_{ij}(t) + \lambda_i u_{ij}(t) = b_{ij} f(t), \qquad j = 1, \cdots, m_i, \quad i = 1, \cdots, n, \tag{12}$$

satisfying the initial condition

$$u_{ij}(0) = (u_0, \phi_{ij}). \tag{13}$$

In (12), b_{ij} is a row vector defined by

$$b_{ij} = (b_{ij}^1, \cdots, b_{ij}^r),$$

where $b_{ij}^k = (b^k, \phi_{ij})$.

Now let us define the following vector and matrix:

$$u_i(t) = \begin{bmatrix} u_{i1}(t) \\ \vdots \\ u_{im_i}(t) \end{bmatrix}, \qquad \tilde{B}_i = \begin{bmatrix} b_{i1} \\ \vdots \\ b_{im_i} \end{bmatrix} = \begin{bmatrix} b_{i1}^1 & \cdots & b_{i1}^r \\ \vdots & & \vdots \\ b_{im_i}^1 & \cdots & b_{im_i}^r \end{bmatrix}. \tag{14}$$

Then (12) can be written as

$$\dot{u}_i(t) + \lambda_i u_i(t) = \tilde{B}_i f(t), \qquad i = 1, \cdots, n. \tag{15}$$

Furthermore, let

$$L = m_1 + \cdots + m_l, \qquad N = m_1 + \cdots + m_n.$$

Since $l \leqq n$, $L \leqq N$. Let us define an L-dimensional vector $x_1(t)$, $L \times r$ matrix B_1, and $L \times L$ diagonal matrix A_1 by

$$x_1(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_l(t) \end{bmatrix}, \qquad B_1 = \begin{bmatrix} \tilde{B}_1 \\ \vdots \\ \tilde{B}_l \end{bmatrix},$$

$$A_1 = \operatorname{diag}(-\lambda_1 I_{m_1}, \cdots, -\lambda_l I_{m_l}), \tag{16}$$

where I_m denotes an $m \times m$ unit matrix. Then from (15) we obtain

$$\dot{x}_1(t) = A_1 x_1(t) + B_1 f(t). \tag{17}$$

Similarly, let us define an $(N-L)$-dimensional vector $x_2(t)$, $(N-L)\times r$ matrix B_2, and $(N-L)\times(N-L)$ diagonal matrix A_2 by

$$\begin{gathered} x_2(t) = \begin{bmatrix} u_{l+1}(t) \\ \vdots \\ u_n(t) \end{bmatrix}, \qquad B_2 = \begin{bmatrix} \tilde{B}_{l+1} \\ \vdots \\ \tilde{B}_n \end{bmatrix}, \\ A_2 = \operatorname{diag}(-\lambda_{l+1} I_{m_{l+1}}, \cdots, -\lambda_n I_{m_n}). \end{gathered} \tag{18}$$

Then we obtain

$$\dot{x}_2(t) = A_2 x_2(t) + B_2 f(t). \tag{19}$$

We see that (17) is controllable if and only if [7], [13]

$$\operatorname{rank} \tilde{B}_i = m_i, \qquad i = 1, \cdots, l. \tag{20}$$

In order for the relation (20) to hold, the number r of the control inputs should satisfy

$$r \geqq \max\{m_1, \cdots, m_l\}. \tag{21}$$

Since

$$u(t) = P_n u(t) + Q_n u(t),$$

the output functions (7) are expressed as

$$y^k(t) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} c_{ij}^k u_{ij}(t) + (Q_n c^k, Q_n u(t)), \qquad k = 1, \cdots, p, \tag{22}$$

where $c_{ij}^k = (c^k, \phi_{ij})$. By defining the matrices

$$\begin{gathered} \tilde{C}_i = \begin{bmatrix} c_{i1}^1 & \cdots & c_{im_i}^1 \\ \vdots & & \vdots \\ c_{i1}^p & \cdots & c_{im_i}^p \end{bmatrix}, \\ C_1 = [\tilde{C}_1, \cdots, \tilde{C}_l], \qquad C_2 = [\tilde{C}_{l+1}, \cdots, \tilde{C}_n], \end{gathered} \tag{23}$$

we can express the output vector function as

$$y(t) = C_1 x_1(t) + C_2 x_2(t) + S_n Q_n u(t), \tag{24}$$

where $S_n Q_n u(t)$ is called the *observation spillover* [2], and the operator S_n mapping $Q_n L^2(\Omega)$ into R^p is defined by

$$S_n u = \begin{bmatrix} (Q_n c^1, u) \\ \vdots \\ (Q_n c^p, u) \end{bmatrix}, \qquad u \in Q_n L^2(\Omega). \tag{25}$$

The system (A_1, C_1) is observable if and only if [7], [14]

$$\operatorname{rank} \tilde{C}_i = m_i, \qquad i = 1, \cdots, l. \tag{26}$$

In order for the rank condition (26) to hold, the number p of sensors should be such that

$$p \geqq \max\{m_1, \cdots, m_l\}. \tag{27}$$

3. Feedback control using observers. First, we construct two kinds of finite dimensional observer defined by

$$\dot{z}_1(t) = (A_1 - G_1C_1)z_1(t) + G_1[y(t) - C_2z_2(t)] + B_1f(t), \tag{28}$$

$$\dot{z}_2 = A_2z_2 + B_2f(t), \tag{29}$$

where $z_1(t)$ is an L-dimensional vector and $z_2(t)$ is an $(N-L)$-dimensional vector which estimate $x_1(t)$ and $x_2(t)$, respectively, and G_1 is an $L \times p$ matrix to be determined.

It is clear from (19) and (29) that

$$x_2(t) - z_2(t) = e^{A_2t}(x_{20} - z_{20}), \tag{30}$$

where $x_{20} = x_2(0)$, and $z_{20} = z_2(0)$. Let us define a $2L$-dimensional vector $q_1(t)$ by

$$q_1(t) = \begin{bmatrix} x_1(t) \\ z_1(t) \end{bmatrix}. \tag{31}$$

Let the control input vector function $f(t)$ be given by

$$f(t) = F_1z_1(t), \tag{32}$$

where F_1 is an $r \times L$ matrix to be determined. Substituting (32) into (10), (17), and (28), and using (24) and (30) gives

$$\frac{d}{dt}\begin{bmatrix} q_1(t) \\ Q_nu(t) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} q_1(t) \\ Q_nu(t) \end{bmatrix} + \begin{bmatrix} \psi(t) \\ 0 \end{bmatrix}, \tag{33}$$

where

$$A_{11} = \begin{bmatrix} A_1 & B_1F_1 \\ G_1C_1 & A_1 - G_1C_1 + B_1F_1 \end{bmatrix}, \qquad A_{12} = \begin{bmatrix} 0 \\ G_1S_n \end{bmatrix},$$
$$A_{21} = [0, Q_nBF_1], \qquad A_{22} = -AQ_n, \tag{34}$$

$$\psi(t) = \begin{bmatrix} 0 \\ G_1C_2\, e^{A_2t}(x_{20} - z_{20}) \end{bmatrix}. \tag{35}$$

We see that

$$\begin{bmatrix} I_L & 0 \\ -I_L & I_L \end{bmatrix}\begin{bmatrix} A_1 & B_1F_1 \\ G_1C_1 & A_1 - G_1C_1 + B_1F_1 \end{bmatrix}\begin{bmatrix} I_L & 0 \\ -I_L & I_L \end{bmatrix}^{-1}$$
$$= \begin{bmatrix} A_1 + B_1F_1 & B_1F_1 \\ 0 & A_1 - G_1C_1 \end{bmatrix}. \tag{36}$$

Suppose the rank conditions (20) and (26) hold. Then the linear system (A_1, B_1, C_1) is controllable and observable. Consequently, there exist matrices F_1 and G_1 such that all the eigenvalues of the matrices $A_1 + B_1F_1$ and $A_1 - G_1C_1$ take any preassigned values $\{-\nu_1, -\nu_2, \cdots, -\nu_{2L}\}$ [10], [16]. Here, the real numbers $\nu_i > 0$ are such that

$$\sigma < \lambda_{l+1} < \nu_1 < \nu_2 < \cdots < \nu_{2L},$$
$$\nu_1 < \lambda_{n+1}. \tag{37}$$

Let us construct the matrices F_1 and G_1 as stated above. In view of (34) and (36), we see that the matrix A_{11} is similar to the diagonal matrix

$$\operatorname{diag}(-\nu_1, -\nu_2, \cdots, -\nu_{2L}),$$

and the matrix $e^{A_{11}t}$ is also similar to the diagonal matrix

$$\text{diag}\,(e^{-\nu_1 t}, e^{-\nu_2 t}, \cdots, e^{-\nu_{2L} t}).$$

In other words, there is a nonsingular matrix T_1 such that [1], [5]

$$T_1 e^{A_{11}t} T_1^{-1} = \text{diag}\,(e^{-\nu_1 t}, \cdots, e^{-\nu_{2L} t}). \tag{38}$$

From (38) we obtain

$$\|e^{A_{11}t}\| \leqq M_1 e^{-\nu_1 t}, \qquad t \geqq 0, \tag{39}$$

where M_1 is the so-called condition number of a nonsingular matrix T_1 defined by

$$M_1 = \|T_1\| \|T_1^{-1}\| \geqq 1. \tag{40}$$

Since

$$e^{A_{22}t} Q_n u_0 = \sum_{i=n+1}^{\infty} \sum_{j=1}^{m_i} e^{-\lambda_i t} (Q_n u_0, \phi_{ij}) \phi_{ij}, \tag{41}$$

it is clear that

$$\|e^{A_{22}t}\| \leqq e^{-\lambda_{n+1} t}, \qquad t \geqq 0. \tag{42}$$

Let us define the operators

$$\tilde{A} = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \qquad \tilde{B} = \begin{bmatrix} 0 & A_{12} \\ A_{21} & 0 \end{bmatrix}, \tag{43}$$

where $\tilde{A}$ is unbounded, whereas $\tilde{B}$ is bounded. Then

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \tilde{A} + \tilde{B}. \tag{44}$$

Let us introduce infinite dimensional vectors

$$w(t) = \begin{bmatrix} q_1(t) \\ Q_n u(t) \end{bmatrix}, \qquad \tilde{\psi}(t) = \begin{bmatrix} \psi(t) \\ 0 \end{bmatrix}, \tag{45}$$

with the norm

$$\|w(t)\| = [\|q_1(t)\|^2 + \|Q_n u(t)\|^2]^{1/2}.$$

Then (33) can be written as

$$\dot{w}(t) = (\tilde{A} + \tilde{B}) w(t) + \tilde{\psi}(t). \tag{46}$$

It is clear that

$$e^{\tilde{A}t} = \begin{bmatrix} e^{A_{11}t} & 0 \\ 0 & e^{A_{22}t} \end{bmatrix},$$

and that

$$\|e^{\tilde{A}t}\| \leqq \max\,(\|e^{A_{11}t}\|, \|e^{A_{22}t}\|). \tag{47}$$

Since $0 < \nu_1 < \lambda_{n+1}$, and $M_1 \geqq 1$, it follows from (39) and (42) that

$$\|e^{\tilde{A}t}\| \leqq M_1 e^{-\nu_1 t}, \qquad t \geqq 0. \tag{48}$$

From (43) we see that

$$\|\tilde{B}\| \leqq \max\,(\|A_{12}\|, \|A_{21}\|) \leqq \max\,(\|Q_n B\| \|F_1\|, \|G_1\| \|S_n\|), \tag{49}$$

where

$$\|Q_n B\| = (\|Q_n b^1\|^2 + \cdots + \|Q_n b^r\|^2)^{1/2},$$
$$\|S_n\| \leqq (\|Q_n c^1\|^2 + \cdots + \|Q_n c^p\|^2)^{1/2}. \tag{50}$$

The second relation of (50) is derived from (25).

Now, applying the perturbation theory of semigroups [8, p. 495], [12, p. 80], we obtain

$$\|e^{(\tilde{A}+\tilde{B})t}\| \leqq M_1 e^{-(\nu_1 - M_1\|\tilde{B}\|)t}, \qquad t \geqq 0. \tag{51}$$

Let

$$\bar{\nu}_1 = \nu_1 - M_1\|\tilde{B}\|.$$

Since $b^i \in L^2(\Omega)$ $(i = 1, \cdots, r)$, $c^i \in L^2(\Omega)$ $(i = 1, \cdots, p)$, and M_1 is independent of n, in view of (49) and (50), for any small number $\varepsilon > 0$ there is an integer $n (\geqq l)$ such that

$$M_1\|\tilde{B}\| \leqq \varepsilon. \tag{52}$$

Since $\lambda_{l+1} < \nu_1$ from (37), proper choice of n gives the relation

$$\sigma < \lambda_{l+1} < \bar{\nu}_1 < \nu_1. \tag{53}$$

The solution of (46) is clearly given as

$$w(t) = e^{(\tilde{A}+\tilde{B})t} w_0 + \int_0^t e^{(\tilde{A}+\tilde{B})(t-s)} \tilde{\psi}(s)\, ds. \tag{54}$$

By using (51), $\|w(t)\|$ can be estimated as

$$\|w(t)\| \leqq M_1 e^{-\bar{\nu}_1 t}\left[\|w_0\| + \int_0^t e^{\bar{\nu}_1 s}\|\tilde{\psi}(s)\|\, ds\right]. \tag{55}$$

From (18) we see that

$$\|e^{A_2 t}\| = e^{-\lambda_{l+1} t}, \qquad t \geqq 0. \tag{56}$$

Using (35) and (56), we obtain

$$\|\tilde{\psi}(t)\| \leqq \|G_1\|\,\|C_2\|\, e^{-\lambda_{l+1} t}\|x_{20} - z_{20}\|. \tag{57}$$

Substituting (57) into (55) yields

$$\|w(t)\| \leqq M_2 e^{-\lambda_{l+1} t}, \qquad t \geqq 0, \tag{58}$$

where

$$M_2 = M_1[\|w_0\| + \|G_1\|\,\|C_2\|(\bar{\nu}_1 - \lambda_{l+1})^{-1}\|x_{20} - z_{20}\|]. \tag{59}$$

Define a $2(N-L)$-dimensional vector $q_2(t)$ by

$$q_2(t) = \begin{bmatrix} x_2(t) \\ z_2(t) \end{bmatrix}. \tag{60}$$

From (19) and (29) we obtain

$$\dot{q}_2(t) = \hat{A} q_2(t) + \hat{B} z_1(t), \tag{61}$$

where

$$\hat{A} = \begin{bmatrix} A_2 & 0 \\ 0 & A_2 \end{bmatrix}, \qquad \hat{B} = \begin{bmatrix} B_2 F_1 \\ B_2 F_1 \end{bmatrix}. \tag{62}$$

Integrating (61) gives

$$q_2(t) = e^{\hat{A}t} q_{20} + \int_0^t e^{\hat{A}(t-s)} \hat{B} z_1(s)\, ds, \tag{63}$$

where $q_{20} = q_2(0)$. Using the estimates

$$\begin{aligned} &\|e^{\hat{A}t}\| \leqq \|e^{A_2 t}\| = e^{-\lambda_{l+1} t} \leqq e^{-\sigma t}, \qquad t \geqq 0, \\ &\|\hat{B}\| \leqq \sqrt{2}\|F_1\|\,\|B_2\|, \\ &\|z_1(t)\| \leqq \|w(t)\| \leqq M_2\, e^{-\lambda_{l+1} t}, \qquad t \geqq 0, \end{aligned} \tag{64}$$

we obtain the following relation:

$$\|q_2(t)\| \leqq e^{-\sigma t}[\|q_{20}\| + \sqrt{2}\, M_2 \|F_1\|\,\|B_2\| (\lambda_{l+1} - \sigma)^{-1}]. \tag{65}$$

Putting (58) and (65) together, we finally obtain

$$\|w(t)\| \leqq M_1[\|w_0\| + \sqrt{2}\|G_1\|\,\|C_2\| (\bar{\nu}_1 - \lambda_{l+1})^{-1} \|q_{20}\|]\, e^{-\sigma t}, \tag{66}$$

$$\begin{aligned} \|q_2(t)\| \leqq M_1[&\sqrt{2}\|F_1\|\,\|B_2\| (\lambda_{l+1} - \sigma)^{-1} \|w_0\| \\ &+ \{1 + 2\|F_1\|\,\|G_1\|\,\|B_2\|\,\|C_2\| (\lambda_{l+1} - \sigma)^{-1} (\bar{\nu}_1 - \lambda_{l+1})^{-1}\} \|q_{20}\|]\, e^{-\sigma t}, \end{aligned} \tag{67}$$

for $t \geqq 0$. Define an infinite dimensional vector $\tilde{w}(t)$ by

$$\tilde{w}(t) = \begin{bmatrix} q_1(t) \\ q_2(t) \\ Q_n u(t) \end{bmatrix} \in R^{2N} \times Q_n L^2(\Omega).$$

It is obvious that $\tilde{w}(t)$ represents the state of the diffusion system as well as the state of the dynamic controller. From (66) and (67) we see that

$$\|\tilde{w}(t)\| \leqq K e^{-\sigma t} \|\tilde{w}(0)\|, \qquad t \geqq 0, \tag{68}$$

where K is a constant dependent on l, n, etc.

Thus we can summarize what we have discussed so far as follows:

THEOREM. *Given an arbitrary damping constant $\sigma > 0$, suppose that the rank conditions* (20) *and* (26) *hold, where l is an integer satisfying* (8). *Then a finite dimensional feedback dynamic controller, described by* (28), (29), *and* (32), *can be constructed such that the state $\tilde{w}(t)$ of the overall system satisfies* (68), *where K is a constant dependent on l, n, and so on.*

Remark 1. The bounded operator $\tilde{B}$ defined by (43) results from the control and observation spillovers [2]. If

$$\begin{aligned} &b^k(\cdot) \in P_n L^2(\Omega), \qquad k = 1, \cdots, r, \\ &c^k(\cdot) \in P_n L^2(\Omega), \qquad k = 1, \cdots, p, \end{aligned}$$

for some integer n, then $\tilde{B} = 0$.

Remark 2. If $l = n$, (52) does not hold, because in this case the constant M_1 depends on $l = n$ and $M_1 \to \infty$ $(l \to \infty)$, in general. Thus, boundedness of $M_1\|\tilde{B}\|$ with respect to n is not clear. The key point of this paper lies in the introduction of two different integers l and n.

Remark 3. In this paper, an identity observer has been used for estimating the state of the first L modes of the diffusion system. It is also possible to construct a feedback dynamic controller by use of a reduced order observer [10].

Remark 4. Curtain [17] discusses, under some conditions, the case where operator B in (1) and the observation operator C defined by $y(t) = Cu(t)$ are not bounded.

Remark 5. Mitkowski [18] considered the stabilization of linear distributed systems by using an infinite-dimensional observer. The idea in this paper is partially due to Mitkowski.

4. An example of a partial differential equation. In this section, we show an example of a partial differential equation as well as a boundary condition which can be described in abstract form as in (1).

Let us consider the partial differential equation

$$\frac{\partial u(t,x)}{\partial t} - (\Delta - c(x))u(t,x) = \sum_{k=1}^{r} b^k(x) f^k(t), \tag{69}$$

where $x \in \Omega$, Δ denotes the Laplacian, and $c(x)$ is a bounded measurable function. The boundary condition is assumed to be either of the Dirichlet type

$$u(t,x) = 0, \qquad t > 0, \quad x \in \Gamma, \tag{70}$$

or of the third kind

$$\frac{du(t,x)}{dn} + \sigma(x)u(t,x) = 0, \qquad t > 0, \quad x \in \Gamma, \tag{71}$$

where Γ is a sufficiently smooth boundary of Ω, d/dn is the derivative in the direction of the inner normal, and $\sigma(x)$ is a sufficiently smooth function on Γ. In view of (69), let us define the operator A by

$$Au = (-\Delta + c(x))u(\cdot). \tag{72}$$

It is proved in [11] that the operator A defined on the domain

$$D(A) = \left\{ u \in H^2(\Omega) : \frac{du}{dn} + \sigma u = 0 \text{ (or } u = 0) \text{ on } \Gamma \right\} \tag{73}$$

is selfadjoint, where $H^2(\Omega)$ is the Sobolev space of order 2, that $D(A)$ is dense in $L^2(\Omega)$, that the resolvent $(A - \lambda)^{-1}$ exists and is compact for any real λ satisfying $\lambda < \inf_{x \in \Omega} c(x)$, and that A is bounded from below. Therefore, the diffusion system (69) with the boundary condition (70) or (71) can be expressed as in (1).

Acknowledgments. The author wishes to thank Dr. N. Fujii and Dr. T. Nambu for their valuable discussions.

REFERENCES

[1] F. AYRES, JR., *Matrices*, Schaum's Outline Series, McGraw-Hill, New York, 1962.
[2] M. J. BALAS, *Modal control of certain flexible dynamic systems*, this Journal, 16 (1978), pp. 450–462.
[3] ———, *Feedback control of linear diffusion processes*, Internat. J. Control, 29 (1979), pp. 523–533.
[4] ———, *Feedback control of dissipative hyperbolic distributed parameter systems with finite dimensional controllers*, private communication.
[5] F. R. GANTMACHER, *The Theory of Matrices*, Vol. I, Chelsea, New York, 1960.
[6] S. ITO, *Partial Differential Equations*, Baifukan, Tokyo, 1966. (In Japanese.)
[7] R. E. KALMAN, Y. C. HO AND K. S. NARENDRA, *Controllability of linear dynamical systems*, Contributions to Differential Equations, 1 (1963), pp. 189–213.
[8] T. KATO, *Perturbation Theory for Linear Operators*, Springer, Berlin, 1966.
[9] S. G. KREIN, *Linear Differential Equations in Banach Space*, American Mathematical Society, Providence, RI, 1971.

[10] D. G. LUENBERGER, *An introduction to observers*, IEEE Trans. Automat. Control, AC-16 (1971), pp. 596–602.
[11] S. MIZOHATA, *The Theory of Partial Differential Equations*, Cambridge Univ. Press, Cambridge, 1973.
[12] A. PAZY, *Semi-groups of linear operators and applications to partial differential equations*, Lecture Note #10, Univ. of Maryland, College Park, 1974.
[13] Y. SAKAWA, *Controllability for partial differential equations of parabolic type*, SIAM J. Control, 12 (1974), pp. 389–400.
[14] ———, *Observability and related problems for partial differential equations of parabolic type*, SIAM J. Control, 13 (1975), pp. 14–27.
[15] Y. SAKAWA AND T. MATSUSHITA, *Feedback stabilization of a class of distributed systems and construction of a state estimator*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 748–753.
[16] W. H. WONHAM, *On pole assignment in multiple-input controllable linear systems*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 660–665.
[17] R. CURTAIN, *Finite-dimensional compensator design for parabolic distributed systems with point sensors and boundary input*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 98–104.
[18] W. MITKOWSKI, *Stabilization of linear distributed systems*, private communication.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0002 $01.25/0

STRUCTURAL PROPERTIES OF THE LINEAR-QUADRATIC STOCHASTIC CONTROL PROBLEM*

KENKO UCHIDA†

Abstract. This paper treats the problem of optimally controlling a class of linear stochastic systems with noisy observations and bounded controls modeled via the Girsanov transformation. The cost functional is quadratic, and the initial state is non-Gaussian and bounded. It is shown that, under a certain inequality condition for the system matrices and the weighting matrices of the cost functional, there exists a unique optimal control which is linear in the state estimate but which is nonlinear and in particular non-Lipschitzian in the observation. It is also shown that the certainty equivalence property holds.

Key words. stochastic control, linear-quadratic problem, non-Gaussian initial state, certainty equivalence

1. Introduction. The linear-quadratic-Gaussian control problem, which takes a particular position of practical importance in stochastic control theory, has been almost exhaustively studied in various aspects (see, e.g., [1], [8], [12]). One of the features of this problem is the certainty equivalence property: the optimal control is given by the tandem connection of the optimal regulator and the optimal state estimator. Attempts to establish this property were reported in [10] for non-Gaussian initial state and noise distributions as well as for nonlinear measurements. In such extensions, however, the problem is reduced to whether for the certainty equivalence control the system equation has a strong solution. A sufficient condition for the existence of the strong solution is that the state estimate is Lipschitzian in the observations but this seems to be a severe restriction except in the Gaussian case. Another sufficient condition is to assume the existence of nonzero time delays in the observations and/or controls [9], [11].

Our concern in the present paper also lies in establishing the certainty equivalence property in a non-Gaussian case. However, in order to get rid of the technical restrictions stated above, we formulate the problem by making use of the notion of weak solution for stochastic differential equations as generated by the Girsanov transformation as in [2]. In the weak sense formulation it has been shown that the linear–Gaussian control problem has the properties similar to ones obtained for the strong sense formulation: the state estimator is generated by a filter of the Kalman–Bucy type [3] and the separation principle can be established under certain conditions [4]. A feature of our control problem compared with these previous ones is the non-Gaussian character of the initial state, which guarantees generally a nonlinear structure for the state estimator and will thus make our problem difficult to solve. It should be noted here that by the weak sense formulation we can eliminate the control dependence of the observation but we are led to the control dependence of the basic probability measure.

In this paper, we consider a linear stochastic control system with noise free state dynamics and linear observations in additive Gaussian white noise. More precisely, on some interval $[0, T]$,

$$dx(t) = A(t)x(t)\,dt + B(t)u(t)\,dt, \qquad x(0) = z, \tag{1}$$

$$dy(t) = F(t)x(t)\,dt + dw(t), \qquad y(0) = 0, \quad w(0) = 0, \tag{2}$$

* Received by the editors March 3, 1981, and in revised form September 10, 1982.

† Department of Electrical Engineering, Waseda University, Tokyo 160, Japan. Now at Lehrstuhl und Laboratorium für Steuerungs- und Regelungstechnik, Technische Universität München, München, West Germany.

where $\{w(t), 0 \leqq t \leqq T\}$ is an m-dimensional Wiener process independent of the initial state z which takes value in a bounded set $Z = \{z \in R^n : \|z\| \leqq N\}$. Here N is a positive constant and $\|\cdot\|$ is the Euclidean norm. The control process $\{u(t), 0 \leqq t \leqq T\}$ is allowed to depend on the past observation $\{y(s), 0 \leqq s \leqq t\}$ and takes value in a bounded set $U = \{u \in R^r : \|u\| \leqq M\}$, where M is a positive constant. The problem is one of selecting an admissible control which minimizes a quadratic cost functional of the form

$$J(u) = E\left\{x'(T)S_T x(T) + \int_0^T (x'(t)Q(t)x(t) + u'(t)R(t)u(t))\, dt\right\}. \tag{3}$$

It is shown that, if a certain inequality condition for $A(t)$, $B(t)$, S_T, $Q(t)$ and $R(t)$ is satisfied, there exists a unique optimal control which is linear in the state estimate but which is nonlinear and in particular non-Lipschitzian in the observations. It is also shown that the optimal control has the certainty equivalence property.

2. Formulation of the problem. The stochastic differential equations (1) and (2) are solved in the weak sense by making use of the Girsanov transformation: Let C be the space of all continuous functions from $[0, T]$ to R^m and denote by $\mathcal{B}_t(C)$ the σ-field of C generated by the cylinder sets $\{f \in C : f(s) \in B\}$, where $0 \leqq s \leqq t$ and B is a Borel subset of R^m and $\mathcal{B}(C) = \bigvee_{0 \leqq t \leqq T} \mathcal{B}_t(C)$. Let $\mathcal{B}(Z)$ denote the σ-field of the Borel subsets of Z. Introduce the basic underlying probability space $(\Omega, \mathcal{B}, P^*)$ as being the product space $(Z \times C, \mathcal{B}(Z) \otimes \mathcal{B}(C), P_Z \otimes P_W)$, where P_Z is the distribution on Z and P_W is the Wiener measure on C. It should be recalled that once the measure P^* is defined as $P_Z \otimes P_W$, the canonical or evaluation process $\{\pi_t, 0 \leqq t \leqq T\}$, defined by $\pi_t(z, y) = y(t)$ for all pairs (z, y) in Ω, is a standard Wiener process independent of the random variable z under measure P^*. An admissible control law u is any jointly measurable function $u : [0, T] \times C \to U$ such that $u(t, \cdot)$ is $\mathcal{B}_t(C)$-measurable for each $t \in [0, T]$. For each admissible control law u and each $(z, y) \in (Z \times C)$, define the processes $\{x_u(t), 0 \leqq t \leqq T\}$ and $\{w_u(t), 0 \leqq t \leqq T\}$ by setting

$$x_u(t) = \Phi(t, 0)z + \int_0^t \Phi(t, s)B(s)u(s, y)\, ds, \tag{4}$$

$$w_u(t) = y(t) - \int_0^t F(s)x_u(s)\, ds, \tag{5}$$

where $\Phi(t, s)$ is the fundamental solution matrix associated with $A(t)$. The components of the matrices $A(t)$, $B(t)$ and $F(t)$ are real-valued continuous functions. Define the measure P^u on $(Z \times C, \mathcal{B}(Z) \otimes \mathcal{B}(C))$, absolutely continuous with respect to the product measure $P^* = P_Z \otimes P_W$, by its Radon–Nikodym derivative

$$\frac{dP^u}{dP^*} = \exp\left\{\int_0^T (F(t)x_u(t))'\, dy(t) - \frac{1}{2}\int_0^T \|F(t)x_u(t)\|^2\, dt\right\}, \tag{6}$$

for each admissible control law u. $(')$ denotes the transpose of vectors and matrices. Then, by noting that $x_u(t)$ is bounded since $\|z\| \leqq N$ and $\|u(t, y)\| \leqq M$, we can show that $E^*\{dP^u/dP^*\} = 1$, where E^* denotes the expectation with respect to P^*, that is, the measure P^u is thus a probability measure on $(Z \times C, \mathcal{B}(Z) \otimes \mathcal{B}(C))$ and Girsanov's theorem applies [5]:

(i) Under P^u, the process $\{w_u(t), 0 \leqq t \leqq T\}$ is a Wiener process independent of the random variable z.

(ii) The random variable z has the same distribution P_Z under both measures P^* and P^u.

In view of the above facts and of the relations (4) and (5), the processes $\{x_u(t), 0 \leqq t \leqq T\}$ and $\{y(t), 0 \leqq t \leqq T\}$ solve the system of stochastic differential equations

$$dx_u(t) = A(t)x_u(t)\,dt + B(t)u(t, y)\,dt, \qquad x_u(0) = z, \tag{7}$$

$$dy(t) = F(t)x_u(t)\,dt + dw_u(t), \qquad y(0) = 0, \quad w_u(0) = 0, \tag{8}$$

and the probability measure P^u is thus the unique weak solution of (1) and (2) for any admissible control law u. The corresponding cost functional is then taken to be

$$J(u) = E^u\left\{x_u'(T)S_T x_u(T) + \int_0^T (x_u'(t)Q(t)x_u(t) + u'(t)R(t)u(t))\,dt\right\}, \tag{9}$$

where $u(t) = u(t, y)$ and E^u denotes expectation with respect to P^u. The matrix S_T is positive semidefinite symmetric. $Q(t)$ and $R(t)$ are positive semidefinite symmetric and positive definite symmetric matrices, respectively, whose components are real-valued continuous functions. Note again that in our control problem the set of the initial state Z and the control set U are both bounded as $Z = \{z \in R^n : \|z\| \leqq N\}$ and $U = \{u \in R^r : \|u\| \leqq M\}$.

Denote by $\mathcal{Y}_t$ the σ-field of the probability space $Z \times C$ generated by the observation process $\{y(s), 0 \leqq s \leqq t\}$ up to time t: obviously $\mathcal{Y}_t = \{\phi, Z\} \otimes \mathcal{B}_t(C)$ and the $\mathcal{B}_t(C)$-adaptability of the admissible control laws can be seen to be equivalent to the $\mathcal{Y}_t$-adaptability of the corresponding control processes. It is remarkable that in our formulation the σ-fields $\{\mathcal{Y}_t, 0 \leqq t \leqq T\}$ generated by the observation are defined independently of the choice of admissible control laws but, as we see from (6), the basic probability measure P^u depends on the choice of admissible control laws.

3. Structures of the optimal control. In order to solve the stochastic optimal control problem formulated in the previous section, introduce a quadratic functional

$$V(t, x_u(t)) = x_u'(t)S(t)x_u(t),$$

where $\{x_u(t), 0 \leqq t \leqq T\}$ is the state process generated through (7) by an admissible control $u(t)$ and $S(t)$ is the unique solution of the matrix Riccati differential equation [13]

$$\begin{aligned} &\frac{dS(t)}{dt} + A'(t)S(t) + S(t)A(t) + Q(t) - S(t)B(t)R^{-1}(t)B'(t)S(t) = 0, \\ &S(T) = S_T. \end{aligned} \tag{10}$$

Differentiating both sides of $V(t, x_u(t)) = x_u'(t)S(t)x_u(t)$ with respect to t yields

$$\begin{aligned} &V(T, x_u(T)) - V(0, x_u(0)) \\ &\quad = \int_0^T \left(\frac{dx_u'(t)}{dt}S(t)x_u(t) + x_u'(t)\frac{dS(t)}{dt}x_u(t) + x_u'(t)S(t)\frac{dx_u(t)}{dt}\right)dt. \end{aligned} \tag{11}$$

Substituting (7) and (10) into (11), rearranging terms and taking the expectation, we obtain

$$\begin{aligned} J(u) &= E^u\left\{x_u'(T)S_T x_u(T) + \int_0^T (x_u'(t)Q(t)x_u(t) + u'(t)R(t)u(t))\,dt\right\} \\ &= E^u\{z'S(0)z\} \\ &\quad + \int_0^T \operatorname{tr}\,[K'(t)R(t)K(t)E^u\{\tilde{x}_u(t)\tilde{x}_u'(t)\}]\,dt \end{aligned}$$

$$+E^u\left\{\int_0^T (u(t)+K(t)\hat{x}_u(t))'R(t)(u(t)+K(t)\hat{x}_u(t))\,dt\right\},$$

where $K(t)=R^{-1}(t)B'(t)S(t)$, $\tilde{x}_u(t)=x_u(t)-\hat{x}_u(t)$ and $\hat{x}_u(t)=E^u\{x_u(t)|\mathcal{Y}_t\}$, and we used the $\mathcal{Y}_t$-measurability of $u(t)$ for each t. It is noted here that the third term of the above expression is nonnegative since $R(t)$ is positive definite. Therefore, the cost achieves its minimum if and only if

$$u(t)=-K(t)\hat{x}_u(t), \tag{12}$$

under the assumptions that

$$\|K(t)\hat{x}_u(t)\|\leqq M \tag{13}$$

and that the first and second terms are independent of the choice of admissible control laws.

The independence of the first term is shown immediately from the property (ii) of P^u as

$$E^u\{z'S(0)z\}=\int_Z \bar{z}'S(0)\bar{z}P_Z(d\bar{z}).$$

We can also establish the independence of the second term.

LEMMA. *For all t in $[0, T]$, the covariance matrix $E^u\{\tilde{x}_u(t)\tilde{x}_u'(t)\}$ is independent of the admissible control law u used to generate it.*

Proof. Using the linearity of the state equation (7) and the $\mathcal{Y}_t$-measurability of $u(t)$, we have

$$\tilde{x}_u(t)=x_u(t)-\hat{x}_u(t)=\Phi(t,0)(z-E^u\{z|\mathcal{Y}_t\}).$$

By a Bayes formula [6], $E^u\{z|\mathcal{Y}_t\}$ is expressed as

$$E^u\{z|\mathcal{Y}_t\}=\frac{\int_Z \bar{z}\exp\{h(t,x_u(\cdot,\bar{z},y),y)\}P_Z(d\bar{z})}{\int_Z \exp\{h(t,x_u(\cdot,\bar{z},y),y)\}P_Z(d\bar{z})}, \tag{14}$$

where

$$h(t,x_u(\cdot,z,y),y)=\int_0^t (F(s)x_u(s,z,y))'\,dy(s)-\frac{1}{2}\int_0^t \|F(s)x_u(s,z,y)\|^2\,ds \tag{15}$$

and $x_u(t,z,y)$ is the state at time t given by (4) and evaluated at the sample point (z,y). Let $\{x_0(t,z), 0\leqq t\leqq T\}$ be the state process generated through (7) by the identically zero control $u(t)\equiv 0$, i.e., $x_0(t,z)=\Phi(t,0)z$ and, for any admissible control law u, set

$$\mu_u(t,y)=\int_0^t \Phi(t,s)B(s)u(s,y)\,ds. \tag{16}$$

Then it is clear that

$$x_u(t,z,y)=x_0(t,z)+\mu_u(t,y). \tag{17}$$

Introduce the process $\{y_u(t), 0\leqq t\leqq T\}$ by setting

$$dy_u(t)=dy(t)-F(t)\mu_u(t,y)\,dt$$

and observe that $y_u(t)$ is $\mathcal{Y}_t$-measurable and has the expression

$$dy_u(t)=F(t)x_0(t,z)\,dt+dw_u(t).$$

Now, substituting (17) into (14), rearranging terms by using the definition of $y_u(t)$ and canceling the common factor in the numerator and denominator of (14), we obtain

$$E^u\{z\,|\,\mathcal{Y}_t\} = \frac{\int_Z \bar{z}\,\exp\{h(t, x_0(\cdot, \bar{z}), y_u)\}P_Z(d\bar{z})}{\int_Z \exp\{h(t, x_0(\cdot, \bar{z}), y_u)\}P_Z(d\bar{z})}. \tag{18}$$

The right-hand side of (18) is a function of $\{y_u(s), 0 \leqq s \leqq t\}$ and $y_u(t)$ is, as observed above, a function of z and $w_u(t)$. So we denote the right-hand side of (18) by $G(t, z, w_u)$, where the function $G(\cdot, \cdot, \cdot)$ does not depend on u. Remember here that $w_u(t)$ and z are independent under P^u and the restriction of P^u to $\mathcal{B}(Z) \otimes \{\phi, C\}$ coincides with P_Z. Furthermore, since $w_u(t)$ is the Wiener process under P^u, the restriction of $P^u \circ (\varphi^u)^{-1}$ to $\{\phi, Z\} \otimes \mathcal{B}(C)$, where $\varphi_t^u(z, y) = (z, y(t) - \int_0^t F(s)x_u(s, z, y)\,ds)$ for $(z, y) \in Z \times C$, coincides with the Wiener measure P_W which does not depend on u. Thus we find

$$\begin{aligned} &E^u\{\tilde{x}_u(t)\tilde{x}_u'(t)\} \\ &\quad = E^u\{\Phi(t, 0)(z - G(t, z, w_u))(z - G(t, z, w_u))'\Phi'(t, 0)\} \\ &\quad = \int_Z \int_C \Phi(t, 0)(\bar{z} - G(t, \bar{z}, \bar{w}))(\bar{z} - G(t, \bar{z}, \bar{w}))'\Phi'(t, 0)P_W(d\bar{w})P_Z(d\bar{z}), \end{aligned}$$

which implies that $E^u\{\tilde{x}_u(t)\tilde{x}_u'(t)\}$ is independent of the choice of admissible control laws for each t. □

Remark 1. This lemma is essentially based on the facts (i) and (ii) stated in the previous section and will thus hold in the more general context whenever the Girsanov transformation is valid, i.e., $E^*\{dP^u/dP^*\} = 1$. The boundedness assumption of initial state and control is only a sufficient condition for its validity [5].

Remark 2. Assume that the system (1)–(2) has a strong solution for a given control law. In this case, if $\{y^0(t), 0 \leqq t \leqq T\}$ denotes the observations process for the identically zero control, then the reduction from (14) to (18) implies $E\{z\,|\,\mathcal{Y}_t\} = E\{z\,|\,\mathcal{Y}_t^0\}$, where $\mathcal{Y}_t$ and $\mathcal{Y}_t^0$ are the σ-fields generated by $\{y(s), 0 \leqq s \leqq t\}$ and $\{y^0(s), 0 \leqq s \leqq t\}$, respectively, on a given probability space $(\Omega, \mathcal{F}, P)$. In the strong formulation of the problem, the above lemma follows immediately from the equality $E\{z\,|\,\mathcal{Y}_t\} = E\{z\,|\,\mathcal{Y}_t^0\}$. A reduction procedure similar to the ones validating the passage from (14) to (18) was used in [7] for establishing the innovations-observations equivalence for a linear stochastic control system with a strong solution.

On the basis of the above observations, the following result can be shown.

THEOREM. *For the stated stochastic control problem, there exists a unique optimal control taking the form* (12) *provided the condition*

$$\begin{aligned} &\max_{0 \leqq t \leqq T} \|R^{-1}(t)B'(t)S(t)\| \\ &\qquad \leqq M\left[N \max_{0 \leqq t \leqq T} \|\Phi(t, 0)\| + M \max_{0 \leqq t \leqq T} \int_0^t \|\Phi(t, s)B(s)\|\,ds\right]^{-1} \end{aligned} \tag{19}$$

is satisfied, where $\|\cdot\|$ *denotes the matrix norm induced by the Euclidean norm.*

Proof. It is sufficient to show that, if (19) holds, there exists a unique admissible control which satisfies (12) and (13).

Let $\mathcal{U}$ denote the set of admissible controls, i.e., if $u \in \mathcal{U}$, $u(t)$ is $\mathcal{Y}_t$-measurable and $\|u(t)\| \leqq M$. It is obvious that $\mathcal{U}$ is a closed set under the topology generated by the following notion of convergence: a sequence u_n, $n = 0, 1, 2, \cdots$, of mappings from $[0, T] \times C$ into R^r converges if and only if, for every y in C, the sequence of mappings

$u_n(\cdot, y)$, $n = 0, 1, 2, \cdots$, from $[0, T]$ into R^r converges uniformly in t. Now, for every u in $\mathcal{U}$, define a measurable mapping $\psi(u)$ from $[0, T]\times C$ into R^r by setting

$$[\psi(u)](t, y) = -K(t)[\Phi(t, 0)E^u\{z|\mathcal{Y}_t\} + \mu_u(t, y)], \tag{20}$$

where $E^u\{z|\mathcal{Y}_t\}$ is given by (14) and $\mu_u(t, y)$ is defined by (16). Then, by using the assumption (19) and $\|z\| \leqq N$, it is shown that $\psi_t(u)$ is $\mathcal{Y}_t$-measurable and $\|\psi_t(u)\| \leqq M$ for each $u \in \mathcal{U}$, that is, ψ is a transformation from $\mathcal{U}$ into $\mathcal{U}$. Furthermore, it is shown in the appendix that there exists a positive bounded measurable function $K(t, y)$ such that $K(t, \cdot)$ is $\mathcal{Y}_t$-measurable and yields the relation

$$\|\psi_t(u) - \psi_t(\tilde{u})\|^2 \leqq K(t, y) \int_0^t \|u(s) - \tilde{u}(s)\|^2 \, ds \tag{21}$$

for all u and $\tilde{u}$ in U. It is noted here that, since

$$\begin{aligned} \hat{x}_u(t) &= E^u\{x_u(t)|\mathcal{Y}_t\} \\ &= \Phi(t, 0)E^u\{z|\mathcal{Y}_t\} + \int_0^t \Phi(t, s)B(s)u(s) \, ds, \end{aligned} \tag{22}$$

$\psi_t(u) = -K(t)\hat{x}_u(t)$ and the admissible control satisfying (12) is a fixed point of ψ.

Now, define the sequence $u_n \in \mathcal{U}$, $n = 0, 1, 2, \cdots$, such that $u_0(t) \equiv 0$ and

$$u_{n+1}(t) = \psi_t(u_n), \qquad n = 0, 1, 2, \cdots,$$

then we have from (21)

$$\|u_{n+1}(t) - u_n(t)\|^2 \leqq \frac{[\int_0^t K(s, y) \, ds]^{n-1}}{(n-1)!} K(t, y)Mt, \qquad n = 1, 2, 3, \cdots.$$

This implies that there is a limit $u^* \in \mathcal{U}$ such that u_n converges to u^* uniformly in t, since $\mathcal{U}$ is a closed set. Furthermore, it follows from (21) that $\psi_t(u_n)$ converges to $\psi_t(u^*)$ uniformly in t, and so u^* is a fixed point of ψ, i.e.,

$$u^*(t) = \psi_t(u^*). \tag{23}$$

By the use of (21) we can also prove the uniqueness of the fixed point of ψ. Thus, noting that (23) is equivalent to

$$u^*(t) = -K(t)\hat{x}_{u^*}(t),$$

and that $\|u^*(t)\| \leqq M$, we know that u^* is the unique admissible control which satisfies (12) and (13). □

Remark 3. A priori the condition (19) could very well turn out to be vacuous, since the matrix $S(t)$ in the left-hand side depends on the matrix functions $B(t)$ and $R(t)$ via the Riccati equation (10). As shown in the following, however, there always exists at least one symmetric positive definite matrix function $R(t)$ satisfying (19). Let $S^0(t)$ be the unique solution of the linear matrix equation

$$\frac{dS^0}{dt} + A'S^0 + S^0A + Q = 0, \qquad S^0(T) = S_T,$$

and set $D(t) = S^0(t) - S(t)$ for $0 \leqq t \leqq T$, where $S(t)$ is the solution of the Riccati equation (10) corresponding to a given symmetric positive definite matrix function $R(t)$. It is easy to conclude that

$$\frac{dD}{dt} + A'D + DA = -SBR^{-1}B'S, \qquad D(T) = 0.$$

Left-multiply this latter equation by $\Phi'(t, 0)$ and right-multiply the resulting relation by $\Phi(t, 0)$; it follows that

$$\frac{d}{dt}\Phi'(t, 0)D(t)\Phi(t, 0) = -\Phi'(t, 0)S(t)B(t)R^{-1}(t)B'(t)S(t)\Phi(t, 0)$$

and after integration

$$D(t) = \int_t^T \Phi'(s, t)S(s)B(s)R^{-1}(s)B'(s)S(s)\Phi(s, t)\, ds.$$

The relation $\|S(t)\| \leqq \|S^0(t)\|$ for all t in $[0, T]$ is now immediate from the expression obtained for $D(t)$ and from the positive definiteness of $R^{-1}(t)$. It follows from this inequality that, for all t in $[0, T]$, $\|R^{-1}(t)B'(t)S(t)\| \leqq \|R^{-1}(t)\|\|B(t)\|\|S^0(t)\|$. Here note that $S^0(t)$ is given independently of $R(t)$. Thus, if $R(t)$ is taken as $\|R^{-1}(t)\| \leqq \alpha(\max_{0\leqq s\leqq T}\|B(s)\|\,\|S^0(s)\|)^{-1}$, where α denotes the right-hand side of (19), we have

$$\|R^{-1}(t)B'(t)S(t)\| \leqq \|R^{-1}(t)\| \max_{0\leqq s\leqq T}\|B(s)\|\,\|S^0(s)\| \leqq \alpha,$$

and so we have (19). (If $\|B(t)\|\,\|S^0(t)\| = 0$ for all t in $[0, T]$, (19) is automatically satisfied for any weighting matrix $R(t)$.) A simple way of choosing such an weighting matrix $R(t)$ is as follows: first find a positive number β such that $\beta \leqq \alpha(\max_{0\leqq s\leqq T}\|B(s)\|\,\|S^0(s)\|)^{-1}$ and then set $R(t) = \beta^{-1}I$, where I is the r-dimensional identity matrix.

Remark 4. Consider the corresponding deterministic problem with nonrandom initial conditions and no observations. Then, under the assumption (19), the optimal control is given by $u(t) = -K(t)x(t)$. Comparing this with (12), we conclude that our stochastic control problem has the certainty equivalence property.

Remark 5. To implement the optimal control, we need the state estimate $\{\hat{x}_u(t), 0 \leqq t \leqq T\}$, which is given by (22) with the Bayes formula (14). It follows from the non-Gaussian and bounded character of the initial state that $\hat{x}_u(t)$ is generally nonlinear and in particular non-Lipschitzian in the observation $\{y(s), 0 \leqq s \leqq t\}$: Note that the boundness of the process $\{x_u(t), 0 \leqq t \leqq T\}$, which comes from the boundnesses of the initial state and of the control, implies the same property for the state estimate. This automatically precludes the existence of a linear or even Lipschitz-like functional dependence of $\hat{x}_u(t)$ on the observation $\{y(s), 0 \leqq s \leqq t\}$, since the latter process is a Wiener process under P^* and any linear or Lipschitz transformation on its path will not produce in general a bounded process.

Appendix. We shall prove the inequality (21). For short, write $h_t^u(z) = h(t, x_u(\cdot, z, y), y)$ and $x_t^u(z) = x_u(t, z, y)$. Using $\|z\| \leqq N$, we have from (20)

$$\begin{aligned}\|\psi_t(u) - \psi_t(\tilde{u})\| &\leqq 2\|K(t)\|N\frac{\int_Z \|\exp(h_t^u(z)) - \exp(h_t^{\tilde{u}}(z))\|P_Z(dz)}{\int_Z \exp(h_t^u(z))P_Z(dz)} \\ &\quad + k_1(t)\int_0^t \|u(s) - \tilde{u}(s)\|\, ds,\end{aligned}$$

where $k_1(t) = \max_{0\leqq s\leqq T}\|K(t)\Phi(t, s)B(s)\|$. Further, using the following inequality in the first term of the right-hand side,

$$\exp\xi - \exp\zeta \leqq \tfrac{1}{2}(\exp\xi + \exp\zeta)\|\xi - \zeta\|,$$

for $\xi, \zeta \in R$, we have

$$\begin{aligned}&\|\psi_t(u)-\psi_t(\tilde{u})\| \\ &\leqq \|K(t)\|N \frac{\int_Z (\exp(h_t^u(z))+\exp(h_t^{\tilde{u}}(z)))\|h_t^u(z)-h_t^{\tilde{u}}(z)\|P_Z(dz)}{\int_Z \exp(h_t^u(z))P_Z(dz)} \\ &\quad + k_1(t)\int_0^t \|u(s)-\tilde{u}(s)\|\,ds.\end{aligned} \tag{24}$$

Here note that

$$\begin{aligned}h_t^u - h_t^{\tilde{u}} &= \int_0^t (F(s)(x_s^u - x_s^{\tilde{u}}))'\,dy(s) - \frac{1}{2}\int_0^t (F(s)(x_s^u + x_s^{\tilde{u}}))'(F(s)(x_s^u - x_s^{\tilde{u}}))\,ds \\ &= (x_t^u - x_t^{\tilde{u}})'\int_0^t F'(s)\,dy(s) - \int_0^t \frac{\partial(x_s^u - x_s^{\tilde{u}})'}{\partial s}\int_0^s F'(r)\,dy(r)\,ds \\ &\quad - \frac{1}{2}\int_0^t (F(s)(x_s^u + x_s^{\tilde{u}}))'(F(s)(x_s^u - x_s^{\tilde{u}}))\,ds,\end{aligned} \tag{25}$$

where

$$x_t^u(z) - x_t^{\tilde{u}}(z) = \int_0^t \Phi(t, s)B(s)(u(s)-\tilde{u}(s))\,ds, \tag{26}$$

and also note that $x_t^u(z)$ and $x_t^{\tilde{u}}(z)$ are bounded, since $\|z\| \leqq N$, $\|u(t)\| \leqq M$ and $\|\tilde{u}(t)\| \leqq M$. Substituting (26) into (25) and using the Schwarz inequality, we find bounded $\mathcal{Y}_t$-adpated functions $k_2(t, y)$ and $k_3(t, y)$ such that

$$\|h_t^u(z) - h_t^{\tilde{u}}(z)\| \leqq k_2(t, y)\int_0^t \|u(s)-\tilde{u}(s)\|\,ds + k_3(t, y)\left[\int_0^t \|u(s)-\tilde{u}(s)\|^2\,ds\right]^{1/2}. \tag{27}$$

From (24) and (27) we obtain

$$\begin{aligned}&\|\psi_t(u)-\psi_t(\tilde{u})\| \\ &\leqq \|K(t)\|N\left\{1 + \frac{\int_Z \exp(h_t^{\tilde{u}}(z))P_Z(dz)}{\int_Z \exp(h_t^u(z))P_Z(dz)}\right\}\left\{\begin{matrix}\text{right-hand} \\ \text{side of (27)}\end{matrix}\right\} \\ &\quad + k_1(t)\int_0^t \|u(s)-\tilde{u}(s)\|\,ds.\end{aligned} \tag{28}$$

To bound the first factor on the right-hand side of (28), note that by the same procedure as in deriving (27) we can find a bounded $\mathcal{Y}_t$-adapted function $k_4(t, y)$ such that

$$\left\|\int_0^t (F(s)(x_s^u(z) - x_s^{\tilde{u}}(z)))'\,dy(s)\right\| \leqq k_4(t, y).$$

Since $x_t^u(z)$ and $x_t^{\tilde{u}}(z)$ have the same upper bound, denote $\|F(t)x_t^u(z)\| \leqq L$ and $\|F(t)x_t^{\tilde{u}}(z)\| \leqq L$. Then, we find

$$\begin{aligned}&\frac{\int_Z \exp(h_t^{\tilde{u}}(z))P_Z(dz)}{\int_Z \exp(h_t^u(z))P_Z(dz)} \\ &\leqq \frac{\int_Z \exp\{\int_0^t (F(s)x_s^u(z))'\,dy(s) + k_4(t, y) + \frac{1}{2}tL^2\}P_Z(dz)}{\int_Z \exp\{\int_0^t (F(s)x_s^u(z))'\,dy(s) - \frac{1}{2}tL^2\}P_Z(dz)} \\ &= \exp(k_4(t, y) + tL^2).\end{aligned} \tag{29}$$

Now it follows from (28) and (29) that

$$(30)\quad \|\psi_t(u)-\psi_t(\tilde{u})\| \leqq K_1(t, y)\int_0^t \|u(s)-\tilde{u}(s)\|\,ds + K_2(t, y)\left[\int_0^t \|u(s)-\tilde{u}(s)\|^2\,ds\right]^{1/2},$$

where

$$K_1(t, y) = \|K(t)\|Nk_2(t, y)(1+\exp(k_4(t, y)+tL^2))+k_1(t),$$
$$K_2(t, y) = \|K(t)\|Nk_3(t, y)(1+\exp(k_4(t, y)+tL^2)).$$

Thus, from (30), we obtain

$$\|\psi_t(u)-\psi_t(\tilde{u})\|^2 \leqq 2(tK_1^2(t, y)+K_2^2(t, y))\int_0^t \|u(s)-\tilde{u}(s)\|^2\,ds,$$

which is the desired inequality (21).

Acknowledgment. The author wishes to thank the referees for their comments and suggestions. In particular, he is deeply indebted to one of the referees for the comment on the original version of Remark 3.

REFERENCES

[1] A. V. BALAKRISHNAN, *A note on the structure of optimal stochastic controls*, Appl. Math. Optim., 1 (1974), pp. 87–94.
[2] V. E. BENEŠ, *Existence of optimal stochastic control laws*, this journal, 9 (1971), pp. 446–472.
[3] M. H. A. DAVIS AND P. P. VARAIYA, *Information sets in linear stochastic systems*, J. Math. Anal. Appl., 37 (1972), pp. 384–402.
[4] M. H. A. DAVIS, *The separation principle in stochastic control via Girsanov solution*, this Journal, 14 (1976), pp. 176–188.
[5] I. V. GIRSANOV, *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*, Theory Prob. Appl., 5 (1960), pp. 285–301.
[6] G. KALLIANPUR AND C. STRIEBEL, *Estimation of stochastic processes: arbitrary system process with additive white noise error*, Ann. Math. Stat., 39 (1968), pp. 785–801.
[7] G. KALLIANPUR, *A linear stochastic system with discontinuous control*, Proc. International Symposium on SDE, Kyoto, 1976, Kinokuniya Book-Store, Co., Ltd., Tokyo, Japan, 1978.
[8] A. LINDQUIST, *On feedback control of linear stochastic systems*, this Journal, 11 (1973), pp. 323–343.
[9] ———, *Comments on "A simple proof of the separation theorem for linear stochastic systems with time delays,"* IEEE Trans. Automat. Contr., AC-25 (1980), pp. 274–275.
[10] K. UCHIDA AND E. SHIMEMURA, *On certainty equivalence in linear-quadratic control problem with nonlinear measurements*, Inform. Control, 41 (1979), pp. 119–135.
[11] K. UCHIDA, *On optimal control of the stochastic systems with delayed controls and delayed measurements*, J. Math. Anal. Appl., 75 (1980), pp. 454–464.
[12] W. M. WONHAM, *On the separation theorem of stochastic control*, this Journal, 6 (1968), pp. 312–326.
[13] ———, *On a matrix Riccati equation of stochastic control*, this Journal, 6 (1968), pp. 681–697.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0003 $01.25/0

LIE BRACKETS AND LOCAL CONTROLLABILITY: A SUFFICIENT CONDITION FOR SCALAR-INPUT SYSTEMS*

HÉCTOR J. SUSSMANN†

Abstract. We develop a general formalism, based on exponential Lie series, which can be used to study control variations. We apply the formalism to prove a result, conjectured by H. Hermes, that gives a sufficient condition for small-time local controllability for single-input systems.

1. Introduction. This paper deals with control systems with a scalar input which enters linearly in the dynamical equations. We prove that a certain condition involving Lie brackets is sufficient for local controllability from a point x_0. The condition considered here was first introduced by H. Hermes, and we will refer to it as the *Hermes local controllability condition* (HLCC). Hermes conjectured that this condition would be sufficient for local controllability, and the primary goal of our paper is to prove that conjecture. (Hermes himself proved the conjecture for systems in the plane, and he has recently announced partial results for more general systems, cf. [13] and [14].)

There are, however, other reasons why the results presented here may be of interest, and this is why we have chosen to present some of our preliminary results in a more general form than is actually needed for the specific problem considered here. It has been clear for a long time (at least since R. Hermann's work [1]) that many properties of control systems should be characterizable in terms of Lie bracket configurations. A precise explanation of why this should be so is provided by *Nagano's theorem*: if $\{X_i: i \in I\}$, $\{Y_i: i \in I\}$ are two transitive families of analytic vector fields on manifolds M, M', and if m, m' are points of M, M' such that there is a linear isomorphism $L: T_mM \to T_{m'}M'$ between the tangent spaces of M at m and of M' at m', which maps every vector $(\operatorname{ad} X_{i_1}) \cdots (\operatorname{ad} X_{i_k})(X_{i_{k+1}})(m)$ to the corresponding vector $(\operatorname{ad} Y_{i_1}) \cdots (\operatorname{ad} Y_{i_k})(Y_{i_{k+1}})(m')$, then there is a diffeomorphism Λ from a neighborhood of m to a neighborhood of m', such that $\Lambda_*(X_i) = Y_i$ for $i \in I$. (A family $\{X_i: i \in I\}$ is *transitive* if the evaluation at every point of the Lie algebra generated by the X_i is the full tangent space at the point.) When this is applied to control systems

$$\dot{x} = f(x, u), \qquad x \in M, \quad u \in U, \tag{1.1}$$

$$\dot{y} = g(y, u), \qquad y \in M', \quad u \in U', \tag{1.2}$$

with f, g analytic, this gives an important conclusion. Define the *Lie configuration* of (1.1) at x_0 to be the specification of all relations at x_0 between the Lie brackets of the vector fields $f(\cdot, u)$ (equivalently, the kernel of the map $\operatorname{Lie}(U) \to T_{x_0}M$ which sends each $[X_{u_1}, [X_{u_2}, [\cdots [X_{u_{m-1}}, X_{u_m}] \cdots]]]$ to the result of plugging in $f(\cdot, u_i)$ for X_{u_i} and then evaluating at x_0. Here $\operatorname{Lie}(U)$ is the free Lie algebra in a set of indeterminates $\{X_u: u \in U\}$). Then if (1.1) and (1.2) have equivalent configurations at x_0, y_0 (the definition of equivalence being the obvious one), it follows that there is a diffeomorphism that maps trajectories of (1.1) from x_0 to trajectories of (1.2) from y_0. This shows, roughly, that all "diffeomorphism invariant" properties of an analytic system with a given initial state x_0 should be describable in terms of the Lie bracket configuration of the system at x_0. Examples of such properties are: accessibility (i.e.

* Received by the editors September 1, 1981, and in revised form August 30, 1982. This research was partially supported by the National Science Foundation under grant MCS 78-02442.

† Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903.

whether it is possible to reach an open set), local controllability (i.e. whether one can reach a neighborhood of x_0), small-time local controllability (i.e. whether it is possible to reach a full neighborhood of x_0 in arbitrarily small time), "bang-bang theorem" (i.e. whether everything reachable from x_0 can actually be reached by a bang-bang trajectory), various properties of reachable sets (e.g. whether the reachable set from x_0 is a locally finite union of manifolds), of time-optimal trajectories (e.g. whether every time-optimal trajectory from x_0 is bang-bang, or piecewise analytic, or piecewise smooth), of the optimal cost function (e.g. whether it is "piecewise analytic" in some sense) and of the time-optimal feedback (e.g. whether it is smooth except on a locally finite family of manifolds that are "switching loci").

These considerations suggest a general program of research, whose aim is to characterize properties such as the ones listed above in terms of Lie configurations. Much has actually been done in this direction, e.g. in the work on "Chow's theorem" (Hermann [1], Lobry [5], Krener [3], Sussmann–Jurdjevic [8]), on bang-bang theorems (Krener [3], Sussmann [9]), on controllability about a reference trajectory (Hermes [2]), on high order optimality conditions (Krener [4]), and on local controllability (Sussmann [10]). However, the peculiar fact that control trajectories are not time-reversible makes it hard to figure out what the Lie algebraic conditions should be and, until now, there was no satisfactory result where this positivity of time was really brought into play.

Our work here should be regarded within the context of the program outlined above. The small-time local controllability (STLC) question is trivial if "time" is made "reversible" by allowing motion along trajectories in either time direction. (The answer is provided by "Chow's theorem".) Our result considers the problem without allowing motion "backwards in time", so it is a genuine "positive-time" theorem. In our view, the main interest of the result given here is that it suggests the possibility of further "positive-time" results for other problems of a similar nature. For this reason, we start the paper by developing a general "exponential Lie series" formalism for arbitrary systems with a control entering linearly, even though we only apply it here to the scalar-input case. The formalism makes it possible to study control variations in a systematic way and, at least for the problem studied here, it avoids cumbersome inductive constructions.

Even for the class of problems considered here, the STLC question is not completely settled. It is not true, for instance, that the HLCC is necessary as well as sufficient for STLC. In § 5, we prove that some particularly simple violations of the HLCC imply noncontrollability. But, in § 6, we show that it is possible for a system to be STLC even if the HLCC fails. So, it is not clear what the true necessary and sufficient condition is.

It is also unclear how to generalize the results to the multi-input case. But the most exciting challenge is to attempt to establish a link between our work and the problems of controllability about a reference trajectory (CART), and of high-order optimality conditions. The STLC problem is, after all, a particular case of the CART problem, the reference trajectory being just a point.

The paper is organized as follows: in § 2 we state the main result, and we discuss how it relates to the linear criterion; in § 3 we introduce the general exponential Lie series formalism; in § 4 we show that the exponential Lie series gives an asymptotic expansion for the trajectories, as $t \to 0$; in § 5 we use the formalism to prove the main result. Finally, in § 6 we show that at least some consequences of the HLCC are necessary conditions for STLC, but we give an example (due to B. Jakubczyk) which shows that HLCC is not necessary for STLC.

2. Statement of the main result. Let f_0, f_1 be smooth vector fields on a manifold M. Let Lie (f_0, f_1) denote the Lie algebra of vector fields generated by f_0 and f_1. For each integer $k \geqq 0$, we define a linear subspace $\mathscr{S}^k(f_0, f_1)$ of Lie (f_0, f_1) as follows: $\mathscr{S}^k(f_0, f_1)$ is the linear span of all Lie monomials in f_0 and f_1 which involve f_1 at most k times. Thus $\mathscr{S}^0(f_0, f_1) = \{\alpha f_0 : \alpha \in \mathbb{R}\}$, $\mathscr{S}^1(f_0, f_1)$ is the linear span of f_0, f_1, and the brackets $(\operatorname{ad} f_0)^j f_1$ for $j \geqq 1$, etc.

Let us consider a control system

$$\frac{dx}{dt} = f_0(x) + uf_1(x), \qquad |u(t)| \leqq A \tag{2.1}$$

where the state x belongs to M, and where the control is subject to the inequality constraint $-A \leqq u \leqq A$. For any point $x_0 \in M$, we let Reach (x_0, t) denote the set of all points that can be reached from x_0 by means of a trajectory of the system (2.1) in no more than t units of time. We say that (2.1) is *small-time locally controllable* (STLC) *from* x_0, if for every $t > 0$ the set Reach (x_0, t) contains a neighborhood of x_0.

We say that x_0 is an *equilibrium point* for (2.1) if there is a $\bar{u} \in \mathbb{R}$, $|\bar{u}| \leqq A$, such that $f_0(x_0) + \bar{u} f_1(x_0) = 0$. We say that x_0 is a *regular equilibrium point* if $f_0(x_0) + \bar{u} f_1(x_0) = 0$ for a $\bar{u}$ such that $|\bar{u}| < A$. Except for the trivial situation when $f_0(x_0) = f_1(x_0) = 0$ (in which case Reach $(x_0, t) = \{x_0\}$ for all t), the $\bar{u}$ of the preceding equations, if it exists, must be unique.

We say that (2.1) satisfies the *Hermes local controllability condition* (HLCC) at x_0 if

(HLCC 1) x_0 is a regular equilibrium point,

(HLCC 2) dim Lie $(f_0, f_1)(x_0) =$ dim M and

(HLCC 3) the increasing sequence of subspaces $\{\mathscr{S}^k(f_0 + \bar{u} f_1, f_1) : k = 0, 1, \cdots\}$ satisfies

$$\mathscr{S}^k(f_0 + \bar{u} f_1, f_1)(x_0) = \mathscr{S}^{k+1}(f_0 + \bar{u} f_1, f_1)(x_0)$$

whenever k is odd.

In conditions (HLCC 2), (HLCC 3) we have used $V(x_0)$, whenever V is a set of vector fields, to indicate the set of all values at x_0 of the elements of V. Also, in (HLCC 3), $\bar{u}$ is the unique number such that $f_0(x_0) + \bar{u} f_1(x_0) = 0$, and therefore $|\bar{u}| < A$.

Our main result is

THEOREM 2.1. *If* (2.1) *satisfies the* HLCC *at* x_0, *then* (2.1) *is small-time locally controllable from* x_0.

The remaining sections will be devoted to the proof of Theorem 2.1. Here we will describe how the result relates to the well-known sufficient condition in terms of the linearization of the system. Choose coordinates in a neighborhood of x_0, so that f_0 and f_1 are, simply, vector-valued functions. The linearization of (2.1) about $(x_0, \bar{u})$ is the system

$$\frac{d\xi}{dt} = A\xi + vb, \tag{2.2}$$

where $\xi = x - x_0$, $v = u - \bar{u}$, b is the vector $f_1(x_0)$, and A is the Jacobian matrix at x_0 of the vector-valued function $x \to f_0(x) + \bar{u} f_1(x)$. Then a simple computation shows that $\mathscr{S}^1(f_0 + \bar{u} f_1, f_1)(x_0)$ is the linear span of the vectors $A^j b$, $j = 0, 1, 2, \cdots$. So, if the linearization of (2.1) about $(x_0, \bar{u})$ is controllable, $\mathscr{S}^1(f_0 + \bar{u} f_1, f_1)(x_0)$ is the whole space, and then conditions (HLCC 2) and (HLCC 3) are trivially satisfied. So our Theorem 2.1 implies in particular the linear criterion: if x_0 is a regular equilibrium

point for (2.1), and if the linearization of (2.1) about $(x_0, \bar{u})$ is controllable ($\bar{u}$ being the unique u for which $f_0(x_0)+uf_1(x_0)=0$), then (2.1) is STLC from x_0.

3. The exponential Lie series formalism. In this section, we present a formalism for the study of the asymptotic behavior of control trajectories as time, or some other parameter on which a control may depend, approaches zero.

To each control $u(\cdot)$ we will associate a series, denoted by Ser (u). This series has also been considered—and used for other purposes—by M. Fliess under the name of "Chen series" (cf. Chen [22] and Fliess [21]).

We let $\mathcal{U}_m$ denote the set of all $\mathbb{R}^m$-valued measurable functions $u(\cdot)$ whose domain is some compact interval of the form $[0, T]$, and which are Lebesgue integrable on $[0, T]$. We identify any two elements of $\mathcal{U}_m$ that are equal almost everywhere. If $u(\cdot):[0, T]\to\mathbb{R}^m$ is an element of $\mathbb{R}^m$, then the time T will be referred to as the *terminal time* of $u(\cdot)$, and denoted by $T(u)$. If $0\leqq t\leqq T(u)$, then one can consider the element of $\mathcal{U}_m$ obtained by restricting $u(\cdot)$ to the interval $[0, t]$. We use $u^t(\cdot)$ to denote this element (so that $T(u^t)=t$). If $u(\cdot)\in\mathcal{U}_m$, we will always use $u_1, \cdots, u_m$ to denote its components. We always write $u_0\equiv 1$.

Now let $\mathbf{X}=(X_0, \cdots, X_m)$ be a finite sequence of objects, that will be called *indeterminates*. We use $A(\mathbf{X})$ to denote the free associative algebra in $X_0, \cdots, X_m$. If $I=(i_1, \cdots, i_k)$ is any finite sequence such that the i_j are integers and $0\leqq i_j\leqq m$, then we define

$$X_I = X_{i_1}X_{i_2}\cdots X_{i_k}.$$

We let $X_\varnothing = 1$. Then the monomials X_I form a basis for $A(\mathbf{X})$. If we let $I*J$ denote the concatenation of I and J (i.e. the sequence obtained by writing first the elements of I, and then those of J), then multiplication in A_m is simply given by

$$X_I X_J = X_{I*J}.$$

We also want to consider the algebra $\hat{A}(\mathbf{X})$ of *formal power series* in $X_0, \cdots, X_m$. The elements of $\hat{A}(\mathbf{X})$ are the formal sums $\sum_I a_I X_I$, where I ranges over all multi-indices.

Given an element P of $\hat{A}(\mathbf{X})$ and a $u(\cdot)\in\mathcal{U}_m$, we can consider the differential equation

$$\frac{dS}{dt} = S\Big(X_0+\sum_{i=1}^{m} u_i X_i\Big) \tag{3.1}$$

with initial condition $S(0)=P$. A *solution* of (3.1) is a function $S:[0, T(u)]\to\hat{A}(\mathbf{X})$ such that $S(0)=P$ and that (3.1) holds for each coefficient. So, if $P=\sum p_I X_I$, and $S(t)=\sum_I s_I(t)X_I$, then $S(\cdot)$ is a solution if and only if, for each I, the conditions

$$S_I(0) = p_I, \tag{3.2a}$$

$$\dot{S}_I(t) = S_J(t)u_i(t) \tag{3.2b}$$

are satisfied, where $I=J*\{i\}$. This shows that, for any given P, the solution of (3.1) with initial condition $S(0)=P$ exists and is unique (because formulas (3.2a) and (3.2b) enable us to compute the S_I recursively).

In the particular case when $P=1$, Formulas (3.2a) and (3.2b) yield $S_\varnothing(t)\equiv 1$, and

$$S_{J*\{i\}}(t) = \int_0^t u_i(\tau)S_J(\tau)\,d\tau.$$

From this one obtains by induction the formulas

$$S_I(t) = \int_0^t u_I \tag{3.3}$$

where $\int_0^t u_I$ denotes the *iterated integral*

$$\int_0^t u_I = \int_0^t \int_0^{\tau_k} \int_0^{\tau_{k-1}} \cdots \int_0^{\tau_2} u_{i_k}(\tau_k) u_{i_{k-1}}(\tau_{k-1}) \cdots u_{i_2}(\tau_2) u_{i_1}(\tau_1)\, d\tau_1\, d\tau_2 \cdots d\tau_k \tag{3.4}$$

if $\varnothing \neq I = (i_1, \cdots, i_k)$, and where we let $\int_0^t u_\varnothing = 1$.

If $u(\cdot) \in \mathcal{U}_m$, we define Ser (u)—the *formal series* of u—to be $S(T(u))$, where $S(\cdot)$ is the solution of (3.1) with initial condition $S(0) = 1$.

The set $\mathcal{U}_m$ is a semigroup under the operation of concatenation. We write $u \# v$ for the concatenation of u and v, i.e.

$$u \# v : [0, T(u) + T(v)] \to \mathbb{R}^m \tag{3.5a}$$

is defined by

$$(u \# v)(t) = \begin{cases} u(t) & \text{for } 0 \leqq t < T(u), \\ v(t - T(u)) & \text{for } T(u) < t \leqq T(u) + T(v). \end{cases} \tag{3.5b}$$

Then we have

LEMMA 3.1. *The map* Ser $: \mathcal{U}_m \to \hat{A}(\mathbf{X})$ *is injective, and satisfies*

$$\text{Ser}\,(u \# v) = \text{Ser}\,(u)\,\text{Ser}\,(v) \tag{3.6}$$

for $u, v \in \mathcal{U}_m$.

Proof. Let $u \in \mathcal{U}_m$, $u : [0, T] \to \mathbb{R}^m$. Then the coefficient of X_0 in Ser (u) is T, so T can be recovered from Ser (u). If $1 \leqq i \leqq m$, and $0 \leqq k$, the coefficient of $X_0^k X_i$ in Ser (u) is

$$\frac{1}{k!} \int_0^T t^k u_i(t)\, dt.$$

Hence Ser (u) determines, for each i, all the integrals $\int_0^T t^k u_i(t)\, dt$, for all k. Therefore each u_i is completely determined by Ser (u), so Ser $(\cdot)$ is one-to-one.

If u, v are in $\mathcal{U}_m$, let $T = T(v)$, $T' = T(u)$. Let $t \to S_1(t)$ be the solution of

$$\dot{S} = S(X_0 + \sum_{i=1}^m v_i X_i) \tag{3.7a}$$

for $0 \leqq t \leqq T$, with initial condition $S_1(0) = 1$. Then $S_1(T) = \text{Ser}\,(v)$. Left multiplication by the constant element Ser (u) implies that $t \to \text{Ser}\,(u) S_1(t)$ is also a solution of (3.7a), with initial condition Ser (u). On the other hand, if we let $S_2(\cdot)$ be the solution of

$$\dot{S} = S(X_0 + \textstyle\sum (u \# v)_i X_i), \qquad S(0) = 1, \tag{3.7b}$$

we see that $t \to S_2(t + T')$, $0 \leqq t \leqq T$, is also a solution of (3.7a) which equals Ser (u) when $t = 0$. Therefore

$$S_2(t + T') = \text{Ser}\,(u) S_1(t)$$

for $0 \leqq t \leqq T$. In particular, if we let $t = T$, we find that Ser $(u \# v) =$ Ser (u) Ser (v). Q.E.D.

In addition to the formal power series Ser $(u) \in \hat{A}(\mathbf{X})$, we also want to consider truncated series. We let $A^N(\mathbf{X})$ denote the free nilpotent associative algebra of order $N + 1$. A basis for $A^N(\mathbf{X})$ consists of all monomials X_I for $|I| \leqq N$ (where $|I|$ denotes

the length of I), and multiplication is defined as in $A(\mathbf{X})$, except that, whenever $|I|+|J|>N$, then the product $X_I X_J$ is set equal to zero.

If $S \in \hat{A}(\mathbf{X})$, we use S_N to denote the series obtained from S by simply deleting all the monomials X_I with $|I|>N$. Then $S_N \in A^N(\mathbf{X})$. If $u \in \mathcal{U}_m$, then we write

$$\mathrm{Ser}_N(u) = [\mathrm{Ser}(u)]_N.$$

The $A^N(\mathbf{X})$-valued map $t \to \mathrm{Ser}_N(u^t)$, $0 \leqq t \leqq T(u)$, is the unique solution of

$$\dot{S} = (R_0 + \sum u_i R_i)(S), \qquad S(0) = 1, \tag{3.8}$$

where $R_i : A^N(\mathbf{X}) \to A^N(\mathbf{X})$ is the linear map defined by right multiplication (in $A^N(\mathbf{X})$) by X_i (i.e. $R_i(S) = SX_i$)).

If S is any formal series in $\hat{A}(\mathbf{X})$, or in $A^N(\mathbf{X})$, we use $\omega(S)$ to denote the *order* of S, i.e. the degree of the monomial X_I of lowest degree which appears in S with a nonzero coefficient. (If $S=0$ we let $\omega(S) = +\infty$.) If $\{a_k\}$ is an arbitrary sequence of numbers, and if $\omega(S) \geqq 1$, then the sum

$$\sum_{k=0}^{\infty} a_k S^k$$

is well defined. In particular, if we use $\hat{A}_0(\mathbf{X})$, $A_0^N(\mathbf{X})$ to denote the set of elements of $\hat{A}(\mathbf{X})$ (or of $A^N(\mathbf{X})$) whose order is not zero, we find that

$$e^S = \sum_{k=0}^{\infty} \frac{1}{k!} S^k \tag{3.9}$$

and

$$\log(1+S) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} S^k \tag{3.10}$$

are well defined for $S \in \hat{A}_0(\mathbf{X})$, or $S \in A_0^N(\mathbf{X})$, and that the identities

$$e^{\log S} = S, \tag{3.11}$$

$$\log e^{S'} = S' \tag{3.12}$$

hold whenever $\omega(S-1)>0$, $\omega(S')>0$.

The algebras $A(\mathbf{X})$, $\hat{A}(\mathbf{X})$, $A^N(\mathbf{X})$ are Lie algebras, with the Lie bracket being defined, as usual, by

$$[S, T] = ST - TS.$$

We use $L(\mathbf{X})$, $L^N(\mathbf{X})$ to denote, respectively, the Lie subalgebras of $A(\mathbf{X})$ and of $A^N(\mathbf{X})$ generated by $X_0, \cdots, X_m$. The elements of $L(\mathbf{X})$, or of $L^N(\mathbf{X})$, will be referred to as *Lie polynomials* in $X_0, \cdots, X_m$. We also want to consider the Lie algebra $\hat{L}(\mathbf{X}) \subseteq \hat{A}(\mathbf{X})$ of all formal sums $\sum_{k=1}^{\infty} P_k$, where each P_k is a homogeneous Lie polynomial of degree k. The elements of $\hat{L}(\mathbf{X})$ are the *Lie series* in $X_0, \cdots, X_m$.

If P and Q are Lie series, then $\omega(P) \geqq 1$ and $\omega(Q) \geqq 1$, so that e^P and e^Q are well defined, and $e^P e^Q - 1 \in \hat{A}_0(\mathbf{X})$. Therefore $\log(e^P e^Q)$ is well defined. We will use the *Campbell–Hausdorff formula* (cf., e.g. Serre [7]), which says that there is a Lie series $\mathrm{CH}(A, B)$ in two indeterminates A, B, with the property that, whenever $P \in \hat{A}_0(\mathbf{X})$ and $Q \in \hat{A}_0(\mathbf{X})$, then

$$\mathrm{CH}(P, Q) = \log(e^P e^Q). \tag{3.13}$$

(An explicit formula for $\mathrm{CH}(A, B)$ is given, e.g., in Serre [7]. The formula up to order three is

$$\mathrm{CH}(A,B) = A + B + \tfrac{1}{2}[A,B] + \tfrac{1}{12}[A,[A,B]] + \tfrac{1}{12}[B,[B,A]] + \cdots . \tag{3.14}$$

Formulas (3.13), (3.14) are also valid for $P, Q \in A_0^N(\mathbf{X})$, and in this case $\mathrm{CH}(P, Q)$ turns out to be in $A_0^N(\mathbf{X})$ again. A particular consequence of the Campbell–Hausdorff formula is that, if P, Q are Lie series, then

$$e^P e^Q = e^R \tag{3.15}$$

where R is again a Lie series.

Let us define $\hat{G}(\mathbf{X})$, $G^N(\mathbf{X})$ to be the sets

$$\hat{G}(\mathbf{X}) = \{P \in \hat{A}(\mathbf{X}) : \log P \in \hat{L}(\mathbf{X})\} \tag{3.16}$$

and

$$G^N(\mathbf{X}) = \{P \in A^N(\mathbf{X}) : \log P \in L^N(\mathbf{X})\}. \tag{3.17}$$

Then $\hat{G}(\mathbf{X})$ and the $G^N(\mathbf{X})$ are clearly groups, because of the Campbell–Hausdorff formula. The elements of $G(\mathbf{X})$ are called *exponential Lie series*. For each N, $A^N(\mathbf{X})$ is a finite-dimensional algebra with identity, and so it can be identified with a subalgebra of the algebra $\mathrm{Lin}(A^N(\mathbf{X}))$ of all linear maps $A^N(\mathbf{X}) \to A^N(\mathbf{X})$ (by identifying each $P \in A^N(\mathbf{X})$ with the map $S \to PS$). Then $G^N(\mathbf{X})$ becomes identified with a closed subgroup of the group of invertible elements of $\mathrm{Lin}(A^N(\mathbf{X}))$, so $G^N(\mathbf{X})$ is a Lie group whose Lie algebra is $L^N(\mathbf{X})$.

We now consider the differential equation

$$\dot{S} = S\left(X_0 + \sum_{i=1}^{m} u_i X_i\right) \tag{3.18}$$

as an equation in $A^N(\mathbf{X})$. Then (3.18) is of the form

$$S = F_0(S) + \sum_{i=1}^{m} u_i F_i(S) \tag{3.19}$$

where $F_0, \cdots, F_m$ are the vector fields on $A^N(\mathbf{X})$ given by

$$F_i(S) = SX_i. \tag{3.20}$$

(Since $A^N(\mathbf{X})$ is a vector space, then a vector field on $A^N(\mathbf{X})$ can be thought of, simply, as a map from $A^N(\mathbf{X})$ to $A^N(\mathbf{X})$.) More generally, if $P \in A^N(\mathbf{X})$, then one can consider the vector field F^P defined by

$$F^P(S) = SP.$$

An easy computation shows that

$$[F^P, F^Q] = F^{[P,Q]} \quad \text{for } P, Q \in A^N(\mathbf{X}). \tag{3.21}$$

Let $\Lambda^N(\mathbf{X})$ be the Lie algebra of vector fields on $A^N(\mathbf{X})$ generated by $F_0, \cdots, F_m$. Then (3.21) shows that $\Lambda^N(\mathbf{X})$ is isomorphic to $L^N(\mathbf{X})$, the map $P \to F^P$ being the isomorphism. It follows from general accessibility theory (cf. [8]) that there is a unique maximal integral manifold I of $\Lambda^N(\mathbf{X})$ through the point 1. (Precisely, I is characterized by the fact that (i) it is a connected submanifold of $A^N(\mathbf{X})$, (ii) its tangent space at each $P \in I$ is the set of all vectors $V(P)$, as V varies over $\Lambda^N(\mathbf{X})$, (iii) $1 \in I$, and (iv) there is no larger J that satisfies (i), (ii) and (iii).) Moreover, I is the *orbit* of $\Lambda^N(\mathbf{X})$ through the point 1 (cf. Sussmann [11]), i.e. the smallest subset J of $A^N(\mathbf{X})$ with the

property that $1 \in J$ and that, whenever $S \in J$ and $V \in \Lambda^N(\mathbf{X})$, then the integral trajectory of V through S is entirely contained in J.

If $V \in \Lambda^N(\mathbf{X})$, and $V = F^P$ for $P \in L^N(\mathbf{X})$, then the integral trajectory of V through an $S \in A^N(\mathbf{X})$ is the curve $t \to S e^{tP}$. Therefore I is the set of all products

$$e^{t_1 P_1} e^{t_2 P_2} \cdots e^{t_k P_k} \tag{3.22}$$

as k varies over all nonnegative integers, and $(P_1, \cdots, P_k)$, $(t_1, \cdots, t_k)$ over all k-tuples of elements of $L^N(\mathbf{X})$, and of real numbers, respectively. The Campbell–Hausdorff formula shows that every product (3.22) is in $G^N(\mathbf{X})$. Conversely, the very definition of $G^N(\mathbf{X})$ shows that every element of $G^N(\mathbf{X})$ is a product of the form (3.22) (with just one factor). So $I = G^N(\mathbf{X})$.

If $P \in I$, then every solution of (3.19) which starts at P must stay in I. Since (3.19) is equivalent to (3.18) we conclude, in particular, that the solutions of (3.18) starting at 1 stay in $G^N(\mathbf{X})$. Therefore $\mathrm{Ser}_N(u) \in G^N(\mathbf{X})$ for every $u \in \mathcal{U}_m$. Since this is true for each N, it follows that $\mathrm{Ser}(u) \in \hat{G}(\mathbf{X})$. We summarize these facts in the following two statements.

PROPOSITION 3.1. *For every $u \in \mathcal{U}_m$, the series* Ser (u) *is an exponential Lie series.*

PROPOSITION 3.2. *For every N, equation* (3.18), *regarded as an evolution equation on $A^N(\mathbf{X})$, is a control system $\dot{S} = F_0(S) + \sum_{i=1}^m u_i F_i(S)$, where the vector fields F_i are given by* (3.20). *The Lie algebra $\Lambda^N(\mathbf{X})$ generated by the F_i is isomorphic to $L^N(\mathbf{X})$, and the maximal integral manifold of $\Lambda^N(\mathbf{X})$ through 1 is $G^N(\mathbf{X})$. In particular, the reachable set from the point 1 for the system* (3.18) *is a subset of $G^N(\mathbf{X})$.*

Actually, the general results of accessibility theory imply a stronger conclusion, which will be important to us. The "positive form" of Chow's theorem (Krener [3]) says that the reachable set from 1 for the system (3.18) actually has a nonempty interior in $G^N(\mathbf{X})$. Then the Sard theorem argument of [8] shows that the reachable set actually contains points which are "normally reachable", i.e. "reachable with full rank". We now make all this precise and, since the proofs are quite short, we present them in full.

Let $\mathbb{R}_+^k$ denote the set of all k-tuples of nonnegative real numbers. Fix a finite sequence

$$\Gamma = (\gamma^1, \cdots, \gamma^r) \tag{3.23}$$

where each γ^i is an element of $\mathbb{R}^m$. The choice of Γ is arbitrary, subject only to the requirement that

$$\mathrm{Aff}(\gamma^1, \cdots, \gamma^r) = \mathbb{R}^m \tag{3.24}$$

where "Aff" means "affine hull". (The affine hull of a set E is the set of all linear combinations of the elements of E with coefficients that add up to 1.) If $m = 1$, then a natural choice is to let $r = 2$, $\gamma^1 = a$, $\gamma^2 = -a$, a being any nonzero number. In general, it is clear that r must be at least $m + 1$.

Given Γ, we assign a meaning to γ^i for an arbitrary positive integer i by letting $i \to \gamma^i$ be periodic with period r, i.e. we let $\gamma^i = \gamma^j$ whenever $1 \leqq j \leqq r$, $i = pr + j$.

For any $\mathbf{t} \in \mathbb{R}_+^k$, $\mathbf{t} = (t_1, \cdots, t_k)$, define a control $\{\Gamma, t\}$ by letting

$$\begin{aligned}
\{\Gamma, \mathbf{t}\}(\tau) &= \gamma^1 \quad \text{if } 0 \leqq \tau < t_1, \\
\{\Gamma, \mathbf{t}\}(\tau) &= \gamma^2 \quad \text{if } t_1 < \tau < t_1 + t_2, \\
&\vdots \qquad\qquad \vdots \\
\{\Gamma, \mathbf{t}\}(\tau) &= \gamma^k \quad \text{if } t_1 + \cdots + t_{k-1} < \tau < t_1 + \cdots + t_k.
\end{aligned}$$

A control which is of the form $\{\Gamma, \mathbf{t}\}$ for some k and some $\mathbf{t} \in \mathbb{R}^k_+$ will be called a Γ-*control*, and we will use $\mathcal{U}_m(\Gamma)$ to denote the set of all Γ-controls. Also, for any given Γ, we let $X^i(\Gamma)$ denote, for each i, the element $X_0 + \sum_{j=1}^m \gamma^i_j X_j$ of $A(\mathbf{X})$ where $\gamma^i = (\gamma^i_1, \cdots, \gamma^i_m)$. We let $\nu^N_{k,\Gamma} : \mathbb{R}^k \to A^N(\mathbf{X})$ denote the map

$$\nu^N_{k,\Gamma}(t_1, \cdots, t_k) = e^{t_1 X^1(\Gamma)} e^{t_2 X^2(\Gamma)} \cdots e^{t_k X^k(\Gamma)}. \tag{3.25}$$

Then it is clear that

$$\nu^N_{k,\Gamma}(\mathbf{t}) = \mathrm{Ser}_N(\{\Gamma, \mathbf{t}\}) \quad \text{if } \mathbf{t} \in \mathbb{R}^k_+, \tag{3.26}$$

and that $\nu^N_{k,\Gamma}$ maps $\mathbb{R}^k$ into $G^N(\mathbf{X})$.

A Γ-control $\{\Gamma, \mathbf{t}^0\}$, $\mathbf{t}^0 \in \mathbb{R}^k_+$, will be called *N-normal* if the differential $d\nu^N_{k,\Gamma}$ of the map $\nu^N_{k,\Gamma}$ has rank at $\mathbf{t}^0$ equal to the dimension of $G^N(\mathbf{X})$, and all the components of $\mathbf{t}^0$ are strictly positive.

PROPOSITION 3.3. *For every $N > 0$, and every Γ such that* (3.24) *holds, there exists an N-normal Γ-control* $\{\Gamma, \mathbf{t}^0\}$.

Proof. Let L be the Lie algebra generated by the $X^i(\Gamma)$. Then $L \subseteq L^N(\mathbf{X})$. On the other hand, every m-tuple $(\alpha_1, \cdots, \alpha_m)$ can be expressed as an affine combination $\sum \lambda_i \gamma^i$, $\sum \lambda_i = 1$, and therefore $X_0 + \sum \alpha_j X_j$ equals $\sum \lambda_i X^i(\Gamma)$, which belongs to L. So $X_0 \in L$, and $X_0 + X_j \in L$ for $j > 0$, and therefore $L = L^N(\mathbf{X})$.

Now, for each k and each $\mathbf{t}^0 \in \mathbb{R}^k$, let $\rho(k, \mathbf{t}^0)$ denote the rank of the differential of $\nu^N_{k,\Gamma}$ at $\mathbf{t}^0$. Choose a k and a $\mathbf{t}^0 \in \mathbb{R}^k$ for which $\rho(k, \mathbf{t}^0)$ has the maximum possible value, and let this value be $\bar{\rho}$. Then $\nu^N_{k,\Gamma}$ has rank $\bar{\rho}$ in a neighborhood of $\mathbf{t}^0$ in $\mathbb{R}^k$, and so there is (by the implicit function theorem) a neighborhood U of $\mathbf{t}^0$ in $\mathbb{R}^k$ which is mapped by $\nu^N_{k,\Gamma}$ onto a $\bar{\rho}$-dimensional submanifold M of $G^N(\mathbf{X})$, and is such that $\nu^N_{k,\Gamma}$ has rank $\bar{\rho}$ throughout U.

For each i, let $V^i(\Gamma)$ denote the vector field $F^{X^i(\Gamma)}$ We claim that the $V^i(\Gamma)$ are tangent to M. Indeed, if this were not so, there would be an i such that, for some $S \in M$, $V^i(\Gamma)$ is not tangent to M at S. Since $i \to \gamma^i$, $i \to X^i(\Gamma)$, $i \to V^i(\Gamma)$ are periodic with period r, we may assume that $i > k$. Also, we have

$$S = \nu^N_{k,\Gamma}(\mathbf{t}^1)$$

for some $\mathbf{t}^1 \in U$. Let $\mathbf{0}$ denote a sequence of $i - k - 1$ zeros. Then

$$\tau \to \nu^N_{i,\Gamma}(\mathbf{t}^1, \mathbf{0}, \tau)$$

is the curve $\tau \to S\, e^{\tau X^i(\Gamma)}$, whose tangent vector at $\tau = 0$ is $V^i(\Gamma)(S)$. On the other hand, the map $\mathbf{t} \to \nu^N_{i,\Gamma}(\mathbf{t}, \mathbf{0}, 0)$ is precisely $\nu^N_{k,\Gamma}$, and so the image of the differential of $\nu^N_{i,\Gamma}$ at $(\mathbf{t}^1, \mathbf{0}, 0)$ contains that of the differential of $\nu^N_{k,\Gamma}$ at $\mathbf{t}^1$, which is the tangent space to M at S. Since $V^i(\Gamma)(S)$ is not tangent to M at S, it follows that the rank of $\nu^N_{i,\Gamma}$ at $(\mathbf{t}^1, \mathbf{0}, 0)$ is at least $\bar{\rho} + 1$, contradicting the fact that $\bar{\rho} = \rho(k, \mathbf{t}^0)$ was maximal.

So all the $V^i(\Gamma)$ are tangent to M, and then every vector field in the Lie algebra Λ generated by the $V^i(\Gamma)$ is also tangent to M. The isomorphism $P \to F^P$ and the identity $L = L^N(\mathbf{X})$ imply that $\Lambda = \Lambda^N(\mathbf{X})$. Since $G^N(\mathbf{X})$ is an integral manifold of $\Lambda^N(\mathbf{X})$, we can conclude that the tangent spaces to M and to $G^N(\mathbf{X})$ at S coincide. So $\bar{\rho} = \dim G^N(\mathbf{X})$.

Therefore the map $d\nu^N_{k,\Gamma}$ has rank $\dim G^N(\mathbf{X})$ at some $\mathbf{t} \in \mathbb{R}^k$. Since $\nu^N_{k,\Gamma}$ is analytic, it follows that there is a $\mathbf{t}^0$ with strictly positive coordinates, such that $d\nu^N_{k,\Gamma}$ has rank equal to $\dim G^N(\mathbf{X})$ at $\mathbf{t}^0$. Q.E.D.

4. Asymptotic properties. In this section we state and prove some elementary lemmas which show how the formal series $\mathrm{Ser}(u)$ gives rise to an asymptotic series

for a control problem. The results given here are explicitly or implicitly contained in earlier work, especially Krener's paper [15], the articles by Brockett, Gilbert and Lesiak–Krener on the convergence of the Volterra series (Brockett [16], Gilbert [17], Lesiak and Krener [18]), and Fliess's work (cf. especially Fliess [19], [20] and [21]). However, since the proofs are quite elementary, we prefer to make our paper self-contained by giving them in full, using precisely the symbolism that will be needed later on.

Let $\mathbf{f}=(f_0, \cdots, f_m)$ be an $(m+1)$-tuple of C^∞ vector fields on a manifold M. Consider the control system

$$\frac{dx}{dt}=f_0(x)+\sum_{i=1}^{m} u_i f_i(x) \tag{4.1}$$

where the state variable x takes values in M. Each vector field f_i is a first order differential operator on the space $C^\infty(M)$ of C^∞ real-valued functions on M and so, if $I=(i_1, \cdots, i_k)$ is a multi-index, with $0 \leqq i_j \leqq m$ for $j=1, \cdots, k$, then the product

$$f_I = f_{i_1} f_{i_2} \cdots f_{i_k} \tag{4.2}$$

is a partial differential operator on $C^\infty(M)$.

We use Ser $(u)(\mathbf{f})$ to denote the formal series

$$\sum_I \left(\int_0^{T(u)} u_I\right) f_I \tag{4.3}$$

obtained from Ser (u) by substituting for each indeterminate X_i the operator f_i. Then Ser $(u)(\mathbf{f})$ is a *formal series of partial differential operators*. If $\phi \in C^\infty(M)$, then we can apply the operators f_I to ϕ, and so we obtain the series

$$\text{Ser}\,(u)(\mathbf{f})(\phi)=\sum_I \left(\int_0^{T(u)} u_I\right)(f_I\phi) \tag{4.4}$$

which is a *formal series of C^∞ functions* on M.

We also want to consider the *truncated series*

$$Ser_N\,(u)(\mathbf{f})=\sum_{|I|\leqq N} \left(\int_0^{T(u)} u_I\right) f_I \tag{4.5}$$

and

$$\text{Ser}_N\,(u)(\mathbf{f})(\phi)=\sum_{|I|\leqq N} \left(\int_0^{T(u)} u_I\right)(f_I\phi). \tag{4.6}$$

These are no longer merely formal objects. The former is a true partial differential operator, and the latter a smooth function.

We now prove that Ser $(u)(\mathbf{f})(\phi)$ is an asymptotic series for the "propagation of ϕ along the trajectories of (4.1)". Precisely, let us use $t \to \pi(\mathbf{f}, u, t, x_0)$ to denote the trajectory of (4.1) corresponding to the control $u \in \mathcal{U}_m$ and the initial condition $x(0)=x_0$. Then, if $K \subseteq M$ is a compact set, and $A>0$, there is a time $\tau(K, A)>0$ such that $\pi(\mathbf{f}, u, t, x_0)$ is defined for all $x_0 \in K$ and all $t \in [0, T(u)]$, provided that $u \in \mathcal{U}_m$ is such that (i) $\|u(t)\| \leqq A$ for $0 \leqq t \leqq T(u)$, and (ii) $T(u) \leqq \tau(K, A)$. Moreover, if we let $\mathcal{U}_{m,A}$ denote the set of all $u \in \mathcal{U}_m$ such that $\|u(t)\| \leqq A$ for $0 \leqq t \leqq T(u)$, the set

$$K^{A,T}=\{\pi(\mathbf{f}, u, t, x_0): x_0 \in K, 0 \leqq t \leqq T(u) \leqq T, u \in \mathcal{U}_{m,A}\} \tag{4.7}$$

is compact, provided that $T \leqq \tau(K, A)$.

Let $\phi \in C^\infty(M)$. Then the t-derivative of $t \to \phi(\pi(\mathbf{f}, u, t, x_0))$ is $t \to [f_0\phi + \sum_{i=1}^m u_i(t)(f_i\phi)](\pi(\mathbf{f}, u, t, x_0))$. So

$$\phi(\pi(\mathbf{f}, u, t, x_0)) = \phi(x_0) + \sum_{i=0}^m \int_0^t u_i(s)(f_i\phi)(\pi(\mathbf{f}, u, s, x_0))\, ds. \tag{4.8}$$

Also

$$(f_i\phi)(\pi(\mathbf{f}, u, s, x_0)) = (f_i\phi)(x_0) + \sum_{j=0}^m \int_0^s u_j(\sigma)(f_jf_i\phi)(\pi(\mathbf{f}, u, \sigma, x_0))\, d\sigma, \tag{4.9}$$

so that

$$\begin{aligned}\phi(\pi(\mathbf{f}, u, t, x_0)) = \phi(x_0) &+ \sum_{i=0}^m \left[\int_0^t u_i(s)\, ds\right](f_i\phi)(x_0) \\ &+ \sum_{i,j} \int_0^t \int_0^s u_i(s)u_j(\sigma)(f_jf_i\phi)(\pi(\mathbf{f}, u, \sigma, x_0))\, d\sigma\, ds.\end{aligned} \tag{4.10}$$

Iterating this procedure yields

$$\phi(\pi(\mathbf{f}, u, t, x_0)) = \mathrm{Ser}_N\,(u_t)(\mathbf{f})(\phi)(x_0) + R_N(u_t, \mathbf{f}, \phi)(x_0) \tag{4.11}$$

where the remainder $R_N(u_t, \mathbf{f}, \phi)$ is given by

$$R_N(u_t, \mathbf{f}, \phi)(x_0) = \sum_{|I|=N+1} R^I(u_t, \mathbf{f}, \phi)(x_0) \tag{4.12}$$

and where, for $I = (i_1, \cdots, i_k)$, we let

$$\begin{aligned}&R^I(u_t, \mathbf{f}, \phi)(x_0) \\ &= \int_0^t \int_0^{\tau_k} \cdots \int_0^{\tau_3} \int_0^{\tau_2} u_{i_k}(\tau_k) \cdots u_{i_2}(\tau_2)u_{i_1}(\tau_1)(f_I\phi)(\pi(\mathbf{f}, u, \tau_1, x_0))\, d\tau_1 \cdots d\tau_k.\end{aligned} \tag{4.13}$$

If we now choose a compact set K, an $A > 0$, and a T not greater than $\tau(K, A)$, then the points $\pi(\mathbf{f}, u, t, x_0)$ belong to the compact set $K^{A,T}$ for all u in $\mathcal{U}_{m,A}$ such that $T(u) \leqq T$, and all $t \in [0, T(u)]$. For each N, choose a constant C_N such that

$$|(f_I\phi)(x)| \leqq C_N \quad \text{for all } x \in K^{A,T}, \quad |I| = N+1.$$

Then, if $B = \max(1, A)$, we have

$$|R^I(u_t, \mathbf{f}, \phi)(x_0)| \leqq \frac{t^{N+1}C_NB^{N+1}}{(N+1)!} \tag{4.14}$$

as long as $u \in \mathcal{U}_{m,A}$, $T(u) \leqq T$, $x_0 \in K$, $0 \leqq t \leqq T(u)$.

Therefore, we have proved

PROPOSITION 4.1. *Consider a system* (4.1), *with* $f_0, \cdots, f_m$ *smooth vector fields on a manifold* M. *Let* $K \subseteq M$ *be compact,* $A > 0$, *and* $\phi \in C^\infty(M)$. *Then for each positive integer* N *there exists a constant* D_N, *depending on* ϕ, A, K, N, *but not on* u, *such that*

$$|\phi(\pi(\mathbf{f}, u, T(u), x_0)) - \mathrm{Ser}_N\,(u)(\mathbf{f})(\phi)(x_0)| \leqq D_N T(u)^{N+1} \tag{4.15}$$

for all $x_0 \in K$, *and all* $u \in \mathcal{U}_{m,A}$ *such that* $T(u) \leqq \tau(K, A)$.

If the vector fields f_i are real analytic, and the function ϕ is also analytic, then Proposition 4.1 can be strengthened. One can actually prove that the series $\mathrm{Ser}\,(u)(\mathbf{f})(\phi)$ converges.

The key step in the proof of this is the following:

LEMMA 4.2. *Let $f_0, \cdots, f_m$ be real analytic vector fields on an analytic manifold M, and let $\phi: M \to \mathbb{R}$ be analytic. Let $K \subseteq M$ be compact. Then there exists a constant $C > 0$ such that the estimate*

$$|f_{i_1} \cdots f_{i_r} \phi(x)| \leqq r! C^r \tag{4.16}$$

holds for all $x \in K$, all integers $r \geqq 0$, and all choices of the indices $i_1, \cdots, i_r$, $0 \leqq i_j \leqq m$.

Proof. The conclusion clearly holds for a finite union of compact sets if it holds for each one of them.

So it suffices to prove that every point $p \in M$ has a neighborhood on which (4.16) holds. We may then assume that M is an open subset of $\mathbb{R}^n$, that $p = 0 \in M$, and that the f_i, ϕ actually extend to complex analytic maps on some complex polydisc

$$D(n, \alpha) = \{z : z = (z_1, \cdots, z_n) \in \mathbb{C}^n, |z_i| < \alpha, i = 1, \cdots, n\}. \tag{4.17}$$

Now consider the control system

$$\frac{dz}{dt} = \sum_{i=0}^{m} v_i \tilde{f}_i(z), \qquad v_i \in \mathbb{C}, \tag{4.18}$$

where the $\tilde{f}_i$ are the analytic extensions to $D(n, \alpha)$ of the f_i. The system (4.18) can also be expressed as a real control system in $D(n, \alpha)$, regarded as a subset of $\mathbb{R}^{2n}$. Hence all the usual properties of such systems apply. In particular, there exists a time $T > 0$ such that, whenever $v = (v_0, \cdots, v_m)$ is a control with values in the polydisc $D(m+1, 1)$, and defined on an interval $[0, t]$, $t \leqq T$, then the unique trajectory of v which goes through any $q \in D(n, \alpha/3)$ at time 0 is defined on the whole interval $[0, t]$, and stays in $D(n, 2\alpha/3)$.

We now fix T, and consider a particular class of controls. Ley $I = (i_1, \cdots, i_r)$ be a multiindex of arbitrary length, with $0 \leqq i_j \leqq m$ for $j = 1, \cdots, r$. Define a control $v^{I,\mathbf{z}}$, depending on a parameter $\mathbf{z} = (z_1, \cdots, z_r) \in D(r, 1)$, and with domain $[0, T]$, by

$$v^{I,\mathbf{z}}(t) = z_i e_{i_r} \quad \text{for } \frac{i-1}{r} T \leqq t < \frac{iT}{r},$$

where $(e_0, \cdots, e_m)$ is the canonical basis of $\mathbb{C}^{m+1}$. Let $\xi_q^{I,\mathbf{z}}$ denote the trajectory of $v^{I,\mathbf{z}}$ which goes through q at time 0. Then $\xi_q^{I,\mathbf{z}}$ is defined on $[0, T]$, and takes values in $D(n, 2\alpha/3)$, for all $q \in D(n, \alpha/3)$. Now, for each complex vector field $g: D(n, \alpha) \to \mathbb{C}^n$, let $\{\Phi^g(t)\}$ denote the flow of g, so that $\Phi^g(0)(q) = q$, and

$$\frac{d}{dt}(\Phi^g(t)(q)) = g(\Phi^g(t)(q)), \tag{4.19}$$

for all q, t.

Then

$$\xi_q^{I,\mathbf{z}}\left(\frac{T}{r}\right) = \Phi^{z_1 \tilde{f}_{i_1}}\left(\frac{T}{r}\right)(q),$$

$$\xi_q^{I,\mathbf{z}}\left(\frac{2T}{r}\right) = \Phi^{z_2 \tilde{f}_{i_2}}\left(\frac{T}{r}\right) \Phi^{z_1 \tilde{f}_{i_1}}\left(\frac{T}{r}\right)(q),$$

$$\vdots$$

$$\xi_q^{I,\mathbf{z}}(T) = \Phi^{z_r \tilde{f}_{i_r}}\left(\frac{T}{r}\right) \cdots \Phi^{z_1 \tilde{f}_{i_1}}\left(\frac{T}{r}\right)(q).$$

Clearly, if g is a complex analytic vector field, the map $(z, q) \to \Phi^{zg}(t)(q)$ is analytic in z and q for every t. (For instance, one can obtain the trajectories by successive approximations, and it is clear that all the successive approximations are analytic in z, q. Or one can consider the vector field $h = (0, zg)$ in $\mathbb{C}^{n+1}$—i.e. one can regard z as a new variable—and then it is clear that $(z, \Phi^{zg}(t)(q)) = \Phi^{h}(t)(z, q)$.) So $\xi_q^{I,\mathbf{z}}(iT/r)$ is analytic in q and $\mathbf{z}$, for $q \in D(n, \alpha/3)$, $\mathbf{z} \in D(r, 1)$. Moreover, $\xi_q^{I,\mathbf{z}}(iT/r)$ belongs to $D(n, 2\alpha/3)$ for all such $\mathbf{z}, q$.

If ψ is a complex-valued analytic function on $D(n, \alpha)$, then we can consider the function $\psi_I^* : D(n, \alpha/3) \times D(r, 1) \to \mathbb{C}$ defined by $\psi_I^*(q, \mathbf{z}) = \psi(\xi_q^{I,\mathbf{z}}(T))$. Then a simple computation shows that

$$\frac{\partial^r \psi_I^*}{\partial z_r \cdots \partial z_2 \, \partial z_1}(q, \mathbf{0}) = (\tilde{f}_{i_1} \cdots \tilde{f}_{i_r} \psi)(q) \times \left(\frac{T}{r}\right)^r.$$

On the other hand, since $\psi_I^*(q, \mathbf{z})$ is analytic for $|z_i| \leqq 1$, we have bounds

$$\left|\frac{\partial^r \psi_I^*}{\partial z_r \cdots \partial z_1}(q, \mathbf{0})\right| \leqq \max\left\{|\psi(q')| : q' \in D\left(n, \frac{2\alpha}{3}\right)\right\}.$$

In particular, we may let ψ be the analytic extension of ϕ to $D(n, \alpha)$. Then we find that, if $C' = \max\{|\psi(q')| : q' \in D(n, 2\alpha/3)\}$, then

$$|(\tilde{f}_{i_1} \cdots \tilde{f}_{i_r} \psi)(q)| \leqq C' r^r T^{-r}$$

for all $q \in D(n, \alpha/3)$. In particular, let $K = D(n, \alpha/3) \cap \mathbb{R}^n$. Then we have proved that

$$|(f_{i_1} \cdots f_{i_r} \phi)(q)| \leqq C' r^r T^{-r} \tag{4.20}$$

for $q \in K$. Note that (4.21) holds for all r, $(i_1, \cdots, i_r)$, with a fixed constant C'. By Stirling's formula, there is a constant C'', such that

$$r^r \leqq C'' e^r r! \quad \text{for all } r.$$

But then

$$|(f_{i_1} \cdots f_{i_r} \phi)(q)| \leqq C^r r!$$

for $q \in K$, if $C = C'C''(eT^{-1} + 1)$. Q.E.D.

Using Lemma 4.2, we can prove the convergence of the exponential Lie series.

PROPOSITION 4.3. *Consider a system* (4.1), *and assume that*

(i) *M is a real analytic manifold and $f_0, \cdots, f_m$ are real analytic vector fields,*

(ii) *$A > 0$,*

(iii) *$K \subseteq M$ is compact,*

(iv) *$\phi : M \to \mathbb{R}$ is real analytic.*

Then there exists a time $T > 0$ such that, for every $x \in K$ and every $u \in \mathcal{U}_{m,A}$ for which $T(u) \leqq T$: (a) *the curve $t \to \pi(\mathbf{f}, u, t, x)$ is defined for $0 \leqq t \leqq T(u)$, and* (b) *the series* $\mathrm{Ser}\,(u_t)(\mathbf{f})(\phi)(x)$ *converges to $\phi(\pi(\mathbf{f}, u, t, x))$. Moreover, the convergence is uniform as long as $x \in K$, $u \in \mathcal{U}_{m,A}$, $T(u) \leqq T$.*

Proof. First choose a $T' > 0$ and a compact set $K' \supseteq K$, such that $\pi(\mathbf{f}, u, t, x)$ is defined and belongs to K' for $x \in K$, $0 \leqq t \leqq T(u)$, whenever $u \in \mathcal{U}_{m,A}$, $T(u) \subseteq T'$.

The difference between $\phi(\pi(\mathbf{f}, u, t, x))$ and the Nth partial sum of the series $\mathrm{Ser}\,(u_t)(\mathbf{f})(\phi)(x)$ is given by (4.11). Using (4.11) together with (4.12), (4.13), and observing that the number of multiindices I for which $|I| = N + 1$ is $(m+1)^{N+1}$, we find

$$|\phi(\pi(\mathbf{f}, u, t, x)) - \mathrm{Ser}_N\,(u_t)(\mathbf{f})(\phi)(x)|$$
$$\leqq \frac{(At)^{N+1}}{(N+1)!}(m+1)^{N+1} \sup\{|f_I \phi)(\pi(\mathbf{f}, v, s, x))|\}$$

where the sup ranges over all $x \in K$, $v \in \mathcal{U}_{m,A}$ such that $T(v) \leqq T'$, s such that $0 \leqq s \leqq T(v)$, and I such that $|I| = N+1$.

By Lemma 4.2, there is a constant $C > 0$ such that $|(f_I \phi)(p)| \leqq C^{N+1}(N+1)!$ for all $p \in K'$, and I such that $|I| = N+1$. Since $\pi(\mathbf{f}, v, t, x) \in K'$ when $v \in \mathcal{U}_{m,A}$, $0 \leqq t \leqq T(v) \leqq T'$, $x \in K$, we conclude that

$$|\phi(\pi(\mathbf{f}, u, t, x)) - \mathrm{Ser}_N(u)(\mathbf{f})(\phi)(x)| \leqq [\mathrm{CAt}(m+1)]^{N+1}. \tag{4.21}$$

Now choose T such that $\mathrm{CAT}(m+1) < 1$. Then (4.21) shows that $\mathrm{Ser}_N(u)(\mathbf{f})(\phi)(x)$ converges to $\phi(\pi(\mathbf{f}, u, t, x))$, and that the convergence is uniform as long as $x \in K$, $0 \leqq t \leqq T(u) \leqq T$, $u \in \mathcal{U}_{m,A}$. Q.E.D.

5. Proof of Theorem 2.1. The result to be proved is local, so we may assume, without loss of generality, that $M = \mathbb{R}^n$ and that $x_0 = 0$. Also, if we let $\tilde{f}_0 = f_0 + \bar{u} f_1$, then the system

$$\frac{dx}{dt} = \tilde{f}_0(x) + u f_1(x), \qquad |u| \leqq \alpha, \tag{5.1}$$

has the property that $\tilde{f}_0(0) = 0$, and that every trajectory of (5.1) is also a trajectory of (2.1), provided that we choose α in such a way that

$$0 < \alpha \leqq A - |\bar{u}|. \tag{5.2}$$

On the other hand, it is clear that (5.1) also satisfies the HLCC at 0. If we prove that (5.1) is STLC from 0, it will follow that (2.1) is STLC from 0 as well. Therefore we may assume that our original system satisfies the HLCC with $\bar{u} = 0$. Finally, we may replace f_1 by αf_1, and so the inequality constraint on u becomes $|u| \leqq 1$.

So, from now on we assume that f_0, f_1 are C^∞ vector fields in $\mathbb{R}^n$, that $f_0(0) = 0$, that $\dim \mathrm{Lie}\,(f_0, f_1)(0) = n$, and that

$$\mathcal{S}^k(f_0, f_1)(0) = \mathcal{S}^{k+1}(f_0, f_1)(0) \quad \text{whenever } k \text{ is odd.} \tag{5.3}$$

Our goal is to show that, for arbitrarily small $t > 0$, the reachable set $\mathrm{Reach}\,(0, t)$ for the system

$$\frac{dx}{dt} = f_0(x) + u f_1(x), \qquad |u| \leqq 1, \tag{5.4}$$

contains a neighborhood of 0.

Throughout this section, we let $\mathbf{X} = (X_0, X_1)$. For each noncommutative polynomial μ in the indeterminates X_0, X_1, we will use either one of the notations $\mu(f)$, $\tilde{\mu}$ to denote the result of substituting f_0 for X_0 and f_1 for X_1 in μ. (Then $\tilde{\mu}$ is a partial differential operator and, in the particular case when μ is a Lie monomial, $\tilde{\mu}$ is a vector field.)

Whenever μ belongs to one of the algebras $A(\mathbf{X})$, $A^N(\mathbf{X})$, we can talk about the *degree* $\delta(\mu)$ of μ (which is the degree of the monomial of highest degree that appears in μ) and also of the *degree* of μ in X_1, which we denote by $\delta_1(\mu)$ (so that, e.g., if $\mu = X_0^2 X_1^3 + X_1 X_0 X_1 X_0^4$, then $\delta(\mu) = 7$, $\delta_1(\mu) = 3$). Similarly, we can talk about the *order* $\omega(\mu)$ of μ (defined earlier) and the *order in* X_1, which we denote by $\omega_1(\mu)$. We let $\omega(0) = +\infty$, and we let $\omega_1(\mu) = +\infty$ if μ does not contain X_1. Then it is clear that

$$\omega(\mu_1 \mu_2) \geqq \omega(\mu_1) + \omega(\mu_2) \tag{5.5}$$

and that

$$\omega_1(\mu_1 \mu_2) \geqq \omega_1(\mu_1) + \omega_1(\mu_2) \tag{5.6}$$

for all μ_1, μ_2.

Let us choose, once and for all, a fixed sequence $\mu_1, \cdots, \mu_n$ of Lie monomials in X_0, X_1, such that the vectors $\tilde{\mu}_1(0), \cdots, \tilde{\mu}_n(0)$ form a basis of $\mathbb{R}^n$. (The existence of the μ_i follows from (HLCC 2).) Let D be the largest of the degrees $\delta(\mu_i)$, and D_1 the largest of the $\delta_1(\mu_i)$.

The union of the increasing sequence of subspaces $\mathscr{P}^j(f_0, f_1)(0)$ is $\mathbb{R}^n$, and so there is an integer E such that $\mathscr{P}^E(f_0, f_1)(0) = \mathbb{R}^n$.

For each i, j, let Mon (i, j) denote the set of all Lie monomials η in X_0, X_1, whose total degree $\delta(\eta)$ is equal to i, and whose degree in X_1 is j. (Precisely, we define

$$\text{Mon } (1, 0) = \{X_0\}, \qquad \text{Mon } (1, 1) = \{X_1\},$$

$$\text{Mon } (i, j) = \bigcup_{\substack{i_1+i_2=i \\ j_1+j_2=j}} \{[A, B]: A \in \text{Mon } (i_1, j_1), B \in \text{Mon } (i_2, j_2)\}.)$$

Let $\{\eta_{ijk}: 1 \leqq k \leqq \bar{k}(i, j)\}$ be, for each i, j, a fixed sequence of elements of Mon (i, j) which is a basis of the linear span of Mon (i, j).

In particular, for each $j \leqq E$, we can choose a finite set $a(j)$ of pairs i, k such that $\mathscr{P}^j(f_0, f_1)(0)$ is spanned by $\mathscr{P}^{j-1}(f_0, f_1)(0)$ together with the $\tilde{\eta}_{ijk}(0)$, $(i, k) \in a(j)$. So, if we let F be the largest degree of all the η_{ijk} for all the $j \leqq E$, and all $(i, k) \in a(j)$, we can conclude that, for every $j \geqq 0$, every element of $\mathscr{P}^j(f_0, f_1)(0)$ is a linear combination of $\tilde{\eta}_{isk}(0)$ with $i \leqq F$, $s \leqq j$, $s \leqq E$.

We now choose Q to be an integer such that $Q > F$, and then choose an N such that

$$D + QD_1 < N.$$

Then every element $Z \in L^N(\mathbf{X})$ can be written uniquely in the form

$$Z = \sum_{i \leqq N} \sum_{j \leqq i} \sum_{k} Z_{ijk} \eta_{ijk}. \tag{5.7}$$

Let $\Gamma = (\gamma^1, \gamma^2)$, where $\gamma^1 = 1$, $\gamma^2 = -1$.

Let us call a point $S \in G^N(\mathbf{X})$ "good" if there exists an integer $p > 0$ and a $\mathbf{t}^0 \in \mathbb{R}^p$, with strictly positive coordinates, such that the Γ-control $\{\Gamma, \mathbf{t}^0\}$ is N-normal and that

$$\text{Ser}_N(\{\Gamma, \mathbf{t}^0\}) = S.$$

Proposition 3.3 implies that a good $S \in G^N(\mathbf{X})$ exists. On the other hand, it is clear that, if S is good, then any point of $G^N(\mathbf{X})$ obtained by right multiplication of S by $\text{Ser}_N(\{\Gamma, \mathbf{t}\})$, for some other Γ-control $\{\Gamma, \mathbf{t}\}$, is also good (even if $\{\Gamma, \mathbf{t}\}$ is not normal).

Let $\lambda: A^N(\mathbf{X}) \to A^N(\mathbf{X})$ be the automorphism which sends X_0 to X_0 and X_1 to $-X_1$. Then the elements of $A^N(\mathbf{X})$ that are invariant under λ are those that are *even* in X_1, i.e. those that are linear combinations of monomials that involve X_1 an even number of times. Similarly, the elements S that satisfy $\lambda(S) = -S$ are precisely those that are odd in X_1. Any $S \in A^N(\mathbf{X})$ can be written in a unique way as a sum $S_{\text{even}} + S_{\text{odd}}$ of an even element and an odd one.

Now let $S_1 \in G^N(\mathbf{X})$ be a good element. It is easy to see that, if $S_1 = \text{Ser}_N(\{\Gamma, \mathbf{t}\})$, then $\lambda(S_1) = \text{Ser}_N(\{\Gamma, (0, \mathbf{t})\})$. So

$$S_2 = S_1 \lambda(S_1)$$

is also good. Now we can write

$$S_1 = e^{Z^1}$$

where Z^1 is a Lie element. Then $Z^1 = Z^1_{\text{even}} + Z^1_{\text{odd}}$, and Z^1_{even}, Z^1_{odd} are Lie elements so that, in particular,

$$\omega(Z^1_{\text{odd}}) \geqq 1.$$

Then the Campbell–Hausdorff formula gives

$$S_2 = e^{Z^1_{\text{even}}+Z^1_{\text{odd}}} e^{Z^1_{\text{even}}-Z^1_{\text{odd}}}$$
$$= e^{2Z^1_{\text{even}}+R^1}$$

where

$$R^1 = \tfrac{1}{2}[Z^1_{\text{even}} + Z^1_{\text{odd}}, Z^1_{\text{even}} - Z^1_{\text{odd}}] + \cdots$$

so that $\omega(R^1) \geqq 2$.

Now, S_2 can also be written as

$$S_2 = e^{Z^2}$$

where Z^2 has a decomposition

$$Z^2 = Z^2_{\text{even}} + Z^2_{\text{odd}}.$$

But

$$Z^2 = 2Z^1_{\text{even}} + R^1. \tag{5.8}$$

Comparison of these two expressions shows that the odd part Z^2_{odd} is contained in R^1, and so

$$\omega(Z^2_{\text{odd}}) \geqq 2.$$

This procedure can be iterated by induction, and one obtains good elements $S_j \in G^N(\mathbf{X})$ which have expressions

$$S_j = e^{Z^j}, \qquad Z^j = Z^j_{\text{even}} + Z^j_{\text{odd}},$$

and satisfy

$$\omega(Z^j_{\text{odd}}) \geqq j.$$

If we now take $j = N+1$, $S = S_{N+1}$, $Z = Z^{N+1}$, we find that $S \in G^N(\mathbf{X})$, that S is good, and that $S = e^Z$ with Z even.

From now on we keep S, Z fixed, and we assume that S is good, $S = e^Z$, and that Z is even. We choose an integer $p > 0$ and a $\mathbf{t}^0 \in \mathbb{R}^p$, with strictly positive coordinates, such that

$$\text{Ser}_N(\{\Gamma, \mathbf{t}^0\}) = S$$

and that $\{\Gamma, \mathbf{t}^0\}$ is N-normal.

We have an expression

$$Z = \sum_{i \leqq N} \sum_{j \leqq i}^{\text{e}} \sum_k Z_{ijk}\eta_{ijk},$$

where "$\sum^{\text{e}}$" means that the sum only runs through even values of j.

Because $\{\Gamma, \mathbf{t}^0\}$ is normal, there exists a neighborhood U of S in $G^N(\mathbf{X})$ and a smooth map

$$\psi : U \to \mathbb{R}^p_+$$

such that

$$\text{Ser}_N(\{\Gamma, \psi(S')\}) = S' \quad \text{for } S' \in U,$$

and that $\psi(S) = \mathbf{t}^0$.

We now use condition (HLCC 3). For each $i \leqq N$, $j \leqq i$, $k \leqq \bar{k}(i,j)$, j even, the vector $\tilde{\eta}_{ijk}(0)$ belongs to $\mathscr{S}^j(f_0, f_1)(0)$. Since j is even, it follows that $\tilde{\eta}_{ijk}(0) \in$

$\mathscr{S}^{j-1}(f_0, f_1)(0)$. So we can find coefficients ξ_{ijk}^{abc} such that

$$\tilde{\eta}_{ijk}(0) + \sum_{abc} \xi_{ijk}^{abc} \tilde{\eta}_{abc}(0) = 0 \tag{5.9}$$

where the sum runs over $a \leqq F$, $b < j$, $b \leqq E$, $c \leqq \bar{k}(a, b)$.

If $\boldsymbol{\varepsilon} = \{\varepsilon_{ijk}^{abc}\}$ is a family of real numbers, for $a \leqq F$, $b < j$, $b \leqq E$, $c \leqq \bar{k}(a, b)$, $i \leqq N$, $j \leqq i$, $k \leqq \bar{k}(i, j)$, and if $\boldsymbol{\rho} = \{\rho_i : i = 1, \cdots, n\}$ is another finite sequence of numbers, then we can define

$$Z'(\boldsymbol{\varepsilon}, \boldsymbol{\rho}) = Z + \sum_{ijk} \sum_{abc} \xi_{ijk}^{abc} \varepsilon_{ijk}^{abc} \eta_{abc} + \sum_i \rho_i \mu_i,$$

$$S'(\boldsymbol{\varepsilon}, \boldsymbol{\rho}) = e^{Z'(\boldsymbol{\varepsilon}, \boldsymbol{\rho})}.$$

Then

$$Z'(0, 0) = Z \quad \text{and} \quad S'(0, 0) = S.$$

Let $|\boldsymbol{\varepsilon}| = \max\{|\varepsilon_{ijk}^{abc}|\}$, $|\boldsymbol{\rho}| = \max\{|\rho_i|\}$. Then there exists a constant $\alpha > 0$ such that

$$S'(\boldsymbol{\varepsilon}, \boldsymbol{\rho}) \in U \quad \text{whenever } |\boldsymbol{\varepsilon}| \leqq \alpha, \quad |\boldsymbol{\rho}| \leqq \alpha.$$

If $|\boldsymbol{\varepsilon}| \leqq \alpha$, $|\boldsymbol{\rho}| \leqq \alpha$, define a control $u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}$ by

$$u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}} = \{\Gamma, \psi(S'(\boldsymbol{\varepsilon}, \boldsymbol{\rho}))\},$$

so that

$$\mathrm{Ser}_N\,(u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}) = S'(\boldsymbol{\varepsilon}, \boldsymbol{\rho}).$$

Let $0 < \delta \leqq 1$, and define $u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}^{\delta}$ to be the control $\delta^Q\{\Gamma, \delta\psi(\boldsymbol{\varepsilon}, \boldsymbol{\rho})\}$. (That is, the times $t_1(\boldsymbol{\varepsilon}, \boldsymbol{\rho}), \cdots, t_p(\boldsymbol{\varepsilon}, \boldsymbol{\rho})$ are multiplied by δ, and then the control itself is multiplied by δ^Q.) It is clear that

$$u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}^{\delta}(t) = \delta^Q u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}(\delta^{-1}t)$$

for $0 \leqq t \leqq \delta T(u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}) = T(u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}^{\delta})$.

Now, multiplying each of the times t_i by δ is equivalent to keeping the times unchanged and multiplying both X_0 and X_1 by δ. On the other hand, multiplication of a control by δ^Q is equivalent to multiplication of X_1 by δ^Q. Therefore, the series for $u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}^{\delta}$ is

$$\mathrm{Ser}_N\,(u_{\boldsymbol{\varepsilon}, \boldsymbol{\rho}}^{\delta}) = e^{Z_\delta(\boldsymbol{\varepsilon}, \boldsymbol{\rho})}$$

where

$$Z_\delta(\boldsymbol{\varepsilon}, \boldsymbol{\rho}) = Z_\delta^1(\boldsymbol{\varepsilon}) + Z_\delta^2(\boldsymbol{\rho}),$$

$$Z_\delta^1(\boldsymbol{\varepsilon}) = \sum_{ijk} \left[Z_{ijk} \delta^{i+Qj} \eta_{ijk} + \sum_{abc} \delta^{a+Qb} \xi_{ijk}^{abc} \varepsilon_{ijk}^{abc} \eta_{abc} \right],$$

$$Z_\delta^2(\boldsymbol{\rho}) = \sum_{i=1}^{n} \delta^{d(i)} \rho_i \mu_i,$$

$$d(i) = \delta(\mu_i) + Q\delta_1(\mu_i).$$

For each ijk, the sum $\sum_{abc}$ runs over indices such that $a \leqq F$, $b < j$. Then, for each abc which occurs in the sum,

$$i + Qj - (a + Qb) \geqq Q + i - a \geqq Q - F \geqq 1.$$

Hence, if we define

$$\boldsymbol{\varepsilon}(\delta) = \{\varepsilon_{ijk}^{abc}(\delta)\}, \qquad \varepsilon_{ijk}^{abc}(\delta) = \delta^{i+Qj-a-Qb},$$

we have

$$\varepsilon_{ijk}^{abc}(\delta) \to 0 \quad \text{as } \delta \to 0$$

for every a, b, c, i, j, k, and

$$Z_\delta^1(\boldsymbol{\varepsilon}(\delta)) = \sum_{ijk} \delta^{i+Qj}\left[\eta_{ijk} + \sum_{abc} \xi_{ijk}^{abc}\eta_{abc}\right]. \tag{5.10}$$

On the other hand, the numbers $d(i)$ all satisfy

$$d(i) < N.$$

So, if we define

$$\boldsymbol{\rho}(\delta, \mathbf{y}) = (\rho_1(\delta, \mathbf{y}), \cdots, \rho_n(\delta, \mathbf{y})), \qquad \rho_i(\delta, \mathbf{y}) = \delta^{N-d(i)}y_i$$

for every $\mathbf{y} = (y_1, \cdots, y_n)$ in the closed unit ball $\mathcal{B}_n$ of $\mathbb{R}^n$, we find that

$$\boldsymbol{\rho}(\delta, \mathbf{y}) \to 0 \quad \text{as } \delta \to 0,$$

the convergence being uniform in $\mathbf{y}$, for $\mathbf{y} \in \mathcal{B}_n$, and that

$$Z_\delta^2(\boldsymbol{\rho}(\delta, \mathbf{y})) = \delta^N \sum_{i=1}^n y_i\mu_i.$$

Now let $\bar{\delta}$ be so small that $|\boldsymbol{\varepsilon}(\delta)| \leqq \alpha$ and that $|\boldsymbol{\rho}(\delta, \mathbf{y})| \leqq \alpha$ for $0 < \delta \leqq \bar{\delta}$, $\mathbf{y} \in \mathcal{B}_n$. Then the control

$$v_{\delta,\mathbf{y}} = u^{\delta}_{\boldsymbol{\varepsilon}(\delta),\boldsymbol{\rho}(\delta,\mathbf{y})}$$

is well defined for $0 < \delta \leqq \bar{\delta}$, $\mathbf{y} \in \mathcal{B}_n$. Moreover, we have

$$\mathrm{Ser}_N(v_{\delta,\mathbf{y}}) = \exp\left(\delta^N\left(\sum_{i=1}^n y_i\mu_i\right) + Z_\delta^1(\boldsymbol{\varepsilon}(\delta))\right) \tag{5.11}$$

where $Z_\sigma^1(\boldsymbol{\varepsilon}(\delta))$ is given by (5.10).

If we expand the right-hand side of (5.11), it turns out to be equal to $1 + \delta^N(\sum_{i=1}^n y_i\mu_i)$ plus a sum of terms each of which is either (I) a power of $Z_\delta^1(\boldsymbol{\varepsilon}(\delta))$ or (II) a product of factors $\delta^N(\sum_{i=1}^n y_i\mu_i)$, $Z_\delta^1(\boldsymbol{\varepsilon}, \delta)$, containing at least two factors, and at least one factor of the first type.

Now, the construction of the coefficients ξ_{ijk}^{abc} shows that $Z_\delta^1(\boldsymbol{\varepsilon}, \delta)(\mathbf{f})(0)$ vanishes, and therefore every power $[Z_\delta^1(\boldsymbol{\varepsilon}, \delta)]^m(\mathbf{f})(0)$ vanishes as well. Also, every type II product contains at least a factor δ^{N+1} in it. So, we have

$$\mathrm{Ser}_N(v_{\delta,\mathbf{y}}) = 1 + \delta^N\left(\sum_{i=1}^n y_i\mu_i\right) + Y_1 + \delta^{N+1}Y_2$$

where Y_1 and Y_2 are sums which may depend on δ, $\mathbf{y}$, but which remain bounded as $\delta \to 0$, and satisfy $Y_1(\mathbf{f})(0) = 0$.

Using Proposition 4.1, we conclude that, if ϕ is a C^∞ function on a neighborhood of 0, and if we write

$$\pi^*(\delta, \mathbf{y}) = \pi(\mathbf{f}, v_{\delta,\mathbf{y}}, T(v_{\delta,\mathbf{y}}), 0),$$

then

$$\phi(\pi^*(\delta, \mathbf{y})) = \phi(0) + \delta^N\Big(\sum_{i=1}^{n} y_i \tilde{\mu}_i(0)\phi\Big) + O(\delta^{N+1}), \tag{5.12}$$

the $O(\delta^{N+1})$ being uniform in $\mathbf{y}$. (Note that $T(v_{\delta,\mathbf{y}}) = O(\delta)$.)

We now apply (5.12) with the coordinate functions x_j playing the role of ϕ. Clearly $\tilde{\mu}_i(0)x_j$ is the jth component of the vector $\tilde{\mu}_i(0)$, and so we find

$$\pi^*(\delta, \mathbf{y}) = \delta^N \sum_{i=1}^{n} y_i \tilde{\mu}_i(0) + O(\delta^{N+1}).$$

Now let $g: \mathbb{R}^n \to \mathbb{R}^n$ be the inverse of the linear map $\mathbf{y} \to \sum y_i \tilde{\mu}_i(0)$, and define, for $\mathbf{x} \in g^{-1}(\mathcal{B}_n)$,

$$\sigma(t, \mathbf{x}) = \pi^*(t^{1/N}, g(\mathbf{x})).$$

Then

$$\sigma(t, \mathbf{x}) = t\mathbf{x} + o(t).$$

Let B be a compact ball centered at the origin, and such that $g(B) \subseteq \mathcal{B}_n$.

Then the maps

$$\beta_t: \mathbf{x} \to \frac{1}{t}\sigma(t, \mathbf{x}), \qquad \mathbf{x} \in B,$$

converge uniformly to the identity map of B as $t \to 0$. Therefore $\beta_t(B)$ contains a neighborhood of 0, if t is small enough. Hence the set

$$C_t = \{\sigma(t, \mathbf{x}): \mathbf{x} \in B\}$$

contains a neighborhood of zero, if t is small enough.

Now, if $\mathbf{z} \in C_t$, then $\mathbf{z} = \pi^*(\delta, \mathbf{y})$ for some $\mathbf{y} \in \mathcal{B}_N$, where $\delta = t^{1/N}$. Therefore z is reachable from the origin in time $T(v_{\delta,\mathbf{y}})$. If $\bar{\tau}$ is an upper bound for the times $T(\{\Gamma, \psi(S'(\boldsymbol{\varepsilon}, \boldsymbol{\rho}))\})$ for $|\boldsymbol{\varepsilon}| \leqq \alpha$, $|\boldsymbol{\rho}| \leqq \alpha$, then $T(v_{\delta,\mathbf{y}}) \leqq \delta\bar{\tau}$. So z is reachable from 0 in time $t^{1/N}\bar{\tau}$. Hence

$$C_t \subseteq \text{Reach}\,(0, t^{1/N}\bar{\tau}).$$

Therefore Reach $(0, s)$ contains a neighborhood of 0, if s is small enough. This completes the proof.

6. Necessary conditions. We show that the sufficient condition of Theorem 2.1 is not necessary for small-time local controllability, but that some parts of it are, at least for analytic systems.

First we prove:

PROPOSITION 6.1. *Suppose the vector fields f_0, f_1 are real analytic. Then* (HLCC 1) *is a necessary condition for small-time local controllability from x_0.*

Proof. We may assume that M is an open set in $\mathbb{R}^n$, and $x_0 = 0 \in M$. Assume that the system (2.1) is small-time locally controllable from 0.

We first prove that 0 must be an equilibrium point. Assume this is not so. Then the segment joining the vectors $f_0(0) - Af_1(0)$, $f_0 + Af_1(0)$ does not contain the origin. This implies the existence of a linear functional $\lambda: \mathbb{R}^n \to \mathbb{R}$ such that $\langle \lambda, f_0(0) - Af_1(0)\rangle > 0$ and $\langle \lambda, f_0(0) + Af_1(0)\rangle > 0$. Let $\phi: \mathbb{R}^n \to \mathbb{R}$ be a smooth function such that $\phi(0) = 0$ and that $(d\phi)(0) = \lambda$. Then, on some neighborhood U of 0, there is a fixed $\alpha > 0$ such that $(f_0\phi - Af_1\phi)(x) \geqq \alpha$ and $(f_0\phi + Af_1\phi)(x) \geqq \alpha$ for all $x \in U$. Let $T > 0$ be such

that no trajectory of (2.1) from 0 leaves U in time $\leqq T$. Then, if $x(\cdot):[0,t]\to M$, $0<t\leqq T$, is any such trajectory, corresponding to a control $u(\cdot)$, we have

$$\phi(x(t))=\phi(0)+\int_0^t[(f_0\phi)(x(s))+u(s)(f_1\phi)(x(s))]\,ds\geqq\alpha t\geqq 0.$$

So no point $p\in M$, for which $\phi(p)<0$, is reachable in time $\leqq T$. Since ϕ has a nonzero gradient at 0, there are points p arbitrarily close to 0 for which $\phi(p)<0$. But then (2.1) is not STLC from 0, which is a contradiction. So 0 is an equilibrium point.

Now we must exclude the possibility that $f_0(0)+\bar{u}f_1(0)=0$ for $\bar{u}=A$ or $\bar{u}=-A$. After some obvious transformations, this is equivalent to proving that a system

$$\dot{x}=f(x)+vg(x),\qquad 0\leqq v\leqq 1, \tag{6.1}$$

(f, g analytic) cannot be STLC from 0 if $f(0)=0$. So, let us assume that (6.1) is STLC from 0, and that $f(0)=0$. Clearly, $g(0)\neq 0$. Let $\lambda:\mathbb{R}^n\to\mathbb{R}$ be a linear functional such that $\langle\lambda, g(0)\rangle=1$, and let $\phi:\mathbb{R}^n\to\mathbb{R}$ be analytic, and such that $\phi(0)=0$, $(d\phi)(0)=\lambda$.

Let U be open, such that $0\in U$, that the closure of U is a compact subset of M, and that $(g\phi)(x)\geqq\frac{1}{2}$ for $x\in U$. Then there is a constant $C>0$ such that the estimate

$$|(h_{i_1}\cdots h_{i_r}\phi)(x)|\leqq C^r r! \tag{6.2}$$

holds for all $i_1,\cdots,i_r$, and all $x\in U$. (Here we let $h_0=f$, $h_1=g$.)

Let $T>0$ be such that all the trajectories of (6.1) from 0 stay in U up to time T. If $v(\cdot):[0,t]\to[0,1]$ is a control with $t\leqq T$, and $x(\cdot):[0,t]\to U$ a corresponding trajectory, with $x(0)=0$, we have

$$\phi(x(t))=\int_0^t(f\phi)(x(s))\,ds+\int_0^t v(s)g(x(s))\,ds,$$

$$(f\phi)(x(s))=(f\phi)(0)+\int_0^s(f^2\phi)(x(\sigma))\,d\sigma+\int_0^s v(\sigma)(gf\phi)(x(\sigma))\,d\sigma$$

so that (since $(f\phi)(0)=0$):

$$\phi(x(t))=\int_0^t\int_0^s(f^2\phi)(x(\sigma))\,d\sigma\,ds+\int_0^t\int_0^s v(\sigma)(gf)(x(\sigma))\,d\sigma\,ds$$
$$+\int_0^t v(s)(g\phi)(x(s))\,ds.$$

Continuing in this fashion, one proves by induction that

$$\phi(x(t))=\int_0^t\int_0^{s_1}\cdots\int_0^{s_{m-1}}(f^m\phi)(x(s_m))\,ds_m\cdots ds_1$$
$$+\sum_{k=1}^m\int_0^t\int_0^{s_1}\cdots\int_0^{s_{k-1}}v(s_k)(gf^{k-1}\phi)(x(s_k))\,ds_k\cdots ds_1.$$

Since $x(s)\in U$ for $s\in[0,t]$, estimate (6.2) gives

$$\left|\int_0^t\int_0^{s_1}\cdots\int_0^{s_{m-1}}(f^m\phi)(x(s_m))\,ds_m\cdots ds_1\right|\leqq(Ct)^m. \tag{6.3}$$

Let T' be such that $CT'<1$, $T'\leqq T$. Then, if $t\leqq T'$, (6.3) yields

$$\phi(x(t))=\sum_{k=1}^{\infty}B_k(t),$$

where

$$B_k(t) = \int_0^t \int_0^{s_1} \cdots \int_0^{s_{k-1}} v(s_k)(gf^{k-1}\phi)(x(s_k))\, ds_k \cdots ds_1.$$

Inequality (6.2) yields the estimate

$$|B_k(t)| \leqq C^k k! \int_0^t \int_0^{s_1} \cdots \int_0^{s_{k-1}} v(s_k)\, ds_k \cdots ds_1,$$

so that

$$|B_k(t)| \leqq C^k k! \int_0^t v(s) \frac{(t-s)^{k-1}}{(k-1)!}\, ds$$

and therefore

$$|B_k(t)| \leqq kC(Ct)^{k-1} \int_0^t v(s)\, ds.$$

Let

$$D(t) = \sum_{k=2}^{\infty} B_k(t).$$

Since the series

$$\sum_{k=2}^{\infty} k(Ct)^{k-2}$$

converges uniformly for $0 \leqq t \leqq T'$, we have a bound

$$|D(t)| \leqq Et \int_0^t v(s)\, ds$$

for some fixed constant E.

On the other hand

$$B_1(t) = \int_0^t v(s)(g\phi)(x(s))\, ds \geqq \frac{1}{2} \int_0^t v(s)\, ds.$$

So:

$$\phi(x(t)) \geqq \left(\frac{1}{2} - Et\right) \int_0^t v(s)\, ds.$$

Let $T'' \leqq T'$ be such that $2ET' < 1$. Then, if $0 \leqq t \leqq T''$, we see that $\phi(x(t)) \geqq 0$. Therefore, no point p for which $\phi(p) < 0$ is reachable from 0 in time $\leqq T''$. Since $(d\phi)(0) \neq 0$, we conclude that (6.1) is not STLC from 0. Q.E.D.

We can also show

PROPOSITION 6.2. *Suppose the vector fields f_0, f_1 are real analytic. Then* (HLCC 2) *is a necessary condition for small-time local controllability from x_0.*

Proof. If (2.1) is STLC from x_0 then, in particular, the reachable set from x_0 has a nonempty interior. That is, (2.1) has the accessibility property from x_0 (cf. Sussmann–Jurdjevic [8]). For analytic systems $\dot{x} = f(x, u)$, the accessibility property from a point x_0 is equivalent to the condition that the Lie algebra generated by the vector fields $x \to f(x, u)$ be of full rank at x_0 (cf. [8]). So our conclusion follows. Q.E.D.

We now study the simplest possible way in which (HLCC 3) could be violated. That is, we consider an analytic system (2.1) for which

$$[f_1, [f_0, f_1]](x_0) \notin \mathscr{S}^1(f_0 + \bar{u}f_1, f_1)(x_0). \tag{6.4}$$

PROPOSITION 6.3. *Consider an analytic system* (2.1), *and a point* $x_0 \in M$ *for which* (6.4) *holds. Then* (2.1) *is not* STLC *from* x_0.

Proof. The conclusion is clear if either (HLCC 1) or (HLCC 2) is violated. So we may assume that (HLCC 1) and (HLCC 2) hold. Also, (6.4) remains unchanged if $[f_1, [f_0, f_1]]$ is replaced by $[f_1, [f_0 + \bar{u}f_1, f_1]]$, where $\bar{u}$ is such that $|\bar{u}| < A$, $(f_0 + \bar{u}f_1)(x_0) = 0$. Moreover, it is clear that we can enlarge the control set and make it symmetric about $\bar{u}$. Then we can rename $f_0 + \bar{u}f_1$ and call it f_0, i.e., we can assume that $\bar{u} = 0$. Finally, we can assume $A = 1$, M open in $\mathbb{R}^n$, $x_0 = 0 \in M$.

So we have a system

$$\dot{x} = f_0(x) + uf_1(x), \qquad |u| \leqq 1, \tag{6.5}$$

for which $f_0(0) = 0$, and we assume that

$$[f_1, [f_0, f_1]](0) \notin \text{Linear span } \{(\text{ad } f_0)^j(f_1)(0): j = 0, 1, \cdots\}.$$

We now choose an analytic function ϕ, defined in a neighborhood V of 0, which satisfies $\phi(0) = 0$, $(g\phi)(0) = 0$ for all $g \in \mathscr{S}^1(f_0, f_1)$, $([f_1, [f_1, f_0]]\phi)(0) = 1$, and $f_1\phi \equiv 0$ on V. (To show that ϕ exists, let $g_1, \cdots, g_n$ be vector fields such that $g_1(0), \cdots, g_n(0)$ is a basis of $\mathbb{R}^n$, that $g_1(0), \cdots, g_k(0)$ is a basis of $\mathscr{S}^1(f_0, f_1)(0)$, for some k, that $g_1 = f_1$ and $g_{k+1} = [f_1, [f_1, f_0]]$. If Φ^{g_i} denotes the flow of g_i, then there is a neighborhood V of 0 such that every $p \in V$ has a unique expression

$$p = \Phi^{g_1}(t_1)\Phi^{g_2}(t_2) \cdots \Phi^{g_n}(t_n)(0).$$

Then we define $\phi(p) = t_{k+1}$.)

Next we choose a $T > 0$ such that, for every control u for which $T(u) \leqq T$, and every $t \in [0, T(u)]$, the series

$$\text{Ser } (\mathbf{f})(\phi)(u_t)(0) = \sum_I \left(\int_0^t u_I\right)(f_I\phi)(0) \tag{6.6}$$

converges to $\phi(\pi(\mathbf{f}, u, t, 0))$, and the convergence is uniform in u, t.

The terms of the series (6.6) are of four types:

(1) terms corresponding to a multiindex $I = (i_1, \cdots, i_r)$ for which $i_1 = 0$ or $i_r = 1$;
(2) terms for which I is of the form $(1, 0, \cdots, 0)$;
(3) the term that corresponds to $I = (1, 1, 0)$;
(4) the terms for which (i) I contains at least two 1's, (ii) I begins with a 1 and ends with a 0, and (iii) $|I| \geqq 4$.

The Type I terms vanish because $f_0(0) = 0$, $f_1\phi \equiv 0$. The Type 2 terms vanish as well, for the following reason: if $I = (1, 0, \cdots, 0)$, then $f_I\phi$ is of the form $f_1 f_0^k \phi$. Now

$$f_1 f_0^k = [f_1, f_0]f_0^{k-1} + f_0 f_1 f_0^{k-1},$$

and $(f_0 f_1 f_0^{k-1}\phi)(0) = 0$ because $f_0(0) = 0$. Similarly

$$[f_1, f_0]f_0^{k-1} = [[f_1, f_0], f_0]f_0^{k-2} + f_0[f_1, f_0]f_0^{k-2}$$

and, again, $(f_0[f_1, f_0]f_0^{k-2}\phi)(0) = 0$. Continuing in this way, we find that

$$(f_1 f_0^k \phi)(0) = (-1)^k (\text{ad } f_0)^k (f_1)(\phi)(0),$$

and therefore $(f_1 f_0^k \phi)(0) = 0$, because $(\text{ad } f_0)^k (f_1) \in \mathscr{S}^1(f_0, f_1)$.

We now compute the contribution of the Type 3 term.

We get:

$$\left[\int_0^t \int_0^s u(\sigma) \int_0^\sigma u(\tau)\, d\tau\, d\sigma\, ds\right](f_1^2 f_0 \phi)(0),$$

i.e.,

$$\left[\frac{1}{2}\int_0^t v(s)^2\, ds\right](f_1^2 f_0 \phi)(0),$$

where

$$v(t) = \int_0^t u(s)\, ds.$$

But

$$f_1^2 f_0 = f_1[f_1, f_0] + f_1 f_0 f_1 = [f_1, [f_1, f_0]] + 2[f_1, f_0] f_1 + f_0 f_1^2.$$

Since $f_0(0) = 0$, and $f_1\phi \equiv 0$, we conclude that

$$(f_1^2 f_0 \phi)(0) = ([f_1, [f_1, f_0]]\phi)(0) = 1.$$

So the contribution of the Type 3 term is, simply, $\frac{1}{2}\int_0^t v(s)^2\, ds$. Therefore:

$$\phi(\pi(\mathbf{f}, u, t, 0)) = \frac{1}{2}\int_0^t v(s)^2\, ds + B,$$

where B is the sum of all Type 4 terms.

Now let I be a multi-index corresponding to a Type 4 term. Then $I = (1, 0_k, 1, J)$, where 0_k denotes a sequence of k zeros ($k \geqq 0$), and J is a multi-index such that $|J| + k + 2 = |I|$, and $|J| \geqq 1$. If $J = (j_1, \cdots, j_r)$, we have

$$\int_0^t u_I = \int_0^t \int_0^{s_r} \cdots \int_0^{s_2} u_{j_r}(s_r) \cdots u_{j_1}(s_1) W_k(s_1)\, ds_r \cdots ds_1, \tag{6.7}$$

where

$$W_k(t) = \int_0^t u(\tau_1) \int_0^{\tau_1} \cdots \int_0^{\tau_{k+1}} u(\tau_{k+2})\, d\tau_{k+2} \cdots d\tau_1.$$

Then

$$W_k(t) = \frac{1}{k!}\int_0^t u(s)(s-\sigma)^k u\, d\sigma\, ds.$$

If $k = 0$, we have

$$W_0(t) = \frac{v(t)^2}{2}.$$

If $k > 0$, then we can write $W_k(t)$ as

$$\frac{1}{k!}\int_0^t u(s)\left[\int_0^s (s-\sigma)^k u(\sigma)\, d\sigma\right] ds,$$

and we can use integration by parts to conclude that

$$W_k(t) = \frac{v(t)}{k!}\int_0^t (t-\sigma)^k u(\sigma)\, d\sigma - \frac{1}{(k-1)!}\int_0^t v(s) \int_0^s (s-\sigma)^{k-1} u(\sigma)\, d\sigma\, ds.$$

Another integration by parts yields

$$\int_0^t (t-\sigma)^k u(\sigma)\,d\sigma = k\int_0^t v(\sigma)(t-\sigma)^{k-1}\,d\sigma.$$

Also, if $k>1$

$$\int_0^s (s-\sigma)^{k-1}u(\sigma)\,d\sigma = (k-1)\int_0^s v(\sigma)(s-\sigma)^{k-2}\,d\sigma,$$

and, if $k=1$

$$\int_0^s (s-\sigma)^{k-1}u(\sigma)\,d\sigma = v(s).$$

So we get

$$W_1(t) = v(t)\int_0^t v(s)\,ds - \int_0^t v(s)^2\,ds,$$

and, if $k>1$,

$$W_k(t) = \frac{v(t)}{(k-1)!}\int_0^t (t-\sigma)^{k-1}v(\sigma)\,d\sigma - \frac{1}{(k-2)!}\int_0^t v(s)\int_0^s (s-\sigma)^{k-2}v(\sigma)\,d\sigma\,ds.$$

If we let $|v|_{2,t}$ denote the L^2 norm of v on $[0,t]$, we find

$$\left|\int_0^s (s-\sigma)^k v(\sigma)\,d\sigma\right| \leqq \frac{s^{k+1/2}}{(2k+1)^{1/2}}|v|_{2,t},$$

using the Schwarz inequality.

Then, if $k>1$, we get

$$|W_k(t)| \leqq |v(t)||v|_{2,t}\frac{t^{k-1/2}}{(k-1)!\,(2k-1)^{1/2}} + \frac{|v|_{2,t}^2 t^{k-1}}{(k-2)!\,(2k-3)^{1/2}(2k-2)^{1/2}}.$$

Then (6.7) yields (since $r\geqq 1$):

$$\left|\int_0^t u_I\right| \leqq \int_0^t\int_0^{s_r}\cdots\int_0^{s_2} |W_k(s_1)|\,ds_1\cdots ds_r$$

$$= \frac{1}{(r-1)!}\int_0^t (t-s)^{r-1}|W_k(s)|\,ds.$$

Then we get the estimate

$$\left|\int_0^t u_I\right| \leqq \frac{Ft^{r+k-1}}{(r-1)!\,(k-2)!}|v|_{2,t}^2$$

for some constant F, and for $k>1$.

When $k=1$, we find

$$\left|\int_0^t u_I\right| \leqq \frac{Ft^r}{(r-1)!}|v|_{2,t}^2.$$

Finally, when $k=0$ we have

$$\left|\int_0^t u_I\right| \leqq \frac{Ft^{r-1}}{(r-1)!}|v|_{2,t}^2.$$

On the other hand, we have the bound

$$|(f_I\phi)(0)| \leqq C^{r+k+2}(r+k+2)!,$$

for a fixed constant C. Therefore we get

$$\left|\left(\int_0^t u_I\right)(f_I\phi)(0)\right| \leqq FC^{r+k+2}t^{r+k-1}\frac{(r+k+2)!}{(r-1)!\,(k-2)!}|v|_{2,t}^2$$

where, if $k=0$ or $k=1$, the factorial $(k-2)!$ is replaced by a 1. This implies that

$$\left|\left(\int_0^t u_I\right)(f_I\phi)(0)\right| \leqq FC^{r+k+2}t^{r+k-1}\frac{(r+k)!\,(r+k+2)^5}{r!\,k!}|v|_{2,t}^2.$$

For any given r, there are 2^{r-1} choices for J. Therefore the contribution of all the Type 4 terms is bounded by

$$\beta = \sum_{\substack{k\geqq 0\\ r\geqq 1\\ k+r\geqq 2}} FC^{r+k+2}(r+k+2)^5\frac{(r+k)!}{r!\,k!}2^{r-1}t^{r+k-1}|v|_{2,t}^2.$$

But

$$\beta \leqq \sum_{\rho=2}^{\infty}\sum_{k+r=\rho} FC^3(\rho+2)^5(2Ct)^{\rho-1}\frac{\rho!}{r!\,k!}|v|_{2,t}^2,$$

so that

$$\beta \leqq 2FC^3|v|_{2,t}^2\sum_{\rho=2}^{\infty}(\rho+2)^5(4Ct)^{\rho-1}.$$

From this it follows that β satisfies

$$\beta \leqq Ht|v|_{2,t}^2 \quad \text{for } 0\leqq t\leqq T$$

for some constant H, if T is small enough. Choosing T even smaller, if necessary, we can have $HT\leqq\frac{1}{2}$. Then, if $0\leqq t\leqq T$, $\phi(\pi(\mathbf{f}, u, t, 0))$ is the sum of $\frac{1}{2}|v|_{2,t}^2$ and of the contribution of the Type 4 terms, which is bounded in absolute value by $\frac{1}{2}|v|_{2,t}^2$. So $\phi(\pi(\mathbf{f}, u, t, 0))\geqq 0$. This is true for every control, and every $t\leqq T$. Therefore no point p where $\phi(p)<0$ is reachable from 0 in time $\leqq T$. So (6.1) is not STLC from 0. Q.E.D.

So far, we have described what happens in a simple situation when (HLCC 3) fails. (HLCC 3) requires, in particular (assuming, for simplicity, that $\bar{u}=0$), that

$$\mathscr{S}^2(f_0, f_1)(x_0) = \mathscr{S}^1(f_0, f_1)(x_0). \tag{6.8}$$

The simplest violation of (6.8) occurs if

$$[f_1, [f_0, f_1]](x_0) \in \mathscr{S}^1(f_0, f_1)(x_0) \tag{6.9}$$

fails. In this case, we have seen that (2.1) cannot be STLC from x_0. The next possibility would be for (6.9) to hold, while

$$[f_1, [f_0, [f_0, f_1]]](x_0) \in \mathscr{S}^1(f_0, f_1)(x_0) \tag{6.10}$$

fails. However, it is easy to see that this cannot happen. If $f_0(x_0)=0$, and if (6.9) holds, then (6.10) automatically follows. Indeed, let $g\in\mathscr{S}^1(f_0, f_1)$ be such that

$$g(x_0) = [f_1, [f_0, f_1]](x_0).$$

Then

$$[f, g](x_0) = [f_0, [f_1, [f_0, f_1]]](x_0),$$

because both f_0 and $g - [f_1, [f_0, f_1]]$ vanish at x_0, and therefore their Lie bracket vanishes as well.

But $[f_0, [f_1, [f_0, f_1]]] = [f_1, [f_0, [f_0, f_1]]]$, and $[f_0, g] \in \mathscr{P}^1(f_0, f_1)$. So (6.10) holds.

This shows that, if we want to pursue the study of the violations of HLCC in an orderly fashion, the next case to look at is when

$$[f_1, [f_0, [f_0, [f_0, f_1]]]](x_0) \in \mathscr{P}^1(f_0, f_1)(x_0) \tag{6.11}$$

fails. If HLCC were necessary, as well as sufficient, for STLC from x_0, then it would follow, in particular, that a system for which (6.11) fails to hold cannot be STLC from x_0. Unfortunately, this is false. We show this by means of a counterexample.

Let $M = \mathbb{R}^3$, with coordinates x, y, z, and consider the system

$$\dot{x} = u, \tag{6.12a}$$

$$\dot{y} = x, \tag{6.12b}$$

$$\dot{z} = x^3 + y^2, \tag{6.12c}$$

with the control constraint $|u| \leqq 1$. Then

$$f_0 = x\frac{\partial}{\partial y} + (x^3 + y^2)\frac{\partial}{\partial z}, \qquad f_1 = \frac{\partial}{\partial x}.$$

The Lie brackets of interest are

$$[f_1, f_0] = \frac{\partial}{\partial y} + 3x^2\frac{\partial}{\partial z}, \qquad [f_0, [f_1, f_0]] = -2y\frac{\partial}{\partial z},$$

$$[f_1, [f_1, f_0]] = 6x\frac{\partial}{\partial z}, \qquad [f_0, [f_0, [f_1, f_0]]] = -2x\frac{\partial}{\partial z},$$

$$[f_0, [f_1, [f_1, f_0]]] = 0, \qquad [f_0, [f_0, [f_0, [f_1, f_0]]]] = 0,$$

$$[f_1, [f_1, [f_1, f_0]]] = 6\frac{\partial}{\partial z}, \qquad [f_1, [f_0, [f_0, [f_1, f_0]]]] = -2\frac{\partial}{\partial z}.$$

Therefore $\mathscr{P}^1(f_0, f_1)(0)$ is two-dimensional, and

$$[f_1, [f_1, f_0]](0), \qquad [f_0, [f_1, [f_1, f_0]]](0)$$

vanish (so that (6.9) and (6.10) hold at 0). However, (6.11) fails at 0. But the vector $[f_1, [f_1, [f_1, f_0]]](0)$ is linearly independent from $\mathscr{P}^1(f_0, f_1)(0)$.

We now show that the system (6.12) is STLC from x_0. Let $\mathbf{X} = (X_0, X_1)$, and let us work in $A^4(\mathbf{X})$, the free associative algebra in the noncommuting indeterminates X_0, X_1, where all monomials of degree $\geqq 5$ are declared to vanish. Let $\Gamma = (1, -1)$. By Proposition 3.3, there exists a 4-normal Γ-control $\{\Gamma, \mathbf{t}^0\}$. Using the "odd-even" trick, as in the proof of Theorem 2.1, we can find a 4-normal Γ-control $\{\Gamma, \mathbf{t}_0\}$ such that

$$\mathrm{Ser}_4(\{\Gamma, \mathbf{t}_0\}) = e^{Z_0}$$

where Z_0 is a linear combination of Lie monomials in X_0, X_1 that contain an even number of X_1's. Therefore Z_0 is a linear combination of X_0, $[X_1, [X_0, X_1]]$ and $[X_1, [X_0, [X_0, X_1]]]$. This implies that $Z_0(\mathbf{t})(0) = 0$.

Exactly as in the proof of Theorem 2.1, the fact that $(\Gamma, \mathbf{t}_0)$ is 4-normal implies that, if $\varepsilon > 0$ is small enough, then we can find a family of controls u_η, depending on the parameter $\eta = (\eta_1, \eta_2, \eta_3)$, $|\eta_j| \leqq \varepsilon$, such that $u_0 = \{\Gamma, \mathbf{t}_0\}$, and that

$$\mathrm{Ser}_4(u_\eta) = e^{Z(\eta)}$$

where $Z(\eta) = Z_0 + \eta_1 X_1 + \eta_2 [X_1, X_0] + \eta_3 [X_1 [X_1, [X_1, X_0]]]$. Since $T(u_\eta)$ is the coefficient of X_0 in $Z(\eta)$, the time $T(u_\eta)$ is actually independent of η, and equal to a number $T > 0$. If we let, for $\delta > 0$,

$$u_\eta^\delta(t) = u_\eta\left(\frac{t}{\delta}\right) \quad \text{for } t \in [0, \delta T],$$

we find that

$$\mathrm{Ser}_4(u_\eta^\delta) = e^{Z_\delta(\eta)}$$

where

$$Z_\delta(\eta) = Z_0^\delta + \eta_1 \delta X_1 + \eta_2 \delta^2 [X_1, X_0] + \eta_3 \delta^4 [X_1, [X_1, [X_1, X_0]]],$$

and where Z_0^δ is a linear combination of X_0, $[X_1, [X_0, X_1]]$, and $[X_1, [X_0, [X_0, X_1]]]$, with δ-dependent coefficients. Therefore $Z_0^\delta(\mathbf{f})(0) = 0$.

Now let B denote the ball of radius ε in $\mathbb{R}^3$, centered at the origin. For $y = (y_1, y_2, y_3) \in B$, and $0 < \delta < 1$, define

$$v_{\delta, y} = u_{\eta(\delta, y)}^\delta$$

where

$$\eta(\delta, y) = (y_1 \delta^3, y_2 \delta^2, y_3).$$

Then $T(v_{\delta,y}) = \delta T$, and

$$\mathrm{Ser}_4(v_{\delta,y}) = e^{Z_\delta(\eta(\delta,\eta))}$$

where

$$Z_\delta(\eta(\delta, y)) = Z_0^\delta + \delta^4 (y_1 X_1 + y_2 [X_1, X_0] + y_3 [X_1, [X_1, [X_1, X_0]]]).$$

Therefore

$$\mathrm{Ser}_4(v_{\delta,y}) = 1 + \delta^4 (y_1 X_1 + y_2 [X_1, X_0] + y_3 [X_1, [X_1, [X_1, X_0]]]) + Y$$

where Y is the sum of an $O(\delta^5)$ and of a sum of powers of Z_0^δ. If we plug in the f_i for the X_i, apply the result as a differential operator to a smooth function ϕ, and evaluate at 0, we find, using Proposition 4.1 plus the fact that $Z_0^\delta(\mathbf{f})(0) = 0$:

$$\phi(\pi(\mathbf{f}, v_{\delta,y}, \delta T, 0)) = \phi(0) + \delta^4 \left(\sum_{i=1}^3 y_i \langle d\phi(0), v_i \rangle \right) + O(\delta^5), \tag{6.13}$$

where

$$v_1 = f_1(0), \quad v_2 = [f_1, f_0](0), \quad v_3 = [f_1, [f_1, [f_1, f_0]]](0).$$

Applying (6.13) to the coordinate functions, we find

$$\pi(\mathbf{f}, v_{\delta,y}, \delta T, 0) = \delta^4 \left(\sum_{i=1}^3 y_i v_i \right) + O(\delta^5). \tag{6.14}$$

Since $\{v_1, v_2, v_3\}$ is a basis of $\mathbb{R}^3$, (6.14) implies, exactly as in the proof of Theorem 2.1, that (6.12) is STLC from 0.

Acknowledgments. The author is grateful to H. Hermes, who is not only the author of the conjecture proved here but who, in addition, persuaded him that the conjecture had to be true and might even be provable. Thanks are also due to the author's St. Louis friends (W. Boothby, D. Elliott, N. Kaloupsidis) who patiently listened to, and commented on, an earlier version of this proof, to P. Brunovsky and, especially, to B. Jakubczyk, who proposed the example of § 6.

REFERENCES

[1] R. HERMANN, *On the accessibility problem in control theory*, in International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, pp. 325–332.

[2] H. HERMES, *Lie algebras of vector fields and local approximation of attainable sets*, this Journal, 16 (1978), pp. 715–727.

[3] A. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control systems*, this Journal, 12 (1974), pp. 43–52.

[4] ———, *The high order maximal principle and its application to singular extremals*, this Journal, 15 (1977), pp. 256–293.

[5] C. LOBRY, *Controlabilité des systèmes non linéaires*, this Journal, 8 (1970), pp. 573–605.

[6] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 398–404.

[7] J. P. SERRE, *Lie Algebras and Lie Groups*, W. A. Benjamin, New York, 1965.

[8] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95–116.

[9] H. J. SUSSMANN, *A bang-bang theorem with bounds on the number of switchings*, this Journal, 17 (1979), pp. 629–651.

[10] ———, *A sufficient condition for local controllability*, this Journal, 16 (1978), pp. 790–802.

[11] ———, *Orbits of families of vector fields and integrability of distributions*, Trans. Amer. Math. Soc., 180 (1973), pp. 171–188.

[12] H. HERMES, *On local controllability*, this Journal, 20 (1982), pp. 211–220.

[13] ———, *Local controllability and sufficient conditions in singular problems*, J. Differential Equations, 20 (1976), pp. 213–232.

[14] ———, *Control systems which generate decomposable Lie algebras*, submitted to J. Differential Equations.

[15] A. J. KRENER, *Local approximation of control systems*, J. Differential Equations, 19 (1975), pp. 125–133.

[16] R. W. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167–176.

[17] E. G. GILBERT, *Functional expansions for the response of nonlinear differential systems*, IEEE Trans. Automat. Control, 22 (1977), pp. 909–921.

[18] C. LESIAK AND A. J. KRENER, *The existence and uniqueness of Volterra series for nonlinear systems*, IEEE Trans. Automat. Control, 23 (1978), pp. 1090–1095.

[19] M. FLIESS, *Développements fonctionnels en indéterminées non commutatives des solutions d'équations différentielles non linéaires forcées*, C.R. Acad. Sci. Paris, Sér. A, 287 (1978), pp. 1133–1135.

[20] ———, *Séries de Volterra et séries formelles non commutatives*, C.R. Acad. Sci. Paris, Sér. A, 280 (1975), pp. 965–967.

[21] ———, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.

[22] K. T. CHEN, *Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula*, Ann. Math., 65 (1957), pp. 163–178.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0004 $01.25/0

LOCAL CONTROLLABILITY, REST STATES AND CYCLIC POINTS*

ROSA MARIA BIANCHINI†

Abstract. The problem of steering a neighbourhood of the state x of a linear time-invariant control system, to the point itself is considered. The set of points for which this property holds when the control is subject to magnitude constraints, is characterised. Moreover it is proved that this set is strictly related to the set of points that can be reached from themselves by an admissible control map.

Key words. linear autonomous systems, local controllability, periodic trajectories

Introduction. Let us consider a linear autonomous control system (A, Γ)

$$\frac{d}{dt}x(t) = Ax(t) - c(t),$$

where A is a real $n \times n$ matrix, and $c(\cdot)$ is a locally integrable map whose values belong to Γ a nonempty subset of $\mathbb{R}^n$. The system (A, Γ) is said to be *locally controllable* if and only if there exists a neighbourhood of the origin whose points can be steered to zero in finite time along trajectories of (A, Γ).

If 0 belongs to the convex closure of Γ, cl co Γ, it is known [2] that (A, Γ) is locally controllable if and only if there exists a neighbourhood W of the origin and a time $T > 0$ such that each point of W can be steered to 0 in the same time T. Moreover [2], [4], [6] the conditions (a) and (b) below are, given the hypothesis $0 \in \text{cl co } \Gamma$, necessary and sufficient for local controllability:

(a) There is no eigenvector y, $y \neq 0$, of A^* such that y^*c is constant for c in Γ.

(b) There is no real eigenvector y, $y \neq 0$, of A^* such that $\sup_{c \in \Gamma} y^*c \leqq 0$.

It is easy to see that $0 \in \text{cl co } \Gamma$ is not a necessary condition to get local controllability. In § 1 we study this property without any assumption on Γ. We prove that, as in the case when $0 \in \text{cl co } \Gamma$, local controllability implies the existence of a neighbourhood of the origin whose points can be steered to zero along the trajectories of (A, Γ) in the same time, but in general conditions (a) and (b) are no longer sufficient to get local controllability, although they are necessary. In § 2 we consider the local controllability of (A, Γ) at a point x^0, i.e., the local controllability of the system $(A, \Gamma - Ax^0)$. We study the set $C(A, \Gamma)$ of points x^0 at which (A, Γ) is locally controllable. We prove that $C(A, \Gamma)$ is an open, convex set and we give necessary and sufficient conditions for this set to be nonempty.

This study will point out the important role played in local controllability, by the interior rest states of (A, Γ), i.e., the points x such that $0 \in \text{int}_R \text{ co } (\Gamma - Ax)$. It will be shown that $C(A, \Gamma)$ coincides with the set of points that can be reached from an interior rest state and from which the system (A, Γ) can be steered to an interior rest state. This property is used to give an explicit representation of the set $C(A, \Gamma)$ and then to give a necessary and sufficient condition for local controllability in the general case.

In § 3 we consider the cyclic points of the system (A, Γ). A point x is a cyclic point of (A, Γ) if there exists a control that steers x to x in time $T > 0$. We prove that if $C(A, \Gamma)$ is not empty, its closure contains all the cyclic points.

---

* Received by the editors September 10, 1979, and in final revised form September 3, 1982. This research was performed under the auspices of CNR GNAFA.

† Istituto Matematico U. Dini, Viale Morgagni 67/a, 50134 Firenze, Italy.

1. Local controllability. Let us first introduce some notation. Let Y be a subset of $\mathbb{R}^n$; then cl Y denotes the closure of Y, int Y the interior of Y, co Y the convex hull of Y, $\operatorname{int}_R Y$ the interior of Y relative to the minimal affine manifold containing Y. Let A be a matrix; then A^* is the transpose matrix of A. Let $V(t, A, \Gamma)$ denote the set of points that can be steered to 0 in time t along the trajectories of (A, Γ), i.e., the points which can be reached in time t, from 0, by solutions of the system

$$\dot{x} = -Ax + c(t).$$

Thus

$$V(t, A, \Gamma) = \left\{ \int_0^t e^{-As} c(s)\, ds \, c(\cdot) \in L^1_{\mathrm{loc}}(\mathbb{R}, \Gamma) \right\}.$$

Let $V(A, \Gamma) = \bigcup_{t>0} V(t, A, \Gamma)$.

DEFINITION 1.1. (A, Γ) is *locally controllable* if and only if

$$0 \in \operatorname{int} V(A, \Gamma).$$

The following example shows that if $0 \notin \operatorname{cl} \operatorname{co} \Gamma$, conditions (a) and (b) are no longer sufficient to get local controllability. Let

$$A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}, \qquad \Gamma = \{(x_1, x_2): (x_1 - 1)^2 + (x_2 - 1)^2 < \tfrac{1}{2}\}.$$

Then (A, Γ) satisfies conditions (a) and (b) but it is not locally controllable.

The conditions are still necessary. In fact:

THEOREM 1.1. *If* (A, Γ) *is locally controllable, then* (a) *and* (b) *hold.*

Proof. Let $y \neq 0$ be a real eigenvector of A^* such that $\sup_{c \in \Gamma} y^* c \leqq 0$ and let $x \in V(A, \Gamma)$

$$y^* x = \int_0^t y^* e^{-As} c(s)\, ds = \int_0^t e^{-\lambda s} y^* c(s)\, ds \leqq 0.$$

Then $V(A, \Gamma)$ is contained in a half space and (A, Γ) is not locally controllable. Let y, Im $y \neq 0$, be a complex eigenvector of A^* such that $y^* c = k$ for each c in Γ. Re y and Im y are two real linearly independent vectors. Let Y be the subspace spanned by Re y and Im y and let P be the orthogonal projector of $\mathbb{R}^n$ onto Y. Then $PV(A, \Gamma) \subset V(A, P\Gamma)$ and since $P\Gamma$ is a point, 0 is not an interior point of $V(A, P\Gamma)$ relative to Y. This implies that (A, Γ) is not locally controllable.

Remark 1.1. Let us remark that condition (a) is equivalent to int $V(t, A, \Gamma) \neq \varnothing$, for all $t > 0$. Then if (a) holds int $V(A, \Gamma) \neq \varnothing$. The converse is not true, as the following example shows:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \Gamma = \{(d, 1): d \in \mathbb{R}\},$$

$V(A, \Gamma) = \{(x_1, x_2): x_2 > 0\}$. Then int $V(A, \Gamma) \neq \varnothing$. But $(0, 1)$ is an eigenvector of A^* such that $y^* c = 1$ for all $c \in \Gamma$.

From the preceding theorem: Theorem 1.2 follows.

THEOREM 1.2. (A, Γ) *is locally controllable if and only if there exists* $T > 0$ *such that* $0 \in \operatorname{int} V(T, A, \Gamma)$.

Proof. Since (A, Γ) is locally controllable, Theorem 1.1 implies that (a) holds: Then int $V(t, -A, -\Gamma) \neq \varnothing$ for all $t > 0$. Let W be a neighbourhood of 0 contained in $V(A, \Gamma)$. If t_1 is sufficiently small

$$H = W \cap \operatorname{int} V(t_1, -A, -\Gamma) \neq \varnothing.$$

Let $z \in H$. Since $z \in V(A, \Gamma)$, there exists $c_1(\cdot) \in L^1_{\text{loc}}(\mathbb{R}, \Gamma)$ and $t_2 > 0$ such that

$$z = \int_0^{t_2} e^{-As} c_1(s)\, ds. \tag{1.1}$$

Let $K = e^{-t_1 A}(z - H)$; K is an open neighbourhood of the origin. Moreover (1.1) implies that $K \subset V(t_1 + t_2, A, \Gamma)$ and the proof is complete.

It is known that, for all $t > 0$, the following relation holds

$$\text{cl } V(t, A, \Gamma) = \text{cl } V(t, A, \text{cl co } \Gamma).$$

Then Theorem 1.2 implies:

COROLLARY 1.1. *A control system* (A, Γ) *is locally controllable if and only if the same is true for the system* $(A, \text{cl co } \Gamma)$.

The obvious relation $V(t, -A, -\Gamma) = -e^{At} V(t, A, \Gamma)$, implies also:

COROLLARY 1.2. *A control system* (A, Γ) *is locally controllable if and only if the same is true for* $(-A, -\Gamma)$.

Let us consider a more restrictive class of admissible controls. Let $K(\mathbb{R}, \Gamma)$ be the collection of all piecewise constant maps $c(\cdot) : \mathbb{R} \to \Gamma$. A. Bacciotti [1] has proved that if $V^0(t, A, \Gamma)$ is the set of points that can be steered to 0 in time t by means of a control $c(\cdot)$ belonging to $K(\mathbb{R}, \Gamma)$, then

$$\text{int}_R\, V(t, A, \Gamma) = \text{int}_R\, V^0(t, A, \Gamma).$$

Then Theorem 1.2 implies that, if $V^0(A, \Gamma) = \cup_{t>0} V^0(t, A, \Gamma)$, then:

COROLLARY 1.3. *If* (A, Γ) *is locally controllable* $V(A, \Gamma) = V^0(A, \Gamma)$.

Proof. Since $V^0(A, \Gamma) \subset V(A, \Gamma)$, we have only to prove the opposite inclusion. Let $x \in V(A, \Gamma)$, then there exists t' such that

$$x \in V(t', A, \Gamma).$$

From the definition of $V(t, A, \Gamma)$ it follows that

$$V(t' + t'', A, \Gamma) = V(t', A, \Gamma) + e^{-At'} V(t'', A, \Gamma)$$

and then if T is such that $0 \in \text{int } V(T, A, \Gamma)$

$$x \in \text{int } V(t' + T, A, \Gamma).$$

Bacciotti's result [1] implies $x \in V^0(A, \Gamma)$.

Let us now derive an important property of the set $V(A, \Gamma)$ in the case (A, Γ) locally controllable. Let us denote by X^0_A, X^+_A, X^-_A the sum of the radical subspaces associated with the characteristic roots of A having real part equal to 0, real part positive and real part negative, respectively. Kun [7] has proved that if $X^+_A = \{0\}$, then (A, Γ) locally controllable implies $V(A, \Gamma) = \mathbb{R}^n$. In general if P^+ denotes the projection of $\mathbb{R}^n$ onto X^+_A along $X^-_A + X^0_A$, then the following property holds:

PROPOSITION 1.1. *Let* (A, Γ) *be locally controllable. Then*

$$V(A, \Gamma) = P^+ V(A, \Gamma) + X^-_A + X^0_A.$$

Proof. Obviously $V(A, \Gamma) \subset P^+ V(A, \Gamma) + X^-_A + X^0_A$. Let us prove the opposite inclusion. Let $x \in P^+ V(A, \Gamma)$ and $y \in X^-_A + X^0_A$. Then there exist $z \in V(A, \Gamma)$ and $y' \in X^-_A + X^0_A$ such that $z = x + y'$. If $c_1(\cdot)$ is a control that steers z to 0 in time t_1, $c_1(\cdot)$ will steer $(x + y)$ in time t_1 to a point $u \in X^-_A + X^0_A$, since this subspace is A invariant. But Kun [7] has proved that $X^-_A + X^0_A \subset V(A, \Gamma)$, hence $u \in V(A, \Gamma)$, and the proof is complete.

2. The rest states and the set of locally controllable points. Let $x^0 \in \mathbb{R}^n$. Then

DEFINITION 2.1. *x^0 is locally controllable by* (A, Γ) ((A, Γ) is locally controllable at x^0) if and only if $(A, \Gamma - Ax^0)$ is locally controllable.

Let us denote by $C(A, \Gamma)$ the set of points locally controllable by (A, Γ). We want to study this set. Let us introduce the following definition [4]:

DEFINITION 2.2. A point $x \in \mathbb{R}^n$ is a *rest state* of (A, Γ) if $0 \in \operatorname{co}(\Gamma - Ax)$. It is an *interior rest* state if $0 \in \operatorname{int}_R \operatorname{co}(\Gamma - Ax)$.

Let us remark that a rest state x belongs to $C(A, \Gamma)$ if and only if $(A, \Gamma - Ax)$ satisfies conditions (a) and (b); if moreover x is an interior rest state, then condition (a) implies condition (b). So if (A, Γ) satisfies condition (a) each interior rest state belongs to $C(A, \Gamma)$.

PROPOSITION 2.1. *$C(A, \Gamma)$ is not empty if and only if condition* (a) *holds and there exist interior rest states of* (A, Γ).

Proof. Since (A, Γ) satisfies condition (a) if and only if the same is true for $(A, \Gamma - Ax)$, $\forall x \in \mathbb{R}^n$, what has been said above implies that the conditions are sufficient.

Let us prove that they are necessary. Let $x^0 \in C(A, \Gamma)$; then Theorem 1.1 implies that $(A, \Gamma - Ax^0)$ satisfies conditions (a) and (b). Then (A, Γ) satisfies condition (a) and from (b) (see [4]) it follows that $(A, \Gamma - Ax^0)$ and hence (A, Γ), has interior rest states.

Let us denote by $V(x, A, \Gamma)$ the set of points that can be steered to x by means of trajectories of (A, Γ). By definition we have

$$V(x, A, \Gamma) = V(A, \Gamma - Ax) + x.$$

LEMMA 2.1. *If x and y belong to $C(A, \Gamma)$, $x \in V(y, A, \Gamma)$.*

Proof. Since $x \in C(A, \Gamma)$, Corollary 1.2 implies that $x \in C(-A, -\Gamma)$. Then from Proposition 1.1

$$V(x, -A, -\Gamma) = P^- V(x, -A, -\Gamma) + X_A^+ + X_A^0,$$

where P^- is the projection of $\mathbb{R}^n$ onto X_{-A}^+ along $X_{-A}^- + X_{-A}^0$, i.e., the projection of $\mathbb{R}^n$ onto X_A^- along $X_A^+ + X_A^0$. $y \in C(A, \Gamma)$ implies that $V(y, A, \Gamma) = P^+ V(y, A, \Gamma) + X_A^- + X_A^0$, and then

$$V(x, -A, -\Gamma) \cap V(y, A, \Gamma) \neq \varnothing.$$

Let $z \in V(x, -A, -\Gamma) \cap V(y, A, \Gamma)$. Then $x \in V(z, A, \Gamma)$ and $z \in V(y, A, \Gamma)$ which proves that $x \in V(y, A, \Gamma)$.

Remark 2.1. Corollary 1.3 and Lemma 2.1 imply that if x and y belong to $C(A, \Gamma)$, there exists a piecewise constant control that steers x to y.

COROLLARY 2.1. *If x and y belong to $C(A, \Gamma)$, $V(x, A, \Gamma) = V(y, A, \Gamma)$.*

Proof. $x \in V(y, A, \Gamma)$ implies $V(x, A, \Gamma) \subset V(y, A, \Gamma)$. Changing x with y, we obtain the opposite inclusion.

THEOREM 2.1. *If $x \in C(A, \Gamma)$*

$$C(A, \Gamma) = V(x, A, \Gamma) \cap V(x, -A, -\Gamma).$$

Proof. Let $y \in C(A, \Gamma)$. Lemma 2.1 implies $y \in V(x, A, \Gamma)$ and $x \in V(y, A, \Gamma)$. But $x \in V(x, A, \Gamma)$ is equivalent to $y \in V(x, -A, -\Gamma)$ and then $y \in V(x, A, \Gamma) \cap V(x, -A, -\Gamma)$ which proves

$$C(A, \Gamma) \subset V(x, A, \Gamma) \cap V(x, -A, -\Gamma).$$

Let us prove now the opposite inclusion. Let $z \in V(x, A, \Gamma) \cap V(x, -A, -\Gamma)$. Since $z \in V(x, A, \Gamma)$ and since $V(x, A, \Gamma)$ is an open set [8], there exists a neighbourhood

W_z of z such that $W_z \subset V(x, A, \Gamma)$. But $z \in V(x, -A, -\Gamma)$ and then $W_z \subset V(x, A, \Gamma) \subset V(z, A, \Gamma)$ which proves that $z \in C(A, \Gamma)$.

COROLLARY 2.2. *$C(A, \Gamma)$ is an open and convex set.*

Proof. If $C(A, \Gamma)$ is empty, there is nothing to prove. Otherwise Proposition 2.1 implies that there exists an interior rest state z belonging to $C(A, \Gamma)$. Since z is a rest state [8] both $V(z, A, \Gamma)$ and $V(z, -A, -\Gamma)$ are open and convex sets. Then by Theorem 2.1 $C(A, \Gamma)$ is the intersection of two convex and open subsets.

We want to characterise the set of locally controllable points by its support function. Using this characterisation we will be able to give an explicit condition for local controllability.

THEOREM 2.2. *Let $0 \in \operatorname{co} \Gamma$. Then*

$$\text{(2.1)} \qquad \operatorname{int} V(A, \Gamma) = \left\{ x \in \mathbb{R}^n : y^* x < \int_0^{+\infty} \sup_{x \in \Gamma} y^* e^{-As} x \, ds \; y \in \mathbb{R}^n - \{0\} \right\}.$$

Proof. Let us introduce the following notation: If X is a subset of $\mathbb{R}^n$, then $H_X : \mathbb{R}^n \to \mathbb{R}$ is the support function of $\operatorname{co} X$, i.e., $H_X(y) = \sup_{x \in X} y^* x$. It is known that $H_X(\cdot)$ is a convex, lower semicontinuous map. This implies that $s \mapsto H_\Gamma(e^{-A^* s} y)$ is a nonnegative measurable map, and so its Lebesgue integral exists and it is finite or infinite [5]. Let $x \in \operatorname{int} V(A, \Gamma)$; from $0 \in \operatorname{co} \Gamma$ it follows that there exists t' such that $x \in \operatorname{int} V(t', A, \Gamma)$. Let

$$\mathcal{U} = \{ c(\cdot) \in L^1_{\mathrm{loc}}(\mathbb{R}, \Gamma) \} \quad \text{and} \quad y \in \mathbb{R}^n - \{0\}.$$

We have

$$y^* x < \sup_{c(\cdot) \in \mathcal{U}} \left\{ \int_0^{t'} y \, e^{-As} c(s) \, ds \right\} \leqq \int_0^{+\infty} \sup_{x \in \Gamma} y^* e^{-As} x \, ds.$$

Let x be such that

$$y^* x < \int_0^{+\infty} \sup_{x \in \Gamma} y^* e^{-As} x \, ds$$

for all $y \in \mathbb{R}^n - \{0\}$, and let us suppose $x \notin \operatorname{int} V(A, \Gamma)$. Since $0 \in \operatorname{co} \Gamma$ implies that $V(A, \Gamma)$ is a convex set, there exists $z \in \mathbb{R}^n - \{0\}$ such that

$$z^* x \geqq \sup_{\substack{t > 0 \\ c(\cdot) \in \mathcal{U}}} \left\{ \int_0^t z^* e^{-As} c(s) \, ds \right\}.$$

Hence

$$z^* x \geqq \int_0^{+\infty} \sup_{x \in \Gamma} z^* e^{-As} x \, ds,$$

a contradiction.

From Theorems 2.1 and 2.2 it is easy to derive the following:

COROLLARY 2.3. *x is locally controllable by (A, Γ) if and only if the set of interior states is not empty and moreover*

$$y^*(x - x_0) < \min \left\{ \int_0^{+\infty} \sup_{x \in \Gamma} y^* e^{-As} (x - A x_0) \, ds, - \int_0^{+\infty} \inf_{x \in \Gamma} y^* e^{As} (x - A x_0) \, ds \right\}$$

for each y belonging to $\mathbb{R}^n - \{0\}$, where x_0 is an interior rest state.

Remark 2.2. Let us note that if (A, Γ) satisfies condition (a) and x_0 is an interior rest state, then if y does not belong to X_A^+,

$$\int_0^{+\infty} \sup_{x\in\Gamma} y^* e^{-As}(x - Ax_0)\, ds = +\infty,$$

and for each y which does not belong to X_A^-

$$\int_0^{+\infty} \inf_{x\in\Gamma} y^* e^{As}(x - Ax_0)\, ds = -\infty.$$

Moreover if $y \in X_A^+$, then

$$\int_0^{+\infty} \sup_{x\in\Gamma} y^* e^{-As}(x - Ax_0)\, ds = \int_0^{+\infty} \sup_{x\in\Gamma} y^* e^{-As}x\, ds - y^*x_0$$

and if $y \in X_A^-$, then

$$\int_0^{+\infty} \inf_{x\in\Gamma} y^* e^{As}(x - Ax_0)\, ds = \int_0^{+\infty} \inf_{x\in\Gamma} y^* e^{As}x\, ds - y^*x_0.$$

And so Corollary 2.3 can be reformulated in the following way:

COROLLARY 2.3. *x is locally controllable by (A, Γ) if and only if*:

i) *(A, Γ) satisfies condition* (a);

ii) $\operatorname{Im} A \cap \operatorname{int}_R \operatorname{co} \Gamma \neq \varnothing$;

iii) *for each y belonging to $X_A^+ - \{0\}$, $y^*x < \int_0^{+\infty} \sup_{x\in\Gamma} y^* e^{-As}x\, ds$;*

iv) *for each y belonging to $X_A^- - \{0\}$, $y^*x < -\int_0^{+\infty} \inf_{x\in\Gamma} y^* e^{As}x\, ds$.*

3. Cyclic points. Let us denote by $V(t, x, A, \Gamma)$ the set of points that can be steered to x in time t by a trajectory of (A, Γ).

DEFINITION 3.1. A point x is *cyclic* for (A, Γ) if there exists $t > 0$ such that $x \in V(t, x, A, \Gamma)$.

By Theorem 1.2 and by the definition of locally controllable points it follows that each point of $C(A, \Gamma)$ is a cyclic point. Then if $\mathscr{C}(A, \Gamma)$ denotes the set of cyclic points

$$C(A, \Gamma) \subseteq \mathscr{C}(A, \Gamma).$$

This inclusion may be a proper inclusion; in fact it is easy to construct a system for which $C(A, \Gamma) = \varnothing$ but $\mathscr{C}(A, \Gamma)$ is not empty. But if $C(A, \Gamma) \neq \varnothing$, then its closure contains $\mathscr{C}(A, \Gamma)$.

THEOREM 3.1. *If $C(A, \Gamma) \neq \varnothing$, then*

$$\mathscr{C}(A, \Gamma) \subset \operatorname{cl} C(A, \Gamma).$$

Proof. Let y be a cyclic point. Then there exists $t_1 > 0$ such that

$$y \in V(t_1, y, A, \Gamma).$$

Let $z \in \mathbb{R}^n$; $V(t_1, y, A, \Gamma) = V(t_1, z, A, \Gamma) - e^{-At_1}(z - y)$. Then

$$(I - e^{-At_1})y \in V(t_1, z, A, \Gamma) - e^{-At_1}z. \tag{3.1}$$

Let us first consider the case in which A is such that $X_A^0 = \{0\}$. In this case $(I - e^{-At_1})$ is an invertible matrix. Let W be a neighbourhood of y; since $C(A, \Gamma) \neq \varnothing$, $\operatorname{int} V(t_1, z, A, \Gamma) \neq \varnothing$ and then (3.1) implies that

$$H = (I - e^{-At_1})W \cap \{\operatorname{int} V(t_1, z, A, \Gamma) - e^{-At_1}z\}$$

is not empty. Let $x \in H$. Then

$$x \in \operatorname{int}\{V(t_1, z, A, \Gamma) - e^{-At_1}(z-x)\} = \operatorname{int} V(t_1, x, A, \Gamma)$$

which implies $x \in C(A, \Gamma)$.

Consider next the case in which $X_A^0 \neq \{0\}$. Let Y be the quotient space $\mathbb{R}^n/X_A^0$. Y is a vector space, $\dim Y = n - \dim X_A^0 = p$. Let π denote the canonical projection of $\mathbb{R}^n$ onto Y. Since $AX_A^0 \subset X_A^0$, there exists a unique linear operator $\bar{\mathscr{A}}$ such that the following diagram commutes

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{A} & \mathbb{R}^n \\ {\scriptstyle\pi}\downarrow & & \downarrow{\scriptstyle\pi} \\ Y & \xrightarrow{\bar{\mathscr{A}}} & Y \end{array} .$$

Let J be the isomorphism between $\mathbb{R}^p$ and Y and let $\bar{A}$ be the matrix associated with $J^{-1}\bar{\mathscr{A}}J$. By its construction it follows that $\bar{A}$ is an invertible matrix. Let $\bar{\Gamma} = J^{-1}\pi\Gamma$; if $y(t)$ is a solution of (A, Γ), then $J^{-1}\pi y(t)$ is a solution of $(\bar{A}, \bar{\Gamma})$.

Let x be a cyclic point of (A, Γ); $\bar{x} = J^{-1}\pi x$ is a cyclic point of $(\bar{A}, \bar{\Gamma})$ and since $C(\bar{A}, \bar{\Gamma}) \supset J^{-1}\pi C(A, \Gamma) \neq \varnothing$ what we have proved before implies that $\bar{x} \in \operatorname{cl} C(\bar{A}, \bar{\Gamma})$. Let W be a neighbourhood of x, $\bar{W} = J^{-1}\pi W$ is a neighbourhood of $\bar{x}$ and so there exists $\bar{y} \in \bar{W} \cap C(\bar{A}, \bar{\Gamma})$. Let $z \in C(A, \Gamma)$. Since $\bar{z} = J^{-1}\pi z \in C(\bar{A}, \bar{\Gamma})$, by Remark 2.1 there exists a piecewise constant control $\bar{c}_1(\cdot)$: $\mathbb{R} \to \bar{\Gamma}$ such that the solution $\bar{z}(t)$ of $(\bar{A}, \bar{\Gamma})$ relative to the control $\bar{c}_1$, $\bar{z}(0) = \bar{y}$, satisfies the condition $\bar{z}(t') = \bar{z}$, for some $t' > 0$. Let $c_1(\cdot): \mathbb{R} \to \Gamma$ be a piecewise constant map such that $J^{-1}\pi c_1(\cdot) = \bar{c}_1(\cdot)$ and let $z(t)$ be a solution of (A, Γ) relative to the control $c_1(\cdot)$ and such that $z(0) \in \pi^{-1}Jy$. By the construction of $c_1(\cdot)$ it follows that $z(t') \in \pi^{-1}J\bar{z} = z + X_A^0$. But $z + X_A^0 \subset V(z, A, \Gamma)$ and then $\pi^{-1}J\bar{y} \subset V(z, A, \Gamma)$. By Remark 2.1, since $\bar{z}$ and $\bar{y}$ belong to $C(\bar{A}, \bar{\Gamma})$, there exists a piecewise constant control $\bar{c}_2(\cdot): \mathbb{R} \to \bar{\Gamma}$ that steers $\bar{z}$ to $\bar{y}$. Using the same arguments as before, we derive that $\pi^{-1}J\bar{z} \subset V(y', A, \Gamma)$, $\forall y' \in \pi^{-1}J\bar{y}$. But $z \in V(y', A, \Gamma)$ implies $y' \in V(z, -A, -\Gamma)$ and then $\pi^{-1}J\bar{y} \subset V(z, A, \Gamma) \cap V(z, -A, -\Gamma) = C(A, \Gamma)$. Since $\bar{y} \in J^{-1}\pi W$, it follows that $W \cap C(A, \Gamma) \neq \varnothing$ and the proof is complete.

REFERENCES

[1] A. BACCIOTTI, *Linear systems with piecewise constant controls*, Boll. Un. Mat. Ital., 18-A (1981), pp. 102–105.

[2] R. F. BRAMMER, *Controllability in linear autonomous systems with positive controllers*, this Journal, 10 (1972), pp. 339–353.

[3] G. S. GOODMAN, *Support functions and the integration of set-valued mappings*, International Atomic Energy Agency, 2 (1976), pp. 281–296.

[4] M. HEYMANN AND J. R. STERN, *Controllability of linear systems with positive controls: geometric considerations*, J. Math. Anal. Appl., 52 (1975), pp. 36–41.

[5] E. HEWITT AND K. STROMBERG, *Real and Abstract Analysis*, Springer-Verlag, New York, 1975.

[6] V. I. KOROBOV, A. P. MARINIC AND E. N. PODOLSKI, *Controllability of linear autonomous systems in presence of constraints on the control*, Differencial'nije Uravnenija, 11 (1975), pp. 1967–1979. (In Russian.)

[7] L. A. KUN, *Relations between local and global controllability*, Avmtomat. i Telemech, 10 (1977), pp. 12–15. (In Russian.)

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

[9] V. M. POPOV, *Hyperstability of Control Systems*, Grund. d. Math., Springer-Verlag, New York, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0005 $01.25/0

A FINITENESS CRITERION FOR NONLINEAR INPUT-OUTPUT DIFFERENTIAL SYSTEMS*

MICHEL FLIESS† AND IVAN KUPKA‡

Abstract. We solve the following problem which has been dealt with in several publications: when does a system have the same input–output behaviour as a system depending linearly on the state variables? Precise definitions of this question are supplied in the text. A particular case of the problem is the equivalence of a system with a bilinear one. At the end, we illustrate our methods and the scope of our results by a brief study of several examples.

Key words. nonlinear systems, state-affine systems, bilinear systems, representative functions, noncommutative generating power series

Introduction. In this paper, we prove a simple criterion for a general system to define the same input–output map as a subsystem of a system defined on a vector space by a family of affine vector fields and an affine output mapping.

Our criterion can be summed up in the following way: the vector space generated by the output function and its successive Lie derivatives along the vector fields of the system is finite dimensional. We study the problem in both the real analytic and the indefinitely differentiable categories. We also give sufficient local conditions which insure that the above finiteness condition is satisfied. Finally, in the real analytic case, we give a second, faster, proof based on the theory of noncommutative generating series (cf. [2]).

The problem stated above was studied for the first time by Krener [5], and later by Lo [8] and Hijab [3]. The solution using generating series has been extended to discrete-time systems by Normand–Cyrot [10] (see also Sontag [12]).

We end the paper with several examples, one from statistical physics (cf. Schenzle and Brand [11], Suzuki, Kaneko and Sasegewa [13]) and one from the theory of representative functions due to Hochschild and Mostow [4]. This last theory is closely related to our methods. In fact our paper can be considered as a study of finite dimensional representations of transformation pseudogroups.

1. Notation and usual definitions. In all this work, the control space and the output space will be kept fixed. We shall assume that the control space, denoted by C, is an open subset of a finite dimensional vector space and that the output space, denoted by E, is a finite dimensional vector space. Moreover, we assume that we have an admissible set $\mathscr{L}$ of controls. $\mathscr{L}$ will be a set of measurable functions $u : [0, T_u] \to C$ ($T_u > 0$ depending on u), containing the constant controls.

DEFINITION 1. (I). A *polysystem* (cf. Lobry [8]) will be a couple (M, Ξ), where M is a C^∞ (resp. C^ω) connected manifold, countable at infinity, Ξ is a family $\{\Xi_c \mid c \in C\}$ of C^∞ (resp. C^ω) vector fields satisfying the following conditions:

1) The mapping $M \times C \to TM$ (tangent structure of M), $(m, c) \mapsto \Xi_c(m)$ is C^∞ (resp. C^ω).

* Received by the editors June 18, 1981, and in revised form August 27, 1982. This paper was written under the auspices of the R.C.P. 567 of the Centre National de la Recherche Scientifique.

† Laboratoire des Signaux et Systèmes (Laboratoire Propre du C.N.R.S.), Ecole Supérieure d'Electricité, Plateau du Moulon, 91190 Gif-sur-Yvette, France.

‡ Institut Fourier, Laboratoire de Mathématiques Pures (Laboratoire Associé au C.N.R.S.), Université Scientifique et Médicale de Grenoble, B.P. 116,38402 Saint-Martin-d'Hères Cedex, France.

2) For any $u \in \mathscr{L}$ and any $m \in M$, the system $d\varphi(t)/dt = \Xi_{u(t)}(\varphi(t))$, $\varphi(0) = m$, has a unique maximal solution $\varphi_{m,u} : [0, T_{m,u}[\to M$.

(II). An *input–output system* (M, Ξ, h) is a triple where (M, Ξ) is a polysystem and $h : M \to E$ a C^∞ (resp. C^ω) mapping, called the output function.

Let us denote by $\mathrm{AC}(E)$ the space of all absolutely continuous mappings $\psi : [0, T_\psi[\to E$ ($T_\psi > 0$ depends on ψ).

DEFINITION 2. The *input–output mapping* of an input–output system (M, Ξ, h) is the mapping $\tilde{h} : \mathscr{L} \times M \to \mathrm{AC}(E)$, defined as follows: $\tilde{h}(u, m) = h \circ \varphi_{m,u}$.

DEFINITION 3. An input–output system (M, Ξ, h) is called *state-affine* (cf. Sontag [12]) if and only if

1) M is a finite dimensional vector space,

2) the vector fields $\{\Xi_c \mid c \in C\}$ are affine,

3) $h : M \to E$ is an affine mapping.

DEFINITION 4. An input–output system (M, Ξ, h) is called *control-affine* if, and only if,

1) C is a vector space,

2) the mapping $C \to$ vector fields on M, $c \mapsto \Xi_c$, is affine.

DEFINITION 5. An input–output system (M, Ξ, h), which is both state- and control-affine is called *affine*.[1]

2. Subsystems of an input–output system. The definitions that follow give a precise mathematical content to the idea that the input–output behaviour of one system is reflected inside the input–output behaviour of another system. Such a conceptualisation is required since one may want to restrict the behaviour of a system by either cutting down the state space or using a smaller family of control fields. Our concepts are related to those of Krener [6] and Hijab [3].

We denote by $\ll$ the following partial order relation on $AC(E)$: if $\varphi, \psi \in AC(E)$, $\varphi \ll \psi$ if:

- $T_\varphi \leqq T_\psi$,
- on $[0, T_\varphi[$, φ and ψ coincide.

DEFINITION 6. Given two C^∞ (resp. C^ω) input–output systems (M, Ξ, h) and (M', Ξ', h'), an *immersion* of (M, Ξ, h) into (M', Ξ', h') is a C^∞ (resp. C^ω) mapping $\tau : M \to M'$ such that:

1) For any couple $(x, y) \in M \times M$, $h(x) \neq h(y)$ implies $h'(\tau(x)) \neq h'(\tau(y))$.

2) For any $(u, m) \in \mathscr{L} \times M$, $\tilde{h}(u, m) \ll \tilde{h}'(u, \tau(m))$.

In this situation, we say that (M, Ξ, h) can be represented as a *subsystem* of (M', Ξ', h').

DEFINITION 7. (I). Given a C^∞ (resp. C^ω) polysystem (M, Ξ) and a family $\mathscr{F}$ of C^∞ (resp. C^ω) mappings of M into a finite dimensional vector V, the *observation space*[2] of $\mathscr{F}$ is the smallest vector subspace of $C^\infty(M, V)$ (resp. $C^\omega(M, V)$) containing $\mathscr{F}$ and stable under the action of the vector fields of Ξ.

(II). The family $\mathscr{F}$ is called *representative*[3] if the observation space is finite dimensional.

Notation. We shall denote by $O(\mathscr{F})$ the observation space of $\mathscr{F}$.

Remark. $\mathscr{F}$ is the vector subspace of $C^\infty(M, V)$ (resp. $C^\omega(M, V)$) generated by the functions $\{X_1^{\alpha_1} \cdots X_n^{\alpha_n} f \mid f \in \mathscr{F}; X_1, \cdots, X_n \in \Xi; \alpha_1, \cdots, \alpha_n \in \mathbf{N}\}$.

---

[1] Such systems were also called (internally) bilinear or regular.

[2] This terminology is borrowed from Sontag [13].

[3] Compare with [4].

The next definition is more technical, but it is useful in practice. Let us denote by $J_m(M, V)$ the space of all ∞-jets at m of mappings $M \to V$ (i.e., the space of all Taylor series at m) and $j_m : C^\infty(M, V) \to J_m^\infty(M, V)$ the mapping which to each f associates its ∞-jet at m (i.e., its Taylor series). It is well known that C^∞ (resp. C^ω) vector fields on M act on $J_m^\infty(M, V)$.

DEFINITION 8. (I). With the same notation as in Definition 3, the *observation space* of the family $\mathscr{F}$ at a point $m \in M$ is the smallest vector subspace of $J_m^\infty(M, V)$ containing the family $j_m^\infty \mathscr{F}$ and stable under the action of the fields from Ξ.

(II). $\mathscr{F}$ will be called *representative* at m if its observation space at m is finite dimensional.

Notation. We shall denote by $O(\mathscr{F}, m)$ the observation space of $\mathscr{F}$ at m.

3. Main results.

THEOREM 1. (I). *If the output mapping h of a C^∞ (resp. C^ω) input–output system (M, Ξ, h) is representative, then the system is representable as a subsystem of a C^∞ (resp. C^ω) state-affine system.*

(II). *If moreover (M, Ξ, h) is control-affine, it is a subsystem of an affine system.*

(III). *If the class $\mathscr{L}$ of admissible controls contains the piecewise-constant controls, then the converses of* (I) *and* (II) *are true.*

In this theorem, the condition for h to be representative is a global one. But it can be replaced by local ones.

THEOREM 2. *If (M, Ξ, h) is a real analytic input–output system, then h is representative if, and only if, it is representative at one point of M.*

This theorem is trivial. The next is harder.

THEOREM 3. *Let (M, Ξ, h) be a C^∞ input–output system such that Ξ is weakly controllable. Then h is representative if, and only if, it is representative at each point of M.*

In both Theorems 2 and 3, the only if part is trivial; the conditions in Theorem 3 cannot be weakened much. If we drop the weak controllability, it becomes false. If we do not assume that the set of all points at which h is representable is not everywhere dense, it is also false.

4. Proofs of the theorems.

a) *Proof of Theorem* 1. First the "if" part. Let us denote by $\mathscr{L}(\Xi)$ the $\mathbf{R}$-Lie algebra of vector fields generated by the family $\{\Xi_c \mid c \in C\}$. $\mathscr{L}(\Xi)$ operates on the observation space of h by the Lie derivatives: $\theta : \mathscr{L}(\Xi) \to \mathrm{End}\,(O(h))$. Call H the space of all $\mathbf{R}$-linear mappings $O(h) \to E$. H is finite dimensional and $\mathscr{L}(\Xi)$ operates on H in a natural way. Calling $\rho : \mathscr{L}(\Xi) \to \mathrm{End}\,(H)$ the corresponding $\mathbf{R}$-linear representation, it is defined as follows: for $l \in H$, $X \in \mathscr{L}(\Xi)$ and $f \in O(h)$, $(\rho(X)l)[f] = -l(\theta(X)f)$.

The linear vector fields on H associated to the endomorphisms $\{-\rho(\Xi_c) \mid c \in C\}$ define a polysystem Θ on H with control space C. Define a $\mathbf{R}$-linear mapping $\vec{h} : H \to E$ as follows: $\vec{h}(l) = l(h)$. It is then clear that the triple (H, Θ, h) is a state-affine input–output system, which is affine in case the polysystem (M, Ξ) is control-affine.

Presently, we shall show that (M, Ξ, h) is a subsystem of $(H, \Theta, \vec{h})$. Let us define a mapping $\tau : M \to H$ as follows: for $m \in M$ and for $f \in O(h)$, $\tau(m)[f] = f(m)$. It is clear that τ is C^∞ (resp. C^ω). To show that it is an immersion, we have to check the conditions 1 and 2 of Definition 6. Now it is clear that $\vec{h} \circ \tau = h$. Condition 1 follows. As for condition 2 let $(u, m) \in \mathscr{L} \times M$. For simplicity, call φ the trajectory $\varphi_{m,n}$. Let $\psi = \tau \circ \varphi$. By definition, $d\varphi(t)/dt = \Xi_{u(t)}(\varphi(t))$ for almost every $t \in [0, T_{m,n}[$. For any $f \in O(h)$, $d(f \circ \varphi)(t)/dt = (\theta(\Xi_{u(t)})f)(\varphi(t))$. But $f \circ \varphi = \tau(\varphi)[f] = \psi[f]$ and $(\theta(\Xi_{u(t)})f)(\varphi(t)) = -\rho(\Xi_{u(t)})\psi(t))[f] = \Theta_{u(t)}(\psi(t))[f]$. Hence

$(d/dt)(\psi(t)[f]) = \Theta_{u(t)}(\psi(t))[f]$. This being true for all f, one gets $d\psi(t)/dt = \Theta_{u(t)}(\psi(t))$. Moreover, $\psi(0) = \tau(\varphi(0)) = \tau(m)$. Hence, $\tilde{h}(u, \tau(m)) = \tilde{h} \circ \psi = h \circ \varphi$. This shows that $\tilde{h}(u, \tau(m)) = \tilde{h}(u, m)$ on $[0, T_m[$. Condition 2 is satisfied.

Finally, it is clear from the construction that, if (M, Ξ) is control-affine, the input–output system (H, Θ, h) is affine.

To show the converse of (I) and (II), we need the following simple result: given any $(c_1, \cdots, c_n) \in C^n$, there exists an open neighborhood V of $\{0\} \times M$ in $\mathbf{R}^n \times M$ such that the mapping $e : V \to M$, $(t_1, \cdots, t_n, m) \mapsto \exp t_n \Xi_{c_n} \circ \exp r_{n-1} \Xi_{c_{n-1}} \circ \cdots \circ \exp t_1 \Xi_{c_1}(m)$ is defined and C^∞ (resp. C^ω). For any $\bar{t} = (t_1, \cdots, t_n) \in \mathbf{R}_+^n$, define the piecewise-constant controls $u_{\bar{t}}$ as follows: $u_t(t) = c_j$ if $\sum_{k=1}^{j-1} t_k < t \leqq \sum_{k=1}^{j} t_k$. Then for any $(\bar{t}, m) \in V$, $h \circ e(\bar{t}, m) = \tilde{h}(u_{\bar{t}}, m)$. Assume now that $\tau : M \to W$ defines a C^∞ (resp. C^ω) immersion of (M, Ξ, h) into a state-affine system (W, Θ, l). It follows that, for each $\bar{t} \in \mathbf{R}_+^n$, the mapping $m \in V(\bar{t}) = \{m \mid (\bar{t}, m) \in V\} \mapsto h(e(\bar{t}, m)) \in E$ is the restriction to $V(\bar{t})$ of an element from $\tau^*(L)$, where L is the space of all $\mathbf{R}$-linear mappings $W \to E$. Since L is finite-dimensional, for any multi-index $\alpha = (\alpha_1, \cdots, \alpha_n) \in \mathbf{N}^n$ the mapping $m \in M \mapsto (\partial^{|\alpha|} / \partial t_1^{\alpha_1} \cdots \partial t_n^{\alpha_n}) h \circ e(\bar{t}, m)$ belongs to $\tau^* L$. A trivial computation shows that this derivate is $\Xi_{c_1}^{\alpha_1} \Xi_{c_2}^{\alpha_2} \cdots \Xi_{c_n}^{\alpha_n} h$. This ends the proof.

Proof of Theorems 2 *and* 3. Theorem 2 is an immediate consequence of analytic continuation.

The proof of Theorem 3 is harder. We prove a special case first. Let us assume that there exists an integer $d \geqq 0$ such that $\dim_{\mathbf{R}} 0(h, m) = d$ for all $m \in M$. Then we can prove the following lemma.

LEMMA 1. *If* $\dim_{\mathbf{R}} O(h, m) = d$ *for all* $m \in M$, *then*

a) $\dim_{\mathbf{R}} O(h) = d$,

b) $j_m : O(h) \to O(h, m)$ *is bijective for all* $m \in M$,

c) $O(h)$ *is invariant by the pseudogroup* $PS(\Xi)$ *generated by* Ξ.

Proof. The space $J^\infty(M, E)$ of all infinite jets of C^∞ mappings from M into E has a natural vector bundle structure[4] if we endow each fiber $J_m^\infty(M, E)$ with the simple convergence topology; a sequence $\{j_m^\infty f_k \mid k \in \mathbf{N}\}$ converges to $j_m^\infty f$ if, for any C^∞ vector fields $X_1, \cdots, X_l$ on M, the sequence $\{(X_1 \cdots X_l f_k)(m) \mid k \in \mathbf{N}\}$ of vectors in E converges to $(X_1 \cdots X_l f)(m)$. This vector bundle $J^\infty(M, E)$ is endowed with a natural connection D having the following characteristic property: let $s : \omega \to J^\infty(M, E)$ be a C^∞ section defined on an open set ω; then $Ds = 0$ if and only if there exists a function $f \in C^\infty(\omega, E)$ such that $J^\infty f = s$.

It is clear that the union $\cup_{m \in M} O(h, m)$ carries the structure of a d-dimensional vector subbundle $\hat{O}(h)$ of $J^\infty(M, E)$. Since $\hat{O}(h)$ is generated by the sections $j^\infty f$, $f \in O(h)$ and since $D(j^\infty f) = 0$, it follows that D restricts to a connection D' on $\hat{O}(h)$ which is integrable. $j^\infty O(h)$ is a vector subspace of the space of horizontal sections for D'. This last vector space is finite dimensional. Points a) and b) follow from this statement. Since $O(h)$ is finite dimensional and $X \cdot O(h) \subset O(h)$ for all $X \in \Xi$, c) is true.

In the general case, let $\nu : M \to \mathbf{N}$ be the function $m \mapsto \dim_{\mathbf{R}} O(h, m)$. It is clear that ν is lower semicontinuous. Baire's theorem implies then that the set ω of all $m \in M$ such that ν is constant in a neighborhood of m is open and dense in M. On each component ω_0 of ω, ν is constant and we can apply Lemma 1 to the restrictions $\Xi_0 = \Xi|_{\omega_0}$ and $h_0 = h|_{\omega_0}$. The next lemma extends part a) of Lemma 1 to certain points of the boundary of ω_0. Before stating it, let us notice the following: it is clear that the mapping $j_m : O(h) \to O(h, m)$ factors through $O(h_0)$. For simplicity's sake, we shall call the induced mapping j_m also.

[4] See the appendix.

LEMMA 2. *Let m be a point of the boundary of ω_0 and let $X \in \Xi \cup (-\Xi)$ be such that $e^{tX}(m)$ is defined for $t \in [0, a]$ $(a > 0)$ and $e^{tX}(m) \in \omega_0$ for $0 < t \leqq a$. Then the mapping $j_m^\infty : O(h_0) \to O(h, m)$ is bijective and $\nu(m) = \nu(\omega_0)$.*

Proof. Since $X \cdot O(h_0) \subset O(h_0)$, X induces a $\mathbf{R}$-linear endomorphism L of the finite dimensional $\mathbf{R}$-vector space $O(h_0)$. Let us denote by e^{tL} the exponential of L. For simplicity, we shall call m_t the point $e^{tX}(m)$, $0 \leqq t \leqq a$. Composition on the left with $e^{(a-t)X}$ induces $\mathbf{R}$-linear mappings $\varphi_t^* : J_{m_a}^\infty(M, E) \to J_{m_t}^\infty(M, E)$ such that on $O(h_0)$, $\varphi_t^* \circ j_{m_a}^\infty = j_{m_t}^\infty \circ e^{(a-t)L}$ for $-a \leqq -t < 0$. By continuity, it follows that $\varphi_0^* \circ j_{m_a} = j_m^\infty \circ e^{aL}$ on $O(h_0)$. Since $\operatorname{Ker} j_{m_a}^\infty = \{0\}$ and $\operatorname{Ker} \varphi_0^* = \{0\}$ one gets $\operatorname{Ker} j_m^\infty = \{0\}$. This implies that j_m^∞ is an isomorphism of $O(h_0)$ onto $O(h, m)$.

To finish the proof, we shall show that the complement F of ω in M is invariant under $\Xi \cup (-\Xi)$. Since F has an empty interior, this shows that F is empty. Let $X \in \Xi \cup (-\Xi)$; if F is not X-invariant, there is a point $m \in F$ such that $e^{tX}(m)$ is defined for $t \in [0, a]$ and $e^{tX}(m) \notin F$ if $t > 0$. There exists a connected component ω_0 of ω containing the set $\{e^{tX}(m) | 0 < t \leqq a\}$, and, by Lemma 2, $\nu(m) = \nu(\omega_0)$. We will discuss the following two cases disjointly:

1) There is a neighborhood V of m such that $\nu | V - F$ is constant.

2) There is no such neighborhood.

In the first case, ν is constant and equal to $\nu(\omega_0)$ on $V - F$, since $V \cap \omega_0 \subset V - F$. Then for all $x \in V$, $\nu(x) \leqq \nu(\omega_0)$ because F has an empty interior and ν is lower semicontinuous. There exists also a neighborhood W of m such that $\nu \geqq \nu(m) = \nu(\omega_0)$ on W. Hence ν is constant on the neighborhood $V \cap W$ of m. This implies that $m \in \omega$ and contradicts the fact that $m \in F$.

In the second case, any neighborhood of m contains points x such that $x \in \omega$ and $\nu(x) \neq \nu(\omega_0)$. If we choose such a point m_1 sufficiently close to m, $e^{tX}(m_1)$ will be defined for all $t \in [0, a]$ and $e^{aX}(m_1) \in \omega_0$. It is easy to find a ball B of codimension 1 in M, centered at m_1, contained in ω, such that e^{tX} is defined on B for all $t \in [0, a]$ and $e^{aX}(B) \subset \omega_0$. Then $\nu(B) = \nu(m_1)$ and $\nu(e^{aX}(B)) = \nu(\omega_0)$.

Let us denote by T the flow box $\{e^{tX}(B) | 0 \leqq t \leqq a\}$ and by γ_y the arc $\{e^{tX}(y) | 0 \leqq t \leqq a\}$ for any $y \in B$. Since $\nu(y) = \nu(B) = \nu(m_1) \neq \nu(\omega_0) = \nu(e^{aX}(y))$, $\gamma_y \cap F$ is not empty for any $y \in B$. It is clear that the set R of all $y \in B$ such that $\gamma_y \cap F$ has no interior with respect to γ_y, is a Baire subset of B. Let RF be the space $F \cap U\{\gamma_y | y \in R\}$ and let Z be the open subset in RF of all $x \in RF$ such that ν is locally constant on RF at x. RF is homeomorphic to a separable complete metric space, since it is a countable intersection of open sets in F. This shows that Z is dense in RF.

Take a $q \in Z$. There is an open set U in T such that $q \in U$ and ν is constant on $U \cap RF$, and equal to ν_1 say. Calling γ the arc γ_y containing q, $\gamma \cap F$ is totally disconnected. One can find then an open ball D of codimension 1, transversal to X and a number $a > 0$ such that the tube $T_1 = U\{e^{tX}(D) | 0 \leqq t \leqq a\}$ is contained in U and the set $D \cup e^{aX}(D)$ in $T - F$. By Lemma 2, ν is constant and equal to ν_1 on $\gamma_y \cap T_1$ for any $y \in R$ such that $\gamma_y \cap T_1 \cap F$ is not empty. In particular ν is equal to ν_1 on D. If, for some y, $\gamma_y \cap T_1 \cap F$ is empty, then ν is constant on $\gamma_y \cap T$ and equal to $\nu(D)$, that is ν_1. Hence for any $y \in R$, $\nu(\gamma_y \cap T) = \nu_1$. Since any open component of $T_1 - F$ meets some γ_y, $y \in R$, ν is constant and equal to ν_1 on $T_1 - F$. By the first part, T_1 is contained in $T - F$, contradicting the fact that $q \in F$.

This ends the proof of Theorems 2 and 3.

Another proof of Theorem 1 *in the C^ω control-affine case.* In this part, we assume that (M, Ξ, h) is real analytic and that $\Xi = X_0 + \sum_{j=1}^p u_j X_j$, where the $X_0, X_1, \cdots, X_p$ are C^ω vector fields on M. To each $m \in M$, we associate the following formal power series in the noncommutative indeterminates $\xi_0, \xi_1, \cdots, \xi_p$ and coefficients in E (cf. [1])

$$\mathfrak{g}_m = \sum_{\alpha \in \bar{p}^\infty} (X^{\check{\alpha}} \cdot h)(m)\xi_\alpha,$$

where $\bar{p}^\infty = \cup_{r \geqq 0} \{0, 1, \cdots, p\}^r$, $\alpha = (\alpha_1, \cdots, \alpha_r)$, $\check{\alpha} = (\alpha_r, \cdots, \alpha_1)$, $\xi_\alpha = \xi_{\alpha_1} \cdots \xi_{\alpha_r}$, $X^{\check{\alpha}} = X_{\alpha_r} X_{\alpha_{r-1}} \cdots X_{\alpha_1}$.

The connection between the system (M, Ξ, h) and the *generating* series $\mathfrak{g}_m$ is the following: for all $u \in \mathscr{L}$, the L^1-norm of which is sufficiently small, one gets

$$\tilde{h}(m, u) = \sum_\alpha (X^{\check{\alpha}} h)(m) w_\alpha,$$

where w_α is the iterated integral defined by induction as follows:

$$w_0(t) = t, \qquad w_k(t) = \int_0^t u_k(s)\, ds \qquad (k = 1, \cdots, p),$$

$$w_\alpha(t) = \int_0^t w_{\bar{\alpha}}(s)\, dw_{\alpha_1}(s) \qquad (\bar{\alpha} = (\alpha_2, \cdots, \alpha_p)).$$

To help the reader to understand the relation between these concepts and what has been done before, let us mention that the above series play the same role with respect to h as the Taylor series play in the classical case of a function. In fact, the Taylor series are particular (commutative) cases of the preceeding (cf. [1]). A conceptual framework, similar to the jet space theory, can be developed to handle the above expansions not only in the real analytic case, but in the C^∞ one as well. For lack of space, we shall not do it here. Finally, we have the definition and the theorem already stated in the foregoing approach.

DEFINITION 6'. (cf. [2]). Given two C^ω control-affine systems $(M, X_0 + \sum_{j=1}^p u_j X_j, h)$ and $(M', X_0' + \sum_{j=1}^p u_j X_j', h')$, an *immersion* of $(M, X_0 + \sum u_j X_j, h)$ into $M', X_0' + \sum u_j X_j', h')$ is a C^ω mapping $\tau: M \to M'$ such that, at $m \in M$ and $m' = \tau(m) \in M'$, the two systems have the same generating series.

THEOREM 1' (cf. [2]). *Assume $\mathscr{L}$ contains the piecewise constant controls. A C^ω control-affine system $(M, X_0 + \sum_{j=1}^p u_j X_j, h)$ can be immersed in an affine system if, and only if, the observation space of h is finite dimensional.*

Proof. We shall only sketch it, leaving the details to the reader (see [2]). Denote by $\mathscr{A}$ the associative **R**-algebra of differential operators generated by $X_0, X_1, \cdots, X_p$. $\mathscr{A}$ operates in a natural fashion on $O(h)$. Call $\rho: \mathscr{A} \to \mathrm{End}\,(O(h))$ the corresponding **R**-linear representation of $\mathscr{A}$. If H denotes the vector space of all **R**-linear mappings of $O(h)$ into E, $\mathscr{A}$ induces a representation $\rho': \mathscr{A} \to \mathrm{End}\,(H)$. Define a system $(H, Y_0 + \sum_{j=1}^p u_j Y_j, \vec{h})$ as follows: if $\lambda \in H$, $Y_k(\lambda) = \rho'(X_k)\lambda$, $\vec{h}(\lambda) = \lambda(h)$. The system is affine. An immersion of (M, Ξ, h) into this system is defined by $\tau: M \to H$, $\tau(m)[f] = f(m)$, for all $f \in O(h)$.

5. Some examples.

a) *A physical example.* Our first example has its origin in statistical physics (cf. Schenzle and Brand [11], Suzuki, Kaneko and Sasagewa [13]):

$$\dot{x}(t) = ax(t) - bx(t)^\alpha + u_1(t)x(t),$$

$$y(t) = \frac{1}{x(t)^{\alpha - 1}} \qquad (a, b \in \mathbf{R}, \quad \alpha \in \mathbf{N}, \quad \alpha \geqq 2).$$

Here, $M = \mathbf{R} - \{0\}$, $C = \mathbf{R}$, $E = \mathbf{R}$, $\Xi = X + u_1 Y$, where $X = (ax - bx^\alpha)\partial/\partial_2$ and $Y = x\partial/\partial_x$, $h(x) = 1/x^{\alpha - 1}$.

The Lie algebra Lie (Ξ) is two-dimensional: if we denote $x^{\alpha}\partial/\partial x$ by Z, we get $X = aY - bZ$ and $[Y, Z] = (1-\alpha)bZ$. Since $O(h) = \mathbf{R}\cdot 1 \oplus \mathbf{R}\cdot h$, consider the map

$$\tau : \mathbf{R} - \{0\} \to \mathbf{R}^2, \qquad x \mapsto \begin{bmatrix} 1 \\ \dfrac{1}{x^{\alpha-1}} \end{bmatrix}.$$

$$X\begin{bmatrix} 1 \\ \dfrac{1}{x^{\alpha-1}} \end{bmatrix} = \begin{bmatrix} 0 \\ b(\alpha-1) + \dfrac{a(1-\alpha)}{x^{\alpha-1}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ b(\alpha-1) & a(1-\alpha) \end{bmatrix}\begin{bmatrix} 1 \\ \dfrac{1}{x^{\alpha-1}} \end{bmatrix},$$

$$Y\begin{bmatrix} 1 \\ \dfrac{1}{x^{\alpha-1}} \end{bmatrix} = \begin{bmatrix} 0 \\ \dfrac{(1-\alpha)}{x^{\alpha-1}} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1-\alpha \end{bmatrix}\begin{bmatrix} 1 \\ \dfrac{1}{x^{\alpha-1}} \end{bmatrix}.$$

Hence, the above system can be immersed in the following affine system:

$$\dot{x}_1 = 0,$$

$$\dot{x}_2 = b(\alpha-1)x_1 + (1-\alpha)(a+u_1)x_2,$$

$$y = x_2.$$

b) *A simple example showing that $\mathscr{L}(\Xi)$ does not have to be finite dimensional.* Take $C = \mathbf{R}$, $M = \mathbf{R}^3$, $\Xi = \{X + uY \mid u \in C\}$, where $X = \partial/\partial x - e^{y^2}\partial/\partial z$, $Y = \partial/\partial z + e^{x^2}\partial/\partial z$. The function $h : M \to E$ is given by $h(x, y, z) = e^x + e^y$. Then $O(h) = \mathbf{R}\, e^X \oplus \mathbf{R}\, e^y H$ is the dual of $O(h)$; hence $H = \mathbf{R}\varepsilon_1 \oplus \mathbf{R}\varepsilon_2$, where $\{\varepsilon_1, \varepsilon_2\}$ is the basis dual to $\{e^x, e^y\}$. We get $\Theta = \{\rho(X) + u\rho(Y) \mid u \in C\}$, where

$$\rho(X) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \rho(Y) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

$h : H \to \mathbf{R}$ is defined by $h(\varepsilon_1) = h(\varepsilon_2) = 1$. We have verified that the above system can be immersed in an affine system, although its Lie algebra Lie (Ξ) is clearly infinite dimensional.

Representative functions. (i) Consider the one-dimensional system on the real line

$$\dot{x}(t) = u_1(t), \qquad y(t) = h(x(t)).$$

The observation space is spanned by $\{d^{\mu}h/dx^{\mu} \mid \mu \geqq 0\}$. Hence h is representative if and only if it is an exponential polynomial.

(ii) Take $C = \mathbf{R}^2$, $E = \mathbf{R}^{2l+1}$ $(l \geqq 1)$, $M = \{(x, y) \mid x, y \in R,\ x \neq 0\}$, $\Xi = \{Z + uX + vY \mid u, v \in \mathbf{R}\}$, where $X = -\partial/\partial x$, $Y = -2lx\partial/\partial y + x^2\partial/\partial x$, $Z = l\partial/\partial y - x\partial/\partial x$. Define $h : M \to \mathbf{R}^{2l+1}$ as follows: $h(x, y) = (e^y, e^y x^{-1}, e^y x^{-2}, \cdots, e^y x^{2l})$. Then h is representative. The details are left to the reader. We satisfy ourselves with the remark that $\{X, Y, Z\}$ generate a Lie algebra isomorphic to su (2).

Appendix. Vector bundles with infinite dimensional fibers and associated connections. A C^{∞} vector bundle is a quadruple (V_M, π, M, V_0) of a topological space V_M (total space), a C^{∞} manifold M (base space), a continuous surjective mapping $\pi : V_M \to M$ (projection) and a topological vector space V_0 (typical fiber) satisfying the following conditions:

1) For each $m \in M$, there exists an open neighborhood U_m of m and a homeomorphism $\Phi_m : \pi^{-1}(U_m) \xrightarrow{\text{onto}} U_m \times V_0$ such that, for all $v \in \pi^{-1}(U_m)$, $\Phi_m(v) = (\pi(v), \varphi_m(v))$.

2) For any $p, q \in M$, such that $U_p \cap U_q \neq \varnothing$, the mapping $g_{pq}: (U_p \cap U_q) \times V_0 \to V_0$, $(x, y) \mapsto \varphi_p \circ \Phi_q^{-1}(x, y)$ is linear and continuous in y for each given $x \in U_p \cap U_q$ and is a C^∞ mapping $U_p \cap U_q \to V_0$ in x for each given y in V_0.

A C^∞ section s of (V_M, π, M, V_0), defined in an open subset U of M, is a continuous mapping $s: U \to V_M$ such that: 1) $\pi \circ s = \mathrm{Id}_U$, 2) for any $m \in M$, such that $U_m \cap U \neq \varnothing$, the mapping $x \in U_m \cap U \to \varphi_m(s(x)) \in V_0$ is C^∞. For any open set $U \subset M$, the set $\Gamma^\infty(U, V_M)$ of all C^∞ sections defined on U is a vector space.

A linear connection D on (V_M, π, M, V_0) is the data for each open subset U of M of a linear mapping $D_U: \Gamma^\infty(U, V_M) \to \Lambda' U \otimes \Gamma^\infty(U, V_M)$, where $\Lambda' U$ is the space of all C^∞ differential forms of degree 1 on U, satisfying the following conditions: 1) the D_U commute with the restriction mappings, 2) for any $s \in \Gamma^\infty(U, V_M)$ and any $f \in C^\infty(U)$, $D_U(fs) = fD_U s + df \otimes s$.

In the case of $J^\infty(M, E)$, the typical fiber is $\mathscr{F}_d(E)$, the space of all formal power series in $d = \dim M$ commutative variables, with coefficients in E, endowed with the topology of convergence of the coefficients. One proves (Kumpera and Spencer [7]) that there exists a unique connection D on $J^\infty(M, E)$ having the following property: for any C^∞ mapping $f: U \to E$, $D_U j^\infty f = 0$. If $(x_1, \cdots, x_d): U \to R^d$ is any chart of M defined on an open set U, any $s \in \Gamma^\infty(U, J^\infty(M, E))$ is represented by a formal power series in d variables $X_1, \cdots, X_d$, $\sum_{\alpha \in \mathbf{N}^d} s_\alpha X^\alpha$, where the s_α are C^∞ functions $U \to E$. $D_U s$ is then represented by a series $\sum_\alpha \sigma_\alpha X^\alpha$, where the σ_α are C^∞ differential forms on U with values in E:

$$\sigma_\alpha = ds_\alpha - \sum_{j=1}^{d} s_{\alpha+\varepsilon_j}\, dx_j, \qquad \varepsilon_j = (0, \cdots, 0, 1, 0, \cdots, 0).$$

REFERENCES

[1] M. FLIESS, *Fonctionnelles causales non linéaires et indéterminées non commutatives*, Bull. Soc. Math. France, 109 (1981), pp. 3–40.

[2] ———, *Finite-dimensional observation-spaces for non-linear systems*, Joint Workshop on Feedback and Synthesis of Linear and Nonlinear Systems, Bielefeld-Rome, June–July 1981, Lecture Notes in Control and Information Science 39, D. Hinrichsen and A. Isidori, eds, Springer-Verlag, Berlin, 1982, pp. 73–77.

[3] O. B. HIJAB, *Minimum energy estimation*, Ph.D. Thesis, University of California, Berkeley, 1980.

[4] G. HOCHSCHILD AND G. D. MOSTOW, *Representations and representative functions of Lie groups*, Ann. Math., 66 (1957), pp. 495–542.

[5] A. J. KRENER, *Bilinear and nonlinear realizations of input–output maps*, this Journal, 13 (1975), pp. 827–834.

[6] ———, *A decomposition theory for differentiable systems*, this Journal, 15 (1977), pp. 813–829.

[7] A. KUMPERA AND D. SPENCER, *Lie Equations, Vol. I: General Theory*, Ann. Math. Studies 73, Princeton Univ. Press, Princeton, N.J., 1972.

[8] J. T. LO, *Global bilinearization of systems with controls appearing linearly*, this Journal, 13 (1975), pp. 879–885.

[9] C. LOBRY, *Dynamical polysystems and control theory*, in Geometric Control Theory, D. Q. Mayne and R. W. Brockett, eds. D. Reidel, Dordrecht, 1973, pp. 1–42.

[10] D. NORMAND-CYROT, *Une condition de réalisation par systèmes à état-affine discrets*, 5th International Conference on Analysis Optimization of Systems, Versailles, December 1982, Lecture Notes in Control and Information Science, 44, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, 1982, pp. 88–98.

[11] A. SCHENZLE AND H. BRAND, *Multiplicative stochastic processes in statistical physics*, Phys. Rev. A, 20 (1979), pp. 1628–1647.

[12] E. D. SONTAG, *Polynomial Response Maps*, Lecture Notes in Control and Information Science, 13, Springer-Verlag, Berlin, 1979.

[13] M. SUZUKI, K. KANEKO AND F. SASAGEWA, *Phase transition and slowing down in non-equilibrium stochastic processes*, Prog. Theoret. Phys., 65 (1981), pp. 828–849.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0006 $01.25/0

OPTIMAL CONTROL PROBLEMS INVOLVING SECOND BOUNDARY VALUE PROBLEMS OF PARABOLIC TYPE*

Z. S. WU† AND K. L. TEO‡

Abstract. Three classes of optimal control problems involving second boundary value problems of parabolic type are considered. The controls are assumed to act through the forcing terms and through the initial and boundary conditions.

A sufficient condition for optimality is derived for the first optimal control problem. For the second problem, a necessary and sufficient condition for optimality is derived, and a method for constructing an optimal control is given. For the third problem, a necessary and sufficient condition for optimality is derived, a result on the existence of optimal controls is proved, an iterative method for solving this optimal control problem is devised, and, finally, the convergence property of this iterative method is investigated.

Key words. parabolic differential equations, second boundary value problems, optimal control problems, necessary and sufficient conditions, existence theory, algorithms, convergence of algorithms

1. Introduction. In this paper, we consider three classes of optimal control problems involving second boundary value problems of parabolic type with controls appearing both in the forcing terms and on the boundary and initial conditions. The partial differential equations involved are as described in [5] and [2]. The first problem is the most general case in which a convex cost functional is considered. A sufficient condition for optimality is derived for this general optimal control problem. In the second problem, the cost functional is assumed to be linear with respect to the solution of the second boundary value problem. The third problem, which overlaps the second one, involves the convex cost functional and a special case of the second boundary value problem. More precisely, the forcing terms and the initial and boundary conditions of this special second boundary value problem are linear with respect to the control variable.

For the second and third problems, answers to four major questions found in the study of optimal control theory are provided. These four questions concern:

(i) necessary conditions for optimality,
(ii) sufficient conditions for optimality,
(iii) existence of optimal controls,
(iv) methods for constructing an optimal control.

For the second problem, an optimal control can be computed in only one iteration. This is not possible for the third problem. However, the method proposed, which is an iterative method, can be used to construct a minimizing sequence of controls. Furthermore, certain convergence properties of the sequence of controls are also established.

Except for question (ii), answers to the questions are not known for the first problem. Continuing research in this area, which is expected to be heavy going, will certainly lead to some interesting results.

* Received by the editors December 30 1981, and in revised form July 15, 1982. This work was partially supported by the Australian Research Grants Committee and was done during the period when Z. S. Wu was an Honorary Visiting Fellow in the School of Mathematics at the University of New South Wales, Australia.

† Department of Mathematics, Zhongshan University, Guangzhou, China.

‡ Department of Applied Mathematics, The University of New South Wales, Kensington, New South Wales 2033, Australia.

Optimal control problems involving second boundary value problems of parabolic type have also been treated in [3], [4], [5], [9] and many others.

In [3] and [5], only very special cases of our third optimal control problem are considered.

Lions and Magenes [9] deal with the case involving a quadratic cost functional and a linear parabolic systems with constant coefficients. Furthermore, only a necessary and sufficient condition for optimality is derived for this case.

In [7], an abstract convex optimal control problem is considered. A method of estimating the rate of convergence of approximation to this problem is proposed, and a necessary condition for optimality involving projections on the set of admissible controls is derived. These general abstract results are then applied to a class of optimal control problems involving second boundary value problems.

In [4], a class of boundary-distributed linear control systems in Banach space is considered. A necessary condition for optimality is then derived for a convex optimal control problem involving such systems. A duality result is also obtained. These results are applicable to convex optimal control problems involving first or second boundary value problems of parabolic type.

In both [4] and [7], the second boundary value problems concerned are only special cases of our third optimal control problem in the sense that the coefficients of their differential equations are all independent of the time variable.

For our first and second optimal control problems, we note that the controls enter the data for the linear equation in a nonlinear way (in distinction to truly nonlinear problems where the governing equation is itself nonlinear). The problem of optimal control of induction heating discussed in [5, p. 19] is an example of such a problem.

Note that the heavily worked time optimal control problem does not fit the framework of this paper. For references on some time optimal control problems, see [5], [2, Chap. 5] and the relevant articles cited therein.

2. The problem statement. Let Ω be a bounded region in R^n (n-dimensional Euclidean space) with its boundary and closure denoted by S and $\bar{\Omega}$ respectively. It is assumed throughout the paper that Ω has the uniform C^3-regularity property as defined in [1, § 4.6, p. 67]. With this assumption, the boundary S of Ω is of the class C^3 (in the sense defined in [5, p. 10]).

Denote a point in R^n by $x \equiv (x_1, \cdots, x_n)$, and time by t. Let T be a fixed positive real number. Let $Q \equiv \Omega \times (0, T)$, $\bar{Q} \equiv \bar{\Omega} \times [0, T]$ and $S_T \equiv S \times [0, T]$.

Consider the parabolic partial differential operator L defined by

$$(1) \qquad L\zeta \equiv \frac{\partial \zeta}{\partial t} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left[\sum_{j=1}^{n} a_{ij}(x,t)\frac{\partial \zeta}{\partial x_j} + a_i(x,t)\zeta\right] - \sum_{i=1}^{n} b_i(x,t)\frac{\partial \zeta}{\partial x_i} - c(x,t)\zeta,$$

where $a_{ij}(i, j = 1, \cdots, n)$, a_i, b_i $(i = 1, \cdots, n)$ and c are measurable functions from Q into R.

Let U_1, U_2 and U_3 be fixed compact and convex subsets of R^{m_1}, R^{m_2} and R^{m_3} respectively.

Let u_1, u_2 and u_3 be measurable functions from Q, Ω and S_T into U_1, U_2 and U_3 respectively. Then, $u \equiv (u_1, u_2, u_3)$ is called *an admissible control.* We denote by $\mathcal{U} \equiv (\mathcal{U}_1, \mathcal{U}_2, \mathcal{U}_3)$ *the class of all such admissible controls.*

Let $F_j: Q \times U_1 \to R$ $(j = 1, \cdots, n)$, $f: Q \times U_1 \to R$, $\psi_0: \Omega \times U_2 \to R$ and $\psi: S_T \times U_3 \to R$ be measurable functions.

We now consider the following second boundary value problem:

$$
\begin{aligned}
&L\phi(x,t)=\sum_{i=1}^{n}\frac{\partial}{\partial x_i}[F_i(x,t,u_1(x,t))]+f(x,t,u_1(x,t)), && (x,t)\in Q,\\
&\phi|_{t=0}=\psi_0(x,u_2(x)), && x\in\Omega,\\
&\left[\frac{\partial\phi}{\partial\nu_L}+\sigma(s,t)\phi\right]\Big|_{S_T}=\psi(s,t,u_3(s,t)), && (s,t)\in S_T,
\end{aligned}
\tag{2}
$$

where $\partial\phi/\partial\nu_L=\sum_{i,j=1}^{n}(a_{ij}(s,t)\partial\phi/\partial x_j+a_i(s,t)\phi)\cos\alpha_i$, $(s,t)\in S_T$; α_i is the angle formed by the outward normal to S with the x_i axis, and $\sigma: S_T\to R$ is a measurable function.

To specify our optimal control problem, we need to introduce a cost functional. For this, let $G_1: Q\times R\to R$, $G_2:\Omega\times R\to R$, $G_3: S_T\times R\to R$, $H_1: Q\times U_1\to R$, $H_2:\Omega\times U_2\to R$ and $H_3: S_T\times U_3\to R$ be real-valued functions. Furthermore, we assume that $G_1(x,t,\cdot)$, $G_2(x,\cdot)$, $G_3(s,t,\cdot)$, $H_1(x,t,\cdot)$, $H_2(x,\cdot)$ and $H_3(s,t,\cdot)$ are differentiable and convex functions on their respective domains of definition.

Let $\phi(u)$ denote the weak solution (to be defined in Definition 2.1) of the second boundary value problem (2) corresponding to the control $u\equiv(u_1,u_2,u_3)\in\mathcal{U}$.

Our optimal control problem may now be stated as:

Subject to the system (2), find a control $u\in\mathcal{U}$ that minimizes the cost functional

$$
\begin{aligned}
J(u)=&\int_Q\int\{G_1(x,t,\phi(u)(x,t))+H_1(x,t,u_1(x,t))\}\,dx\,dt\\
&+\int_\Omega\{G_2(x,\phi(u)(x,T))+H_2(x,u_2(x))\}\,dx\\
&+\int_{S_T}\int\{G_3(s,t,\phi(u)(s,t))+H_3(s,t,u_3(s,t))\}\,ds\,dt.
\end{aligned}
\tag{3}
$$

For convenience, this optimal control problem will be referred to as the problem (P).

We now itemize the aims of the present paper as follows:

1. We shall derive a sufficient condition for optimality for the problem (P).

2. Consider the problem (P) under the additional assumptions that the functions $G_i(i=1,2,3)$ are linear in ϕ. For this special problem, we shall derive a necessary and sufficient condition for optimality and then detail a method for constructing an optimal control.

3. Consider the problem (P) under the additional assumptions that the functions F_i $(i=1,\cdots,n)$, f, ψ_0 and ψ are linear in their respective components of $u\equiv(u_1,u_2,u_3)$. Then, firstly, we shall show that this special problem has a solution. Secondly, we shall derive a necessary and sufficient condition for optimality. Thirdly, we shall devise an iterative method for solving this problem. Finally, the convergence property of this iterative method will be investigated.

3. Preparatory results. To begin, we shall note that the results of this paper depend heavily on the theory of parabolic partial differential equations provided by [8]. Thus all subsequent definitions and assumptions are drawn from that source.

Let $L_p(\Omega)$ be the space of all pth power integrable functions ϕ from Ω into R, with finite norm

$$
\|\phi\|_{p,\Omega}\equiv\left[\int_\Omega|\phi|^p\,dx\right]^{1/p},
$$

and let $L_\infty(\Omega)$ be the space of all essentially bounded measurable functions ϕ from Ω into R, with finite norm

$$\|\phi\|_{\infty,\Omega} \equiv \operatorname{ess}_{x\in\Omega} \sup |\phi(x)|.$$

Let $L_{q,r}(Q)$, $1 \leqq q, r \leqq \infty$, denote the space of all measurable functions f from Q into R, with finite norm

$$\|f\|_{q,r,Q} \equiv \left[\int_0^T \left\{\int_\Omega |f(x,t)|^q \, dx\right\}^{r/q} dt\right]^{1/r}$$

for $1 \leqq q, r \leqq \infty$;

$$\|f\|_{q,\infty,Q} \equiv \operatorname{ess}_{t\in[0,T]} \sup \{\|f(\cdot,t)\|_{q,\Omega}\}$$

for $1 \leqq q < \infty$, $r = \infty$;

$$\|f\|_{\infty,r,Q} \equiv \left[\int_0^T \{\|f(\cdot,\tau)\|_{\infty,\Omega}\}^r \, dt\right]^{1/r}$$

for $q = \infty$, $1 \leqq r < \infty$;

$$\|f\|_{\infty,\infty,Q} \equiv \operatorname{ess}_{(x,t)\in Q} \sup |f(x,t)|$$

for $q = \infty$, $r = \infty$.

For simplicity, we denote $L_{q,q}(Q)$ by $L_q(Q)$ and the norm $\|\cdot\|_{q,q,Q}$ by $\|\cdot\|_{q,Q}$.

Let $W_2^{1,1}(Q)$ be the Hilbert space of all measurable functions defined on Q with finite scalar product

$$\langle Z, Y\rangle_{W_2^{1,1}(Q)} \equiv \int_Q \int \left\{Z(x,t)Y(x,t) + \sum_{i=1}^n Z_{x_i}(x,t)Y_{x_i}(x,t) + Z_t(x,t)Y_t(x,t)\right\} dx\, dt,$$

where $Z_{x_i} \equiv \partial Z/\partial x_i$ $(i = 1, \cdots, n)$ and $Z_t \equiv \partial Z/\partial t$ are the generalized derivatives, and the same notation is applied to the function Y.

$W_2^{1,0}(Q)$ is the Hilbert space of all measurable functions defined on Q with finite scalar product

$$\langle Z, Y\rangle_{W_2^{1,0}(Q)} \equiv \int_Q \int \left\{Z(x,t)Y(x,t) + \sum_{i=1}^n Z_{x_i}(x,t)Y_{x_i}(x,t)\right\} dx\, dt.$$

$V_2(Q)$ is the Banach space of all those functions ϕ in $W_2^{1,0}(Q)$, with finite norm

$$|\phi|_Q \equiv \|\phi\|_{2,\infty,Q} + \|\phi_x\|_{2,Q},$$

where

$$\|\phi_x\|_{2,Q} \equiv \left[\int_Q \int \left\{\sum_{i=1}^n (\phi_{x_i}(x,t))^2\right\} dx\, dt\right]^{1/2}.$$

$V_2^{1,0}(Q)$ is the Banach space of all those functions ϕ in $V_2(Q)$ that are continuous in t (in the norm of $L_2(\Omega)$) and equipped with the norm

$$\|\phi\|_Q \equiv \sup_{t\in[0,T]} \{\|\phi(\cdot,t)\|\}_{2,\Omega} + \|\phi_x\|_{2,Q}.$$

Besides, let $C^\infty(Q)$ be the space of all those functions, from Q into R, which are continuously partial differentiable of arbitrary order, and let $C_0^\infty(Q)$ be the space of all those functions in $C^\infty(Q)$ with compact support in Q.

We shall denote by ds the surface measure on S, induced by dx. The integral of a function $f(s)$ over the boundary S is to be understood as defined in [1, § 5.21, p. 114] and is to be denoted by $\int_S f(s)\, ds$.

Let $L_{q,r}(S_T)$, $1 \leqq q, r \leqq \infty$, denote the space of all such functions defined on S_T with finite norm which is as given for $\|\cdot\|_{q,r,Q}$ with Q, Ω, dx being replaced by S_T, S, ds respectively.

Throughout the paper, we make the following assumptions.

(A1) $a_{ij} \in L_2(Q)$ $(i, j = 1, \cdots, n)$.

(A2) There exist positive constants α_1 and α_2 such that, for all $\xi \equiv (\xi_1, \cdots, \xi_n) \in R^n$,

$$\alpha_1 \sum_{i=1}^{n} (\xi_i)^2 \leqq \sum_{i,j=1}^{n} a_{ij}(x, t)\xi_i\xi_j \leqq \alpha_2 \sum_{i=1}^{n} (\xi_i)^2,$$

uniformly on Q.

(A3) There exists a positive constant $M_1 \equiv M_1(q_1, r_1)$ such that

$$\left\| \sum_{i=1}^{n} (a_i)^2 \right\|_{q_1,r_1,Q} \leqq M_1, \quad \left\| \sum_{i=1}^{n} (b_i)^2 \right\|_{q_1,r_1,Q} \leqq M_1, \quad \|c\|_{q_1,r_1,Q} \leqq M_1,$$

and

$$\|c\|_{q_1,r_1,Q} \leqq M_1,$$

for a certain pair of constants q_1 and r_1 satisfying

$$\frac{1}{r_1} + \frac{n}{2q_1} = 1,$$

$$q_1 \in \left(\frac{n}{2}, \infty\right], \qquad r_1 \in [1, \infty) \quad \text{for } n \geqq 2,$$

$$q_1 \in [1, \infty], \qquad r_1 \in [1, 2] \quad \text{for } n = 1.$$

(A4) There exists a positive constant $M_2 \equiv M_2(q_2, r_2)$ such that

$$\|\sigma\|_{q_2,r_2,S_T} \leqq M_2,$$

where q_2, r_2 are subject to the requirements

$$\frac{1}{r_2} + \frac{n-1}{2q_2} = \frac{1}{2},$$

$$q_2 \in (2n-1, \infty], \qquad r_2 \in [2, \infty) \quad \text{for } n > 2,$$

$$q_2 \in (1, \infty), \qquad r_2 \in (2, \infty) \quad \text{for } r = 2,$$

$$q_2 = r_2 = 2 \qquad \text{for } n = 1.$$

(A5) $F_j(\cdot, \cdot, u_1(\cdot, \cdot)) \in L_2(Q)$ $(j = 1, 2, \cdots, n)$ for all $u_1 \in \mathcal{U}_1$.

(A6) $f(\cdot, \cdot, u_1(\cdot, \cdot)) \in L_{q_3,r_3}(Q)$, for all $u_1 \in \mathcal{U}_1$, where q_3 and r_3 are subject to the restrictions:

$$\frac{1}{r_3} + \frac{n}{2q_3} = 1 + \frac{n}{4},$$

$$q_3 \in \left[\frac{2n}{n+2}, 2\right], \qquad r_3 \in [1, 2] \quad \text{for } n \geqq 3,$$

$$q_3 \in (1, 2], \qquad r_3 \in [1, 2) \quad \text{for } n = 2,$$

$$q_3 \in [1, 2], \qquad r_3 \in [1, \tfrac{4}{3}] \quad \text{for } n = 1.$$

(A7) $\psi_0(\cdot, u_2(\cdot)) \in L_2(\Omega)$ for all $u_2 \in \mathcal{U}_2$.

(A8) $\psi(\cdot,\cdot,u_3(\cdot,\cdot))\in L_{q_4,r_4}(S_T)$ for all $u_3\in\mathcal{U}_3$, where q_4, r_4 are subject to the restrictions:

$$\frac{1}{r_4}+\frac{n-1}{2q_4}=\frac{n}{4}+\frac{1}{2},$$

$$q_4\in\left[\frac{2n-1}{n},\frac{2n-2}{n-2}\right], \qquad r_2\in[1,2] \quad \text{for } n>2,$$

$$q_4\in(1,\infty], \qquad r_4\in[1,2) \quad \text{for } n=2,$$

$$q_4=r_4=\tfrac{4}{3} \qquad \text{for } n=1.$$

For brevity, we introduce the following notation:

$$\mathscr{L}(\phi,\eta)(t)\equiv\int_\Omega\left\{\sum_{i=1}^n\left(\sum_{j=1}^n a_{ij}(x,t)\phi_{x_j}(x,t)+a_i(x,t)\phi(x,t)\right)\eta_{x_i}(x,t)\right.$$
$$\left.-\sum_{i=1}^n b_i(x,t)\phi_{x_i}(x,t)\eta(x,t)-c(x,t)\phi(x,t)\eta(x,t)\right\}dx, \tag{4}$$

$$\hat{F}(u)(x,t)\equiv\sum_{i=1}^n\left(F_i(x,t,u_1(x,t))\right)_{x_i}+f(x,t,u_1(x,t)) \tag{5}$$

and

$$(\hat{F}(u),\eta)(t)\equiv-\int_\Omega\left\{\sum_{i=1}^n F_i(x,t,u_1(x,t))\eta_{x_i}(x,t)+f(x,t,u_1(x,t))\eta(x,t)\right\}dx\,dt. \tag{6}$$

For the system (2), we introduce the following definition.

DEFINITION 3.1. For each $u\in\mathcal{U}$, a function $\phi(u)$ is said to be a weak solution of the second boundary value problem (2), if

(i) $\phi(u)\in V_2^{1,0}(Q)$,

(ii) for each $\tau\in[0,T]$,

$$\int_\Omega\phi(u)(x,\tau)\eta(x,\tau)\,dx-\int_0^\tau\int_\Omega\phi(u)(x,t)\eta_t(x,t)\,dx\,dt$$
$$+\int_0^\tau\left\{\mathscr{L}(\phi(u),\eta)(t)+(\hat{F}(u),\eta)(t)\right.$$
$$\left.+\int_S(\sigma(s,t)\phi(u)(s,t)-\psi(s,t,u_3(s,t)))\eta(s,t)\,ds\right\}dt$$
$$=\int_\Omega\psi_0(x,u_2(x))\eta(x,0)\,dx \tag{7}$$

for any $\eta\in W_2^{1,1}(Q)$.

LEMMA 3.1. *Consider the second boundary problem* (2). *Suppose that Assumptions* (A1) *to* (A8) *are satisfied. If* $\phi(u)$ *is a weak solution, then it satisfies the estimate*

$$\|\phi(u)\|_Q\leqq K_1\{\|F(\cdot,\cdot,u_1(\cdot,\cdot))\|_{2,Q}+\|f(\cdot,\cdot,u_1(\cdot,\cdot))\|_{q_3,r_3,Q}$$
$$+\|\psi_0(\cdot,u_2(\cdot))\|_{2,\Omega}+\|\psi(\cdot,\cdot,u_3(\cdot,\cdot))\|_{q_4,r_4,S_T}\}, \tag{8}$$

where

$$\|F(\cdot,\cdot,u_1(\cdot,\cdot))\|_{2,Q}\equiv\left(\int_Q\int\left\{\sum_{i=1}^n(F_i(x,t,u_1(x,t)))^2\right\}dx\,dt\right)^{1/2} \tag{9}$$

and the constant K_1 depends only on n, α_1, α_2, M_1, M_2 and the quantities q_1, r_1, q_2 and r_2.

The corresponding conclusion of the lemma is well known for the case involving the first (rather than the second) boundary value problem. Its proof is given in [8, § 2, Chap. 3, pp. 139–143]. By making some modifications, as suggested in [8, § 5, Chap. 3, pp. 169–170], in the arguments used in the proof for the case of the first boundary value problem, we can show that the estimate (8) is valid.

Remark 3.1. Consider the system (2) with f replaced by $\sum_{i=1}^{\nu} f^i$. Suppose that the Assumptions (A1) to (A5) and (A7) to (A8) are satisfied. Furthermore, we assume that $f^i \in L_{\hat{q}_i,\hat{r}_i}(Q)$ $(i = 1, \cdots, \nu)$, where for each $i = 1, \cdots, \nu$, the quantities $\hat{q}_i$ and $\hat{r}_i$ satisfy the same restrictions as those given in (A6) for the quantities q_3 and r_3. Then, by a similar approach as indicated in the statement made in the paragraph following Lemma 3.1, we can show that the estimate (8) with

$$\|f(\cdot, \cdot, u_1(\cdot, \cdot))\|_{q_3,r_3,Q}$$

replaced by

$$\sum_{i=1}^{\nu} \|f^i(\cdot, \cdot, u_1(\cdot, \cdot))\|_{\hat{q}_i,\hat{r}_i Q}$$

is valid.

The next lemma follows immediately from [8, Thm. 5.1, pp. 169–170].

LEMMA 3.2. *Consider the second boundary problem* (2). *Suppose Assumptions* (A1) *to* (A8) *are satisfied. Then, for each $u \in \mathcal{U}$, the problem admits a unique weak solution $\phi(u)$.*

Next, we shall introduce the adjoint system of the problem (P). To begin with, let

$$\nabla G_1(x, t, \hat{\phi}) \equiv \frac{\partial G_1(x, t, \phi)}{\partial \phi}\bigg|_{\phi=\hat{\phi}},$$

$$\nabla G_2(x, \hat{\phi}) \equiv \frac{\partial G_2(x, \phi)}{\partial \phi}\bigg|_{\phi=\hat{\phi}}$$

and

$$\nabla G_3(s, t, \hat{\phi}) \equiv \frac{\partial G_3(s, t, \phi)}{\partial \phi}\bigg|_{\phi=\hat{\phi}},$$

where G_i $(i = 1, 2, 3)$ are given in the definition of the cost functional J.

The following system is called the adjoint system:

$$\begin{aligned} &L^*Z(x, t) = \nabla G_1(x, t, \phi(u)(x, t)), \qquad (x, t) \in Q, \\ &Z(x, T) = \nabla G_2(x, \phi(u)(x, T)), \\ &\left[\frac{\partial Z}{\partial \nu_{L^*}} + \sigma Z\right]\bigg|_{S_T} = \nabla G_3(s, t, \phi(u)(s, t)), (s, t) \in S_T, \end{aligned} \tag{10}$$

where the operator L^* is defined by.

$$L^*\psi \equiv -\frac{\partial \psi}{\partial t} - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{n} a_{ij}(x, t)\psi_{x_j} - b_i(x, t)\psi\right) - \sum_{i=1}^{n} a_i(x, t)\psi_{x_i} - c(x, t)\psi, \tag{11}$$

and

$$\frac{\partial Z}{\partial \nu_{L^*}} \equiv \sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij}(s, t)Z_{x_j} - b_i(s, t)Z\right) \cos \alpha_i, \tag{12}$$

with α_i as defined for the system (2).

To proceed further, additional assumptions on the functions G_i $(i = 1, 2, 3)$ are required. These assumptions are given below.

(A9) (i) For each $u \in \mathcal{U}$, $\nabla G_1(\cdot, \cdot, \phi(u)(\cdot, \cdot)) \in L_{q_3, r_3}(Q)$, where q_3 and r_3 are defined in Assumption (A6).

(ii) For each $u \in \mathcal{U}$, $\nabla G_2(\cdot, \phi(u)(\cdot, T)) \in L_2(\Omega)$, and

(iii) For each $u \in \mathcal{U}$, $\nabla G_3(\cdot, \cdot, \phi(u)(\cdot, \cdot)) \in L_{q_4, r_3}(S_T)$, where q_4 and r_4 are defined in Assumption (A8).

Now we need to show that, for each $u \in \mathcal{U}$, the adjoint problem (10) has a unique weak solution in the sense of the following definition.

DEFINITION 3.2. For each $u \in \mathcal{U}$, a function $Z(u)$ is said to be a weak solution of the adjoint problem (10) if

(i) $Z(u) \in V_2^{1,0}(Q)$;

and

(ii) for each $\tau \in [0, T]$,

$$\begin{aligned} &\int_\Omega Z(u)(x, \tau)\eta(x, \tau)\, dx + \int_\tau^T \int_\Omega Z(u)(x, t)\eta_t(x, t)\, dx\, dt \\ &\quad + \int_\tau^T \Big\{ \mathscr{L}^*(Z(u), \eta)(t) + (\nabla G_1(u), \eta)(t) \\ &\quad + \int_S (\sigma(s, t)Z(u)(s, t) - \nabla G_3(s, t, \phi(u)(s, t)))\eta(s, t)\, ds \Big\} dt \\ &\qquad = \int_\Omega \nabla G_2(x, \phi(u)(x, T))\eta(x, T)\, dx, \end{aligned} \tag{13}$$

for any $\eta \in W_2^{1,1}(Q)$, where

$$\begin{aligned} \mathscr{L}^*(Z(u), \eta)(t) \equiv \int_\Omega \Big\{ &\sum_{i=1}^n \Big(\sum_{j=1}^n a_{ij} Z(u)_{x_j} - b_i Z(u) \Big) \eta_{x_i} \\ &+ \sum_{i=1}^n a_i Z(u)_{x_i} \eta - cZ(u)\eta \Big\} dx, \end{aligned} \tag{14}$$

and

$$(\nabla G_1(u), \eta)(t) \equiv -\int_\Omega \nabla G_1(x, t, \phi(u)(x, t))\eta(x, t)\, dx. \tag{15}$$

LEMMA 3.3. *Consider the adjoint problem* (10). *Suppose that Assumptions* (A1) *to* (A4) *and* (A9) *are satisfied. Then, for each* $u \in \mathcal{U}$, *the adjoint problem admits a unique weak solution* $Z(u)$ *which satisfies the estimate*

$$\begin{aligned} \|Z(u)\|_Q \leqq K_2 \{ &\|\nabla G_1(\cdot, \cdot, \phi(u)(\cdot, \cdot))\|_{q_3, r_3, Q} \\ &+ \|\nabla G_2(\cdot, \phi(u)(\cdot, T))\|_{2, \Omega} + \|\nabla G_3(\cdot, \cdot, \phi(u)(\cdot, \cdot))\|_{q_4, r_4, S_T} \}, \end{aligned}$$

where the constant K_2 *depends only on* n, α_1, α_2, M_1, M_2, T *and the quantities* q_1, q_2, r_1 *and* r_2.

Proof. Letting $t' = T - t$ and then setting $Z(u)(x, T - t') \equiv \hat{Z}(u)(x, t')$, the adjoint system (10) can be reduced to the one involving the function $\hat{Z}(u)(x, t')$. This reduced system is in the form of the system (2). Furthermore, it can be verified that all the assumptions required by Lemma 3.1 and Lemma 3.2 are satisfied. Thus, from the same lemmas and the definition of $\hat{Z}(u)(x, t')$, we obtain the conclusion of the lemma.

In our later analysis, we need to smooth the coefficients and data of the adjoint system (10) so that it admits classical solutions. For this, we shall adopt the following convention:

$$a_{ii}(x,t)\equiv 1, \qquad i=1,2,\cdots,n,$$
$$a_{ij}(x,t)\equiv 0, \qquad i\neq j,\quad i,j=1,2,\cdots,n,$$
$$a_i(x,t)\equiv b_i(x,t)\equiv 0, \qquad i=1,2,\cdots,n,$$

for all $(x,t)\in R^{n+1}\backslash Q$;

$$\nabla G_1(x,t,\phi(u)(x,t))\equiv 0$$

for all $u\in\mathcal{U}$ and for all $(x,t)\in R^{n+1}\backslash Q$, and

$$\nabla G_2(x,\phi(u)(x,T))\equiv 0$$

for all $u\in\mathcal{U}$ and for all $x\in R^n\backslash\Omega$.

For each $u\in\mathcal{U}$, let $\nabla G_1(u)$ and $\nabla G_2(u)$ denote, respectively, $\nabla G_1(\cdot,\cdot,\phi(u)(\cdot,\cdot))$ and $\nabla G_2(\cdot,\phi(u)(\cdot,T))$, and let α denote any of the coefficients a_{ij} $(i,j=1,\cdots,n)$, a_i, b_i $(i=1,\cdots,n)$ and c. Furthermore, let $\nabla G_1^k(u)$ and α^k be, respectively, the integral averages of $\nabla G_1(u)$ and α with a kernel whose support lies in $\{(x,t)\in R^{n+1}: \sum_{i=1}^n (x_i)^2+(t)^2\leqq 1/(k)^2\}$. Similarly, let $\nabla G_2^k(u)$ be the integral average of $\nabla G_2(u)$ with a kernel whose support lies in $\{x\in R^n: \sum_{i=1}^n (x_i)^2\leqq 1/(k)^2\}$.

Remark 3.2. By the well-known properties of integral averages, we conclude that a_{ij}^k $(i,j=1,\cdots,n)$, a_i^k, b_i^k $(i=1,\cdots,n)$, c^k, $\nabla G_1^k(u)$ and $\nabla G_2^k(u)$ converge, respectively, to a_{ij} $(i,j=1,\cdots,n)$, a_i, b_i $(i=1,\cdots,n)$, c, $\nabla G_1(u)$ and $\nabla G_2(u)$ in the norms of the spaces to which they belong.

Remark 3.3. Since Ω has uniform C^3-regularity property, it follows that $S\in H^{l+2}$ for $0<l<1$. (For the definition of H^{l+2}, see [8, p. 10].) Thus, there exists a sequence of functions $\{\nabla G_3^k(u)\}\subset H^{l+1,(l+1)/2}(S_T)$. (For the definition of $H^{l+1,(l+1)/2}(S_T)$, see [8, Chap. II, (3.19), p. 82].) and a sequence of function $\{\sigma^k\}\subset H^{l+1,(l+1)/2}(S_T)$ such that, as $k\to\infty$, $\nabla G_3^k(u)\to\nabla G_3(u)(\equiv\nabla G_3(\cdot,\cdot,\phi(u)(\cdot,\cdot)))$ in the norm of $L_{q_4,r_4}(S_T)$ and $\sigma^k\to\sigma$ in the norm of $L_{q_2,r_2}(S_T)$.

Let $\{\Omega^k\}$ be a sequence of open connected sets with sufficiently smooth boundaries such that $\overline{\Omega^k}\subset\Omega^{k+1}\subset\overline{\Omega^{k+1}}\subset\Omega$ for all integers $k\geqq 1$ and $\lim_{k\to\infty}\Omega^k=\Omega$. For each $k\geqq 1$, let d_k be an element in $C_0^\infty(\bar{\Omega})$ so that $d_k(x)=1$ on $\overline{\Omega^k}$ and $0\leqq d_k(x)\leqq 1$ on $\bar{\Omega}\backslash\Omega^k$. Again, let $\{I^k\}$ be a sequence of open intervals such that $\overline{I^k}\subset I^{k+1}\subset\overline{I^{k+1}}\subset(0,T)$ for all integers $k\geqq 1$ and $\lim_{k\to\infty} I^k=(0,T)$. For each $k\geqq 1$, let $\tilde{d}_k$ be an element in $C_0^\infty([0,T])$ so that $\tilde{d}_k(t)=1$ on I^k and $0\leqq\tilde{d}_k(t)\leqq 1$ on $[0,T]\backslash I^k$.

We now consider the following sequence of second boundary value problems

$$L^{*,k}Z^k(x,t)=\nabla G_1^k(u)(x,t), \tag{16a}$$
$$Z^k(x,T)|_{\bar{\Omega}}=\nabla G_2^k(u)(x)d_k(x), \tag{16b}$$
$$\left[\frac{\partial}{\partial\nu_{L^{*,k}}}Z^k+\sigma^k(s,t)Z^k\right]\Big|_{S_T}=\nabla G_3^k(u)(s,t)\tilde{d}_k(t), \tag{16c}$$

where, for each k, the operators $L^{*,k}$ and $\partial/\partial\nu_{L^{*,k}}$ are as defined by L^* and $\partial/\partial\nu_{L^*}$, in (11) and (12), with a_{ij}, a_i, b_i, c replaced, respectively, by the corresponding integral averages.

It is clear that the system (16) satisfies all the assumptions as required in [8, Thm. 5.3, pp. 320–321]. Thus, it follows from the same theorem that the system (16) admits, for each k, a unique classical solution $Z^k(u)$, that is, there exists a unique

function $Z^k(u)$ which fulfills the following conditions:

(i) $Z^k(u)$ satisfies (16a) everywhere in Q, (16b) everywhere in Ω and (16c) everywhere in S_T, and

(ii) $Z^k(u)$ and $Z^k(u)_{x_i}$ $(i=1,\cdots,n)$ are continuous on $\bar{Q}$, while $Z^k(u)_t$ and $Z^k(u)_{x_ix_j}$ $(i,j=1,\cdots,n)$ are continuous on Q.

Remark 3.4. Note that, for each k, the classical solution of the system (16) is also a weak solution of the same system.

In the next theorem, we shall show that the sequence of the classical solutions $\{Z^k(u)\}$ converges, in the norm of $V_2^{1,0}(Q)$, to the weak solution $Z(u)$ of the system (10).

LEMMA 3.4. *Suppose that the assumptions* (A1) *to* (A4) *and* (A9) *are satisfied. Then, for each* $u\in\mathcal{U}$, $Z^k(u)\to Z(u)$ *in the norm of* $V_2^{1,0}(Q)$ *as* $k\to\infty$.

Proof. In view of Remark 3.4, the classical solutions $Z^k(u)$ of the system (16) are also weak solutions. Thus, it follows from Definition 3.2 that, for each $\tau\in[0,T]$,

$$\begin{aligned}&\int_\Omega Z^k(u)(x,\tau)\eta(x,\tau)\,dx+\int_\tau^T\int_\Omega Z^k(u)(x,t)\eta_t(x,t)\,dx\,dt\\&\quad+\int_\tau^T\Big\{\mathscr{L}^{*,k}(Z^k(u),\eta)(t)+(\nabla G_1^k(u),\eta)(t)\\&\quad+\int_S(\sigma^k(s,t)Z^k(u)(s,t)-\nabla G_3^k(u)(s,t)\tilde{d}_k(t))\eta(s,t)\,ds\Big\}\,dt\\&\qquad=\int_\Omega\nabla G_2^k(u)(x)\,d_k(x)\eta(x,T)\,dx\end{aligned}\tag{17}$$

for any $\eta\in W_2^{1,1}(Q)$, where $\mathscr{L}^{*,k}(Z^k(u),\eta)$ is as defined by $\mathscr{L}^*(Z(u),\eta)(t)$, in (14), with a_{ij}, a_i, b_i, c and $Z(u)$ replaced, respectively, by the corresponding integral averages a_{ij}^k, a_i^k, b_i^k, c^k and $Z^k(u)$, while

$$(\nabla G_1^k(u),\eta)(t)\equiv-\int_\Omega\nabla G_1^k(u)(x,t)\eta(x,t)\,dx.$$

Note that $Z(u)$ is the weak solution of the system (10). Hence it must satisfy (13). Thus, by subtracting (17) from (13) and then setting $Z^k(u)-Z(u)\equiv\hat{Z}^k(u)$, we obtain

$$\begin{aligned}&\int_\Omega\hat{Z}^k(u)(x,\tau)\eta(x,\tau)\,dx+\int_\tau^T\int_\Omega\hat{Z}^k(u)(x,t)\eta_t(x,t)\,dx\,dt\\&\quad+\int_\tau^T\Big\{\mathscr{L}^{*,k}(\hat{Z}^k(u),\eta)(t)+(\hat{G}^k(u),\eta)(t)\\&\quad+\int_S[\sigma^k\hat{Z}^k(u)-(\sigma-\sigma^k)Z(u)-(\nabla G_3^k(u)\tilde{d}_k-\nabla G_3(u))]\eta\,ds\Big\}\,dt\\&\qquad=\int_\Omega(\nabla G_2^k(u)d_k-\nabla G_2(u))\eta(x,T)\,dx,\end{aligned}\tag{18}$$

where $\mathscr{L}^{*,k}(\hat{Z}^k(u),\eta)(t)$ is as defined by $\mathscr{L}^{*,k}(Z(u),\eta)(t)$ with $Z(u)$ replaced by $\hat{Z}^k(u)$, while

$$\begin{aligned}(\hat{G}^k(u),\eta)(t)\equiv\int_\Omega\Big\{&\sum_{i=1}^n\tilde{g}_i^k(x,t)\eta_{x_i}(x,t)-[\tilde{g}^k(x,t)+(\nabla G_1^k(u)(x,t)\\&-\nabla G_1(u)(x,t))]\eta(x,t)\Big\}\,dx,\end{aligned}\tag{19}$$

with

$$\tilde{g}_i^k(x,t) \equiv \sum_{j=1}^{n} (a_{ij}^k(x,t) - a_{ij}(x,t))Z(u)_{x_i}(x,t) - (b_i^k(x,t) - b_i(x,t))Z(u)(x,t), \tag{20}$$

and

$$\tilde{g}^k(x,t) \equiv \sum_{i=1}^{n} (a_i^k(x,t) - a_i(x,t))Z(u)_{x_i}(x,t) - (c^k(x,t) - c(x,t))Z(u)(x,t). \tag{21}$$

By setting $t' \equiv T - t$, (18) can be reduced to the form of (7). Thus, from Lemma 3.1 and Remark 3.1, we have

$$\begin{aligned} \|\hat{Z}^k(u)\|_Q \leqq K\Big\{ & \Big(\int\int_Q \Big[\sum_{i=1}^{n} (\tilde{g}_i^k(x,t))^2 \Big] dx\, dt \Big)^{1/2} \\ & + \|\tilde{g}^k\|_{\lambda_1,\mu_1,Q} + \|\nabla G_1^k(u) - \nabla G_1(u)\|_{q_3,r_3,Q} \\ & + \|\nabla G_2^k(u) d_k - \nabla G_2(u)\|_{2,\Omega} + \|(\sigma^k - \sigma)Z(u)\|_{\lambda_2,\mu_2,S_T} \\ & + \|\nabla G_3^k(u)\tilde{d}_k - \nabla G_3(u)\|_{q_4,r_4,S_T} \Big\} \end{aligned} \tag{22}$$

where $\lambda_i \equiv 2q_i/(q_i+1)$, $\mu_i \equiv 2r_i/(r_i+1)$ $(i = 1, 2)$.

To complete the proof, it remains to show that

$$\|\hat{Z}^k(u)\|_Q \to 0 \quad \text{as } k \to \infty.$$

From Minkowski's inequality and (20), we obtain

$$\begin{aligned} & \Big(\int\int_Q \Big| \sum_{i=1}^{n} (\tilde{g}_i^k(x,t))^2 \Big| dx\, dt \Big)^{1/2} \\ & \quad \leqq \sum_{i=1}^{n}\sum_{j=1}^{n} \|(a_{ij}^k - a_{ij})Z(u)_{x_j}\|_{2,Q} + \sum_{i=1}^{n} \|(b_i^k - b_i)Z(u)\|_{2,Q}. \end{aligned} \tag{23}$$

Using Cauchy's inequality, it follows that

$$\begin{aligned} \sum_{i,j=1}^{n} \|(a_{ij}^k - a_{ij})Z(u)_{x_j}\|_{2,Q} & \leqq \sum_{i,j=1}^{n} \{\|a_{ij}^k - a_{ij}\|_{2,Q} \|Z(u)_{x_j}\|_{2,Q}\} \\ & \leqq \sum_{i,j=1}^{n} \{\|a_{ij}^k - a_{ij}\|_{2,Q} \|Z(u)\|_Q\}. \end{aligned} \tag{24}$$

From Hölder's inequality and [8, (3.8), p. 77], we get

$$\begin{aligned} \sum_{i=1}^{n} \|(b_i^k - b_i)Z(u)\|_{2,Q} & \leqq \sum_{i=1}^{n} \{(\|(b_i^k - b_i)^2\|_{q_1,r_1,Q})^{1/2} \|Z(u)\|_{\bar{\lambda}_1,\bar{\mu}_1,Q}\} \\ & \leqq \sum_{i=1}^{n} \{(\|(b_i^k - b_i)^2\|_{q_1,r_1,Q})^{1/2} (\hat{C}\|Z(u)\|_Q)\} \end{aligned} \tag{25}$$

where the constant $\hat{C}$ is as defined for [8, (3.8), p. 77], and $\bar{\lambda}_1 \equiv 2q_1/(q_1-1)$, $\bar{\mu}_1 \equiv 2r_1/(r_1-1)$.

Combining (23) to (25) and then using Assumptions (A2) and (A3) and Remark 3.1, we obtain

$$\begin{aligned} \Big(\int_Q \int \Big[\sum_{i=1}^{n} \|(\tilde{g}_i^k(x,t))^2 \Big] dx\, dt \Big)^{1/2} & \leqq \sum_{i,j=1}^{n} \{\|a_{ij}^k - a_{ij}\|_{2,Q} \|Z(u)\|_{2,Q}\} \\ & \quad + \sum_{i=1}^{n} \{(\|(b_i^k - b_i)^2\|_{q_1,r_1,0})^{1/2} (\hat{C}\|Z(u)\|_Q)\} \to 0 \end{aligned} \tag{26}$$

as $k \to \infty$.

For $\tilde{g}^k$, it follows from (21) that

$$\|\tilde{g}^k\|_{\lambda_1,\mu_1,0} \leqq \sum_{i=1}^{n} \|(a_i^k - a_i)Z(u)_{x_i}\|_{\lambda_1,\mu_1,Q} + \|(c^k - c)Z(u)\|_{\lambda_1,\mu_1,Q} \tag{27}$$

where $\lambda_1 \equiv 2q_1/(q_1+1)$ and $\mu_1 \equiv 2r_1/(r_1+1)$.

Now, we are required to show that the right-hand side of the above inequality tends to zero as $k \to \infty$. From Hölder's inequality and Assumption (A3), we have

$$\begin{aligned} \sum_{i=1}^{n} \|(a_i^k - a_i)Z(u)_{x_i}\|_{\lambda_1,\mu_1,Q} &\leqq \sum_{i=1}^{n} (\|(a_i^k - a_i)^2\|_{q_1,r_1,Q})^{1/2} \|Z(u)_{x_i}\|_{2,Q} \\ &\leqq \sum_{i=1}^{n} (\|(a_i^k - a_i)^2\|_{q_1,r_1,Q})^{1/2} \|Z(u)\|_Q \to 0 \quad \text{as } k \to 0. \end{aligned} \tag{28}$$

On the other hand, it follows from Hölder's inequality, [8, (3.8), p. 77] and (A3) that

$$\begin{aligned} \|(c^k - c)Z(u)\|_{\lambda_1,\mu_1,Q} &\leqq \|c^k - c\|_{q_1,r_1,Q} \|Z(u)\|_{\bar{\lambda}_1,\bar{\mu}_1,Q} \\ &\leqq \|c^k - c\|_{q_1,r_1,Q} (\hat{C}\|Z(u)\|_Q) \to 0 \end{aligned} \tag{29}$$

where $\bar{\lambda}_1 \equiv 2q_1/(q_1-1)$ and $\bar{\mu}_1 = 2r_1(r_1-1)$.

Combining (27) to (29), we obtain

$$\|\tilde{g}^k\|_{\lambda_1,\mu_1,Q} \to 0 \quad \text{as } k \to \infty. \tag{30}$$

Similarly, we have

$$\|(\sigma^k - \sigma)Z(u)\|_{\lambda_2,\mu_2,S_T} \leqq \|\sigma^k - \sigma\|_{q_2,r_2,S_T} \|Z(u)\|_{\bar{\lambda}_2,\bar{\mu}_2,S_T} \tag{31}$$

where $\lambda_2 \equiv 2q_2/(q_2+1)$, $\mu_2 \equiv 2r_2/(r_2+1)$. $\bar{\lambda}_2 = 2q_2/(q_2-1)$ and $\bar{\mu}_2 = 2r_2/(r_2-1)$. Note that $\bar{\lambda}_2$ together with $\bar{\mu}_2$ satisfies the conditions:

$$\frac{1}{\bar{\mu}_2} + \frac{n-1}{2\bar{\lambda}_2} = \frac{n}{4},$$

$$\begin{aligned} &\bar{\mu}_2 \in [2, \infty], \quad \bar{\lambda}_2 \in \left[\frac{2(n-1)}{n}, \frac{2(n-1)}{n-2}\right] && \text{for } n \geqq 3, \\ &\bar{\mu}_2 \in (2, \infty], \quad \bar{\lambda} \in [1, \infty) && \text{for } n = 2, \\ &\bar{\lambda}_2 = \bar{\mu}_2 = 4 && \text{for } n = 1. \end{aligned}$$

Thus, by making use of [8, (3.11), p. 78], we obtain

$$\|Z(u)\|_{\bar{\lambda}_2,\bar{\mu}_2,S_T} \leqq \tilde{C}\|Z(u)\|_Q, \tag{32}$$

where $\tilde{C}$ is a constant as defined for [8, (3.11)–(3.12), p. 78]. Hence,

$$\|(\sigma^k - \sigma)Z(u)\|_{\lambda_2,\mu_2,S_T} \leqq \|\sigma^k - \sigma\|_{q_2,r_2,S_T} (\tilde{C}\|Z(u)\|_Q) \to 0 \tag{33}$$

as $k \to \infty$.

The other three terms in the right-hand side of (22) also approach zero as $k \to \infty$, by Remarks 3.2 and 3.3 and the properties of the functions d_k and $\tilde{d}_k$. Thus, we conclude that

$$\|\hat{Z}^k(u)\|_Q \to 0 \quad \text{as } k \to 0.$$

This completes the proof of Lemma 3.4.

In the sequel, we need to consider another sequence of second boundary value problems:

$$
\begin{aligned}
&L^k\phi^k(x,t)=\sum_{j=1}^{n}\frac{\partial}{\partial x_j}[F_j(x,t,u_1(x,t))]+f(x,t,u_1(x,t)),\\
&\phi^k(x,t)|_{t=0}=\psi_0(x,u_2(x)),\\
&\left[\frac{\partial\phi^k}{\partial\nu_{L^k}}+\sigma^k(x,t)\phi^k\right]\Big|_{S_T}=\psi(x,t,u_3(x,t)),
\end{aligned}
\tag{34}
$$

where, for each k, the operators L^k and $\partial/\partial\nu_{L^k}$ are as defined, respectively, by L and $\partial/\partial\nu$, in system (2), with the coefficients a_{ij}, a_i, b_i, c replaced by the corresponding integral averages a_{ij}^k, a_i^k, b_i^k, c^k, respectively.

In view of Lemma 3.2, we observe that, for each k, the system (34) admits a unique weak solution $\phi^k(u)$ (in the sense of Definition 3.1). Furthermore, by using the same approach as that given for Lemma 3.4, we can show that, for each $u\in\mathcal{U}$,

$$
\phi^k(u)\to\phi(u) \tag{35}
$$

in the norm of $V_2^{1,0}(Q)$ as $k\to\infty$, where $\phi(u)$ is the weak solution of the second boundary problem (2).

4. A basic inequality. In this section, we shall derive a basic inequality which will then be used to derive a sufficient condition for optimality for the problem (P).

To begin, we need to impose the following additional assumptions on the functions G_i $(i=1,2,3)$ and H_i $(i=1,2,3)$. (These functions are given in the definition of the cost functional J.)

(A10) For each $u\equiv(u_1,u_2,u_3)\in\mathcal{U}$, $G(\cdot,\cdot,\phi(u)(\cdot,\cdot))\in L_1(Q)$ and $H_1(\cdot,\cdot,u_1(\cdot,\cdot))\in L_1(Q)$.

(A11) For each $u\equiv(u_1,u_2,u_3)\in\mathcal{U}$, $G_2(\cdot,\phi(u)(\cdot,T))\in L_1(\Omega)$ and $H_2(\cdot,u_2(\cdot))\in L_1(\Omega)$.

(A12) For each $u\equiv(u_1,u_2,u_3)\in\mathcal{U}$, $G_3(\cdot,\cdot,\phi(u)(\cdot,\cdot))\in L_1(S_T)$ and $H_3(\cdot,\cdot,u_3(\cdot,\cdot))\in L_1(S_T)$.

LEMMA 4.1. *Consider the problem* (P). *Let* $u^0\equiv(u_1^0,u_2^0,u_3^0)\in\mathcal{U}$ *be an admissible control and let* $Z(u^0)$ *be the weak solution of the adjoint system* (10) *with* u *replaced by* u^0. *Then*

$$
\begin{aligned}
J(u)-J(u^0)\geqq\int_Q\int\Big\{&\sum_{i=1}^{n}(F_i(x,t,u_1(x,t))-F_i(x,t,u_1^0(x,t)))Z(u_1^0)_{x_i}(x,t)\\
&+(f(x,t,u_1(x,t))-f(x,t,u_1^0(x,t)))Z(u^0)(x,t)\\
&+(H_1(x,t,u_1(x,t))-H_1(x,t,u^0(x,t)))\Big\}\,dx\,dt\\
+\int_\Omega\{&(\psi_0(x,u_2(x)))-\psi_0(x,u_2^0(x)))Z(u^0)(x,0)\\
&+(H_2(x,u_2(x))-H_2(x,u_2^0(x)))\}\,dx\\
+\int_{S_T}\int\{&(\psi(s,t,u_3(s,t))-\psi(s,t,u_3^0(s,t)))Z(u^0)(s,t)\\
&+(H_3(s,t,u_3(s,t))-H_3(s,t,u_3^0(s,t)))\}\,ds\,dt
\end{aligned}
\tag{36}
$$

for all $u\in\mathcal{U}$.

Proof. Let

$$\Delta H(u, u^0) \equiv \int_Q \int (H_1(x, t, u_1(x, t)) - H_1(x, t, u_1^0(x, t)))\, dx\, dt$$
$$+ \int_\Omega (H_2(x, u_2(x)) - H_2(x, u_2^0(x)))\, dx \tag{37}$$
$$+ \int_{S_T} \int (H_3(s, t, u_3(s, t)) - H_3(s, t, u_3^0(s, t)))\, ds\, dt.$$

By the convexity properties of the functions G_i $(i = 1, 2, 3)$, we obtain

$$J(u) - J(u^0) \geqq \int_Q \int \nabla G_1(u^0)(x, t)(\phi(u)(x, t) - \phi(u^0)(x, t))\, dx\, dt$$
$$+ \int_\Omega \nabla G_2(u^0)(x)(\phi(u)(x, T) - \phi(u^0)(x, T))\, dx \tag{38}$$
$$+ \int_{S_T} \int \nabla G_3(u^0)(s, t)(\phi(u)(s, t) - \phi(u^0)(s, t))\, ds\, dt + \Delta H(u, u^0).$$

From Remarks 3.2 and 3.3, we recall that $\nabla G_i^k(u^0)$ $(i = 1, 2, 3)$ converge, respectively, to $\nabla G_i(u^0)$ $(i = 1, 2, 3)$ in the norms of the spaces to which they belong. Thus, by virtue of (35), it follows from inequality (38) that

$$J(u) - J(u^0) \geqq \lim_{k \to \infty} \left\{ \int_Q \int \nabla G_1^k(u^0)(x, t)(\phi^k(u)(x, t) - \phi^k(u^0)(x, t))\, dx\, dt \right.$$
$$+ \int_\Omega \nabla G_2^k(u^0)(x)(\phi^k(u)(x, T) - \phi^k(u^0)(x, T))\, dx\, dt \tag{39}$$
$$\left. + \int_{S_T} \int \nabla G_3^k(u^0)(s, t)(\phi^k(u)(s, t) - \phi^k(u^0)(s, t))\, ds\, dt \right\} + \Delta H(u, u^0)$$

In view of (16), the above inequality can be written as

$$J(u) - J(u^0) \geqq \lim_{k \to \infty} \left\{ \int_Q \int L^{*,k} Z^k(u^0)(x, t)(\phi^k(u)(x, t) - \phi^k(u^0)(x, t))\, dx\, dt \right.$$
$$+ \int_\Omega Z^k(u^0)(x, T)(\phi^k(u)(x, T) - \phi^k(u^0)(x, T))\, dx \tag{40}$$
$$+ \int_{S_T} \int \left(\frac{\partial}{\partial \nu_{L^{*,k}}} Z^k(u^0)(s, t) + \sigma^k(s, t) Z^k(u^0)(s, t) \right)$$
$$\left. \cdot (\phi^k(u)(s, t) - \phi^k(u^0)(s, t))\, ds\, dt \right\} + \Delta H(u, u^0).$$

Since $Z^k(u^0) \in W_2^{1,1}(Q)$ and $\phi^k(u)$ is the weak solution of the system (34), corresponding to $u \in \mathcal{U}$, it follows from Definition 3.1 that

$$\int_\Omega \phi^k(u)(x, T) Z^k(u^0)(x, T) dx - \int_0^T \int_\Omega \phi^k(u)(x, t) Z^k(u^0)_t(x, t)\, dx\, dt$$
$$+ \int_0^T \left\{ \mathcal{L}^k(\phi^k(u), Z^k(u^0))(t) + (\hat{F}(u), Z^k(u^0))(t) \right.$$
$$\left. + \int_S (\sigma^k(s, t)\phi^k(u)(s, t) - \psi(s, t, u_3(s, t)) Z^k(u^0)(s, t)\, ds \right\} dt \tag{41}$$
$$= \int_\Omega \psi_0(x, u_2(x)) Z^k(u^0)(x, 0)\, dx,$$

where $\mathscr{L}^k(\phi^k(u); Z^k(u^0))$ (resp. $(\hat{F}(u), Z^k(u^0))$) is as defined by (4) (resp. (5)), with a_{ij}, a_i, b_i, c, ϕ and η replaced, respectively, by a_{ij}^k, a_i^k, b_i^k, c^k, $\phi^k(u)$ and $Z^k(u^0)$.

But, as a result of integration by parts with respect to x_i in the appropriate terms, we have

$$
\begin{aligned}
&\int_0^T \mathscr{L}^k(\phi^k(u), Z^k(u^0))(t)\,dt \\
&\quad = \int_0^T \int_\Omega \Big\{ -\sum_{i=1}^n \frac{\partial}{\partial x_i}\Big(\sum_{j=1}^n a_{ij}^k Z^k(u^0)_{x_j} - b_i^k z^k(u^0)\Big) \\
&\qquad\qquad - \sum_{i=1}^n a_i^k Z^k(u^0)_{x_i} - c^k Z^k(u^0)\Big\} \phi^k(u)\,dx\,dt \\
&\quad + \int_0^T \Big\{ \int_S \Big[\sum_{i=1}^n \Big(\sum_{j=1}^n (a_{ij}^k Z^k(u^0)_{x_j} - b_i^k Z^k(u^0))\Big) \cos\alpha_i \Big] \phi^k(u)(s,t)\,ds \Big\}\,dt \\
&\quad = \int_Q \int \Big(L^{*,k} Z^k(u^0) + \frac{\partial}{\partial t} Z^k(u^0)\Big) \phi^k(u)\,dx\,dt \\
&\quad + \int_{S_T} \int \Big(\frac{\partial}{\partial \nu_{L^{*,k}}} Z^k(u^0)\Big) \phi^k(u)\,ds\,dt.
\end{aligned}
\tag{42}
$$

By virtue of (42) and (6), the inequality (41) can be reduced to

$$
\begin{aligned}
&\int_\Omega \phi^k(u)(x,T) Z^k(u^0)(x,T)\,dx + \int_Q \int (L^{*,k} Z^k(u^0)) \phi^k(u)\,dx\,dt \\
&\quad + \int_{S_T} \int \Big(\frac{\partial}{\partial \nu_{L^{*,k}}} Z^k(u^0) + \sigma^k Z^k(u^0)\Big) \phi^k(u)\,ds\,dt \\
&\quad = \int_Q \int \Big\{ \sum_{i=1}^n F_i(x,t,u_1(x,t)) Z^k(u_0)_{x_i}(x,t) + f(x,t,u_1(x,t)) Z^k(u^0)(x,t)\Big\}\,dx\,dt \\
&\quad + \int_\Omega \psi_0(x,u_2(x)) Z^k(u^0)(x,0)\,dx + \int_{S_T} \int \psi(s,t,u_3(s,t)) Z^k(u^0)\,ds\,dt.
\end{aligned}
\tag{43}
$$

Using (43), it follows from (40) that

$$
\begin{aligned}
&J(u) - J(u^0) \\
&\geqq \lim_{k\to\infty} \Big\{ \int \int_Q \Big[\sum_{i=1}^n (F_i(x,t,u_1(x,t)) - F_i(x,t,u_1^0(x,t))) Z^k(u^0)_{x_i}(x,t) \\
&\qquad + (f(x,t,u_1(x,t)) - f(x,t,u_1^0(x,t))) Z^k(u^0)(x,t)\Big]\,dx\,dt \\
&\quad + \int_\Omega (\psi_0(x,u_2(x)) - \psi_0(x,u_2^0(x))) Z^k(u^0)(x,0)\,dx \\
&\qquad + \int_{S_T} \int (\psi(s,t,u_3(s,t)) - \psi(s,t,u_3^0(s,t))) Z^k(u^0)(s,t)\,ds\,dt\Big\} + \Delta H(u,u^0).
\end{aligned}
\tag{44}
$$

By Cauchy's inequality, we have

$$\begin{aligned}
&\left| \int_Q \int \left\{ \sum_{i=1}^{n} (F_i(x, t, u_1(x, t)) - F_i(x, t, u_1^0(x, t))) \right.\right. \\
&\qquad \left.\left. \cdot (Z^k(u^0)_{x_i}(x, t) - Z(u^0)_{x_i}(x, t)) \right\} dx\, dt \right| \\
&\qquad \leqq \sum_{i=1}^{n} \|F_i(\cdot, \cdot, u_1(\cdot, \cdot)) - F_i(\cdot, \cdot, u_1^0(\cdot, \cdot))\|_{2,Q} \|Z^k(u^0)_{x_i} - Z(u^0)_{x_i}\|_{2,Q} \\
&\qquad \leqq \sum_{i=1}^{n} \|F_i(\cdot, \cdot, u_1(\cdot, \cdot)) - F_i(\cdot, \cdot, u_1^0(\cdot, \cdot))\|_{2,Q} \|Z^k(u^0) - Z(u^0)\|_Q.
\end{aligned} \tag{45}$$

Again from Cauchy's inequality, it follows that

$$\begin{aligned}
&\left| \int_\Omega (\psi_0(x, u_2(x)) - \psi_0(x, u_2^0(x)))(Z^k(u^0)(x, 0) - Z(u^0)(x, 0))\, dx \right| \\
&\qquad \leqq \|\psi_0(\cdot, u_2(\cdot)) - \psi_0(\cdot, u_2^0(\cdot))\|_{2,\Omega} \|Z^k(u^0)(\cdot, 0) - Z(u^0)(\cdot, 0)\|_{2,\Omega} \\
&\qquad \leqq \|\psi_0(\cdot, u_2(\cdot)) - \psi_0(\cdot, u_2^0(\cdot))\|_{2,\Omega} \|Z^k(u^0) - Z(u^0)\|_Q.
\end{aligned} \tag{46}$$

By Hölder's inequality and [8, (3.8), p. 77], we obtain

$$\begin{aligned}
&\left| \int_Q \int (f(x, t, u_1(x, t)) - f(x, t, u_1^0(x, t)))(Z^k(u^0)(x, t) - Z(u^0)(x, t))\, dx\, dt \right| \\
&\qquad \leqq \|f(\cdot, \cdot, u_1(\cdot, \cdot)) - f(\cdot, \cdot, u_1^0(\cdot, \cdot))\|_{q_3, r_3, Q} \|Z^k(u^0) - Z(u^0)\|_{\bar{q}_3, \bar{r}_3, Q} \\
&\qquad \leqq \|f(\cdot, \cdot, u_1(\cdot, \cdot)) - f(\cdot, \cdot, u_1^0(\cdot, \cdot))\|_{q_3, r_3, Q} (\hat{C} \|Z^k(u^0) - Z(u^0)\|_Q)
\end{aligned} \tag{47}$$

where $q_3 \equiv q_3/(q_3 - 1)$ and $\bar{r}_3 \equiv r_3/(r_3 - 1)$.

Again, by Holder's inequality and [8, (3.11), p. 78], we have

$$\begin{aligned}
&\left| \int_{S_T} \int (\psi(s, t, u_3(s, t)) - \psi(s, t, u_3^0(s, t)))(Z^k(u^0)(s, t) - Z(u^0)(s, t))\, ds\, dt \right| \\
&\qquad \leqq \|\psi(\cdot, \cdot, u_3(\cdot, \cdot)) - \psi(\cdot, \cdot, u_3^0(\cdot, \cdot))\|_{q_4, r_4, S_T} \|Z^k(u^0) - Z(u^0)\|_{\bar{q}_4, \bar{r}_4, S_T} \\
&\qquad \leqq \|\psi(\cdot, \cdot, u_3(\cdot, \cdot)) - \psi(\cdot, \cdot, u_3^0(\cdot, \cdot))\|_{q_4, r_4, S_T} (\tilde{C} \|Z^k(u^0) - Z(u^0)\|_Q)
\end{aligned} \tag{48}$$

where $\bar{q}_4 = q_4/(q_4 - 1)$ and $\bar{r}_4 = r_4/(r_4 - 1)$.

Combining (45) to (48), and then using Lemma 3.4, we see that the limit of the right-hand side of the inequality (44) exists and is equal to the right-hand side of the inequality (36). This completes the proof.

Remark 4.1. In fact, we showed, in deriving relationships (38)–(43), that

$$\begin{aligned}
&\int_Q \int \nabla G_1(u^0)(x, t)(\phi(u)(x, t) - \phi(u^0)(x, t))\, dx\, dt \\
&\quad + \int_\Omega \nabla G_2(u^0)(x)(\phi(u)(x, T) - \phi(u^0)(x, T))\, dx \\
&\quad + \int_{S_T} \int \nabla G_3(u^0)(s, t)(\phi(u)(s, t) - \phi(u^0)(s, t))\, ds\, dt \\
&\qquad = \int_Q \int \left\{ \sum_{i=1}^{n} (F_i(x, t, u_1(x, t)) - F_i(x, t, u_1^0(x, t)))Z(u^0)_{x_i}(x, t) \right. \\
&\qquad\qquad \left. + (f(x, t, u_1(x, t)) - f(x, t, u_1^0(x, t)))Z(u^0)(x, t) \right\} dx\, dt
\end{aligned}$$

$$+\int_{\Omega}(\psi_0(x, u_2(x))-\psi_0(x, u_2^0(x)))Z(u^0)(x, 0)\,dx$$

$$+\int_{S_T}\int(\psi(s, t, u_3(s, t))-\psi(s, t, u_3^0(s, t)))Z(u^0)(s, t)\,ds\,dt. \tag{49}$$

In closing this section, we present a sufficient condition for optimality for the problem (P) in the following theorem.

THEOREM 4.1. *Consider the problem* (P). $u^*\equiv(u_1^*, u_2^*, u_3^*)\in\mathcal{U}$ *is an optimal control if the following conditions are satisfied*:

$$\sum_{i=1}^{n}F_i(x, t, u_1^*(x, t))Z(u^*)_{x_i}(x, t)+f(x, t, u_1^*(x, t))Z(u^*)(x, t)+H_1(x, t, u_1^*(x, t))$$
$$\leqq\sum_{i=1}^{n}F_i(x, t, v_1)Z(u^*)_{x_i}(x, t)+f(x, t, v_1)Z(u^*)(x, t)+H_1(x, t, v_1) \tag{50}$$

for all $v_1\in U_1$ *and almost all* $(x, t)\in Q$;

$$\psi_0(x, u_2^*(x))Z(u^*)(x, 0)+H_2(x, u_2^*(x))\leqq\psi_0(x, v_2)Z(u^*)(x, 0)+H_2(x, v_2) \tag{51}$$

for all $v_2\in U_2$ *and almost all* $x\in\Omega$, *and*

$$\psi(s, t, u_3^*(s, t))Z(u^*)(s, t)+H_3(s, t, u_3^*(s, t))$$
$$\leqq\psi(s, t, v_3)Z(u^*)(s, t)+H_3(s, t, v_3) \tag{52}$$

for all $v_3\in U_3$ *and almost all* $(s, t)\in S_T$.

Proof. For any $u\in\mathcal{U}$, it follows from Lemma 4.1 and the conditions (50) to (52) that

$$J(u)-J(u^*)\geqq 0.$$

Therefore, u^* is an optimal control for the problem (P).

Remark 4.2. In the proof of Theorem 4.1, we note that the convexity property of H_1, H_2 and H_3 in u is not used. This property is required only in proving the necessary condition for optimality and the existence of optimal controls.

5. An optimal control problem with a linear cost functional. In this section, we shall consider the problem (P) under certain linearity assumptions. More precisely, we shall assume that the cost function J takes the following special form:

$$J_1(u)=\int_Q\int\{\bar{G}_1(x, t)\phi(u)(x, t)+H_1(x, t, u_1(x, t))\}\,dx\,dt$$

$$+\int_{\Omega}\{\bar{G}_2(x)\phi(u)(x, t)+H_2(x, u_2(x))\}\,dx \tag{53}$$

$$+\int_{S_T}\int\{\bar{G}_3(s, t)\phi(u)(s, t)+H_3(s, t, u_3(s, t))\}\,ds\,dt$$

The functions $\bar{G}_i$ $(i=1, 2, 3)$ are assumed, throughout this section, to satisfy the following conditions:

(A13) $\bar{G}_1\in L_{q_3, r_3}(Q)$, where q_3 and r_3 are defined in Assumption (A6).

(A14) $\bar{G}_2\in L_2(\Omega)$.

(A15) $\bar{G}_3\in L_{q_4, r_4}(S_T)$, where q_4 and r_4 are defined in Assumption (A8).

Remark 5.1. Using the same arguments as those given to obtain the inequalities (47), (46) and (48), we can show that all the integrals involving the functions $\bar{G}_1$, $\bar{G}_2$

and $\bar{G}_3$ in the definition of the cost functional J_1 are finite under Assumptions (A13), (A14) and (A15). Furthermore, Assumption (A9) is also guaranteed by these three assumptions.

For convenience, the problem (P) with the cost functional J replaced by the cost functional J_1 will be referred as the problem (P1).

The adjoint system for the problem (P1) takes the following form:

$$
\begin{aligned}
& L^*Z(x,t)=\bar{G}_1(x,t), && (x,t)\in Q,\\
& Z(x,t)=\bar{G}_2(x), && x\in\Omega,\\
& \left[\frac{\partial Z}{\partial \nu_{L^*}}+\sigma(s,t)Z\right]\Big|_{S_T}=\bar{G}_3(s,t), && (s,t)\in S_T,
\end{aligned}
\tag{54}
$$

where L^* and $\partial Z/\partial\nu_{L^*}$ are as defined, respectively, in (11) and (12).

Note that the adjoint system (54) is independent of the control $u\in\mathcal{U}$. Hence, the weak solution of the adjoint system, which is guaranteed to exist by Lemma 3.3 and denoted by Z, is also independent of $u\in\mathcal{U}$.

LEMMA 5.1 *Consider the problem* (P1). *Then*

$$
\begin{aligned}
&J_1(u)-J_1(u^0)\\
&\quad=\int_Q\int\Big\{\sum_{i=1}^n\left(F_i(x,t,u_1(x,t))-F_i(x,t,u_1^0(x,t))\right)Z_{x_i}(x,t)\\
&\qquad+\left(f(x,t,u_1(x,t))-f(x,t,u_1^0(x,t))\right)Z(x,t)\\
&\qquad+\left(H_1(x,t,u_1(x,t))-H_1(x,t,u_1^0(x,t))\right)\Big\}\,dx\,dt\\
&\quad+\int_\Omega\left\{\left(\psi_0(x,u_2(x))-\psi_0(x,u_2^0(x))\right)Z(x,0)\right.\\
&\qquad\left.+\left(H_2(x,u_2(x))-H_2(x,u_2^0(x))\right)\right\}dx\\
&\quad+\int_{S_T}\int\left\{\left(\psi(s,t,u_3(s,t))-\psi(s,t,u_3^0(s,t))\right)Z(s,t)\right.\\
&\qquad\left.+\left(H_1(x,t,u_1(x,t))-H_1(x,t,u_1^0(x,t))\right)\right\}dx\,dt
\end{aligned}
\tag{55}
$$

for any $u\equiv(u_1,u_2,u_3)$, $u^0=(u_1^0,u_2^0,u_3^0)\in\mathcal{U}$.

Proof. The proof is similar to that given for Lemma 4.1, except with the cost functional J and the adjoint system (10) being replaced, respectively, by the cost functional J_1 and the adjoint system (54). Furthermore, inequality (38) becomes, in the present case, an equality.

THEOREM 5.1. *A necessary and sufficient condition for an admissible control* $u^*\equiv(u_1^*,u_2^*,u_3^*)$ *to be an optimal control is that*

$$
\begin{aligned}
&\sum_{i=1}^n F_i(x,t,u_1^*(x,t))Z_{x_i}(x,t)+f(x,t,u_1^*(x,t))Z(x,t)+H_1(x,t,u_1^*(x,t))\\
&\qquad\leqq\sum_{i=1}^n F_i(x,t,v_1)Z_{x_i}(x,t)+f(x,t,v_1)Z(x,t)+H_1(x,t,v_1)
\end{aligned}
\tag{56}
$$

for all $v_1\in U_1$ *and almost all* $(x,t)\in Q$;

$$
\psi_0(x,u_2^*(x))Z(x,0)+H_2(x,u_2^*(x))\leqq\psi_0(x,v_2)Z(x,0)+H_2(x,v_2)
\tag{57}
$$

for all $v_2 \in U_2$ *and almost all* $x \in \Omega$, *and*

$$\psi(s, t, u^*(s, t))Z(s, t) + H_3(s, t, u^*(s, t)) \leqq \psi(s, t, v_3)Z(s, t) + H_3(s, t, v_3) \tag{58}$$

for all $v_3 \in U_3$ *and almost all* $(s, t) \in S_T$.

Proof. The sufficient condition follows, as a special case, from Theorem 4.1. It remains to prove the necessary condition. For this, let $u^* \in \mathcal{U}$ be an optimal control. Let $v_1 \in U_1$ and let (x_0, t_0) be an interior point of Q such that it is also a regular point for all those functions appearing on both sides of the inequality (56). Let $\{S_n\} \subset Q$ be a sequence of spheres with (x_0, t_0) as their center such that $|S_n| \to 0$ as $n \to \infty$ (where $|S_n|$ denotes the Lebesgue measure of S_n). Furthermore, let

$$u_1^n(x, t) = \begin{cases} v_1 & \text{if } (x, t) \in S_n, \\ u_1^*(x, t) & \text{if } (x, t) \notin S_n, \end{cases}$$

and define

$$u^n \equiv (u_1^n, u_2^*, u_3^*),$$

Then, it follows from Lemma 5.1 and the optimality of u^* that

$$\begin{aligned} J_1(u^n) - J_1(u^*) = \int_{S_n} \int \Big\{ & \sum_{i=1}^{n} (F_i(x, t, v_1) - F_i(x, t, u^*(x, t)))Z_{x_i}(x, t) \\ & + (f(x, t, v_1) - f(x, t, u_1^*(x, t)))Z(x, t) \\ & + (H_1(x, t, v_1) - H_1(x, t, u_1^*(x, t))) \Big\} \, dx \, dt \geqq 0. \end{aligned}$$

Thus, we have

$$\begin{aligned} \lim_{n \to \infty} \Big[\frac{1}{|S_n|} \int_{S_n} \int \Big\{ & \sum_{i=1}^{n} (F_i(x, t, v) - F_i(x, t, u^*(x, t)))Z_{x_i}(x, t) \\ & + (f(x, t, v_1) - f(x, t, u_1^*(x, t)))Z(x, t) \\ & + H_1(x, t, v_1) - H_1(x, t, u_1^*(x, t)) \Big\} \, dx \, dt \Big] \\ = & \sum_{i=1}^{n} (F_i(x_0, t_0, v_1) - F_i(x_0, t_0, u_1^*(x_0, t_0)))Z_{x_i}(x_0, t_0) \\ & + (f(x_0, t_0, v_1) - f(x_0, t_0, u_1^*(x_0, t_0)))Z(x_0, t_0) \\ & + (H_1(x_0, t_0, v_1) - H_1(x_0, t_0, u_1^*(x_0, t_0))) \geqq 0. \end{aligned} \tag{59}$$

Note that, for each $v_1 \in U_1$, almost all $(x, t) \in Q$ are regular points, that U_1 has countably dense subsets, and that all functions appearing in the inequality (56) are continuous with respect to v_1. Thus, we conclude from (59) that the inequality (56) holds for every $v_1 \in U_1$ and for almost all $(x, t) \in Q$.

The inequalities (57) and (58) can also be derived similarly. This completes the proof.

The next theorem deals with the existence of optimal controls for the problem (P1).

THEOREM 5.2. *Consider the problem* (P1). *Then there exists an admissible control* $u^* \equiv (u_1^*, u_2^*, u_3^*) \in \mathcal{U}$ *such that*

$$\begin{aligned} & \sum_{i=1}^{n} F_i(x, t, u_1^*(x, t))Z_{x_i}(x, t) + f(x, t, u_1^*(x, t))Z(x, t) + H_1(x, t, u_1^*(x, t)) \\ & = \inf_{v_1 \in U_1} \Big\{ \sum_{i=1}^{n} F_i(x, t, v_1)Z_{x_i}(x, t) + f(x, t, v_1)Z(x, t) + H_1(x, t, v_1) \Big\} \end{aligned} \tag{60}$$

for all $(x, t) \in Q$;

$$\psi_0(x, u_2^*(x))Z(x, 0) + H_2(x, u_2^*(x)) = \inf_{v_1 \in U_2} \{\psi_0(x, v_2)Z(x, 0) + H_2(x, v_2)\} \tag{61}$$

for all $x \in \Omega$, *and*

$$\psi(s, t, u_3^*(s, t))Z(s, t) + H_3(s, t, u_3^*(s, t)) = \inf_{v_3 \in U_3} \{\psi(s, t, v_3)Z(s, t) + H_3(s, t, v_3)\} \tag{62}$$

for all $(s, t) \in S_T$.

Furthermore, u^* *is an optimal control.*

Proof. Let

$$I(x, t) \equiv \inf_{v_1 \in U_1} \left\{ \sum_{i=1}^{n} F_i(x, t, v_1)Z_{x_i}(x, t) + f(x, t, v_1)Z(x, t) + H_1(x, t, v_1) \right\}.$$

Since all the functions in the right-hand side of the above expression are continuous in $v_1 \in U_1$ and measurable in $(x, t) \in Q$, I is measurable in Q. Furthermore, since U_1 is compact, we have

$$I(x, t) \in \left\{ \sum_{i=1}^{n} F_i(x, t, U_1)Z_{x_i}(x, t) + f(x, t, U_1)Z(x, t) + H_1(x, t, U_1) \right\}.$$

Thus, it follows readily from [6, Thm. 3'] that there exists a measurable function $u_1^*(x, t) \in U_1$ such that

$$I(x, t) = \sum_{i=1}^{n} F_i(x, t, u^*(x, t))Z_{x_i}(x, t) + f(x, t, u_1^*(x, t))Z(x, t) + H_1(x, t, u_1^*(x, t)).$$

this means that $u_1^* \in \mathcal{U}_1$ and satisfies (60).

Similarly, it can also be shown that there exist measurable functions $u_2^* \in \mathcal{U}_2$ and $u_3^* \in \mathcal{U}_3$ such that they satisfy (61) and (62) respectively.

Finally, since $u^* \equiv (u_1^*, u_2^*, u_3^*)$ belongs to $\mathcal{U}$ and satisfies the sufficient conditions (56) to (58), it is an optimal control. Thus, the proof is complete.

On the basis of the above theorem , a method for constructing an optimal control for the problem (P1) can be described as follows:

1. Solve (54) to find the function $Z(x, t)$ and its derivatives $Z_{x_i}(x, t)$.

2. Find an admissible control $u^* \equiv (u_1^*, u_2^*, u_3^*)$ (whose existence is ensured by the first part of Theorem 5.2) such that (60) to (62) hold. Then, by virtue of the second part of Theorem 5.2, u^* is an optimal control for the problem (P1).

6. An optimal control problem with a linear system. In this section, we shall consider the problem (P) under certain linearity assumptions on the forcing terms, and the initial and boundary data. More precisely, the system (2) shall take the following special form:

$$\begin{aligned} &L\phi(x, t) = \sum_{i=1}^{n} \frac{\partial}{\partial x_i}[\bar{F}_i(x, t) \cdot u_1(x, t)] + \bar{f}(x, t) \cdot u_1(x, t), && (x, t) \in Q, \\ &\phi|_{t=0} = \bar{\psi}_0(x) \cdot u_2(x), && x \in \Omega, \\ &\left[\frac{\partial \phi}{\partial \nu_L} + \sigma(s, t)\phi\right]\Big|_{S_T} = \bar{\psi}(s, t) \cdot u_3(s, t), && (s, t) \in S_T, \end{aligned} \tag{63}$$

where $\cdot$ denotes the usual inner product in any Euclidean space.

The vector valued functions $\bar{F}_i$ $(i=1,\cdots,n)$, $\bar{f}$, $\bar{\psi}_0$ and $\bar{\psi}$ are assumed, throughout this section, to satisfy the following conditions:

(A16) $\bar{F}_i \in L_2(Q, R^{m_1})$ for $i=1,\cdots,n$.

(A17) $\bar{f} \in L_{q_3,r_3}(Q, R^{m_1})$, where q_3 and r_3 are defined in Assumption (A6).

(A18) $\bar{\psi}_0 \in L_2(\Omega, R^{m_2})$.

(A19) $\bar{\psi} \in L_{q_4,r_4}(S_T, R^{m_3})$, where q_4 and r_4 are defined in Assumption (A8).

For convenience, the problem (P) with the system (2) replaced by the system (63) will be referred to as the problem (P2).

Remark 6.1. Consider the problem (P2). The corresponding version of the inequality (36) may be written as:

$$
\begin{aligned}
J(u)-J(u^0) \geqq \int_Q \Big\{ \sum_{i=1}^{n} (\bar{F}_i(x,t)\cdot(u_1(x,t)-u_1^0(x,t)))Z(u^0)_{x_i}(x,t) \\
+(\bar{f}(x,t)\cdot(u_1(x,t)-u_1^0(x,t)))Z(u^0)(x,t) \\
+(H_1(x,t,u_1(x,t))-H_1(x,t,u_1^0(x,t)))\Big\}\,dx\,dt \\
+\int_\Omega \{(\psi_0(x)\cdot(u_2(x)-u_2^0(x)))Z(u^0)(x,0) \\
+(H_2(x,u_2(x))-H_2(x,u_2^0(x)))\}\,dx \\
+\int_{S_T} \{(\bar{\psi}(s,t)\cdot(u_3(s,t)-u_3^0(s,t)))Z(u^0)(s,t) \\
+(H_3(s,t,u_3(s,t))-H_3(s,t,u_3^0(s,t)))\}\,ds\,dt,
\end{aligned} \tag{64}
$$

where $u \equiv (u_1, u_2, u_3)$, $u^0 \equiv (u_1^0, u_2^0, u_3^0) \in \mathcal{U}$, and $Z(u^0)$ is the weak solution of the adjoint system (10) corresponding to the control u^0.

For the rest of this section, let $\nabla H_1(u_1)(x,t)$ denote the gradient of $H_1(x,t,\cdot)$ evaluated at $u_1(x,t)$. Furthermore, $\nabla H_2(u_2)(x)$ and $\nabla H_3(u_3)(s,t)$ are defined similarly.

THEOREM 6.1. *The control problem* (P2) *has a solution.*

Proof. In view of inequality (64) and the assumptions on those functions appearing on the right-hand side of (64), we note that the cost functional J subject to the system (63) is bounded below on $\mathcal{U}$, that is,

$$
\inf_{u\in\mathcal{U}} J(u) = \sigma > -\infty. \tag{65}
$$

Let $\{u^k\} \subset \mathcal{U}$ be a sequence such that

$$
\lim_{k\to\infty} J(u^k) = \sigma. \tag{66}
$$

Such a sequence is referred to as a minimizing sequence. Define

$$
\begin{aligned}
\tilde{\mathcal{U}} \equiv \{u \equiv (u_1,u_2,u_3):\ & u_1(x,t)\in U_1 \quad \text{for almost all } (x,t)\in Q, \\
& u_2(x)\in U_2 \quad \text{for almost all } x\in\Omega \text{ and} \\
& u_3(s,t)\in U_2 \quad \text{for almost all } (s,t)\in S_T\}.
\end{aligned}
$$

Since U_1, U_2 and U_3 are compact and convex subsets of R^{m_1}, R^{m_2} and R^{m_3} respectively, $\tilde{\mathcal{U}}$ is sequentially compact in the weak* topology of $L_\infty(Q, R^{m_1}) \times L_\infty(\Omega, R^{m_2}) \times L_\infty(S_T, R^{m_3})$. Thus, there exists a function $u^* \in \tilde{\mathcal{U}}$ and a subsequence of the sequence $\{u^k\}$, again indexed by k, such that $u^k \to \tilde{u}^*$, as $k\to\infty$, in the weak* topology mentioned above.

Let

$$u^*(x,t)=\begin{cases}\tilde{u}^*(x,t) & \text{if } \tilde{u}^*(x,t)\in U,\\ v^* & \text{if } \tilde{u}^*(x,t)\notin U,\end{cases}$$

where $v^*\equiv(v_1^*, v_2^*, v_3^*)$ is any fixed vector in $U\equiv U_1\times U_2\times U_3$. Then $u^*\in\mathcal{U}$ and, furthermore, the sequence $\{u^k\}$ also converges to u^* in the same topology as $k\to\infty$.

We shall show that u^* is an optimal control.

In view of Remark 6.1 and the convexity of functions H_i $(i=1,2,3)$, we observe that

$$\begin{aligned}J(u^k)-J(u^*)\geqq\int_Q\int\Big\{\sum_{i=1}^n[\bar{F}_i(x,t)\cdot(u_1^k(x,t)-u_1^*(x,t))]Z(u^*)_{x_i}(x,t)\\ +[\bar{f}(x,t)\cdot(u_1^k(x,t)-u_1^*(x,t))]Z(u^*)(x,t)\\ +\nabla H_1(u_1^*)(x,t)\cdot(u_1^k(x,t)-u_1^*(x,t))\Big\}\,dx\,dt\\ +\int_\Omega\{[\bar{\psi}_0(x)\cdot(u_2^k(x)-u_2^*(x))]Z(u^*)(x,0)\\ +\nabla H_2(u_2^*)(x)\cdot(u_2^k(x)-u_2^*(x))\}\,dx\\ +\int_{S_T}\int\{[\bar{\psi}(s,t)\cdot(u_3^k(s,t)-u_3^*(s,t))]Z(u^*)(s,t)\\ +\nabla H_3(u_3^*)(s,t)\cdot(u_3^k(s,t)-u_3^*(s,t))\}\,ds\,dt.\end{aligned}$$

Since $\{u_k\}$ converges to u^* in the weak* topology, the right-hand side of the above inequality converges to zero as $k\to\infty$. Thus,

$$\lim_{k\to\infty}(J(u^k)-J(u^*))\geqq 0,$$

and hence from (65) and (66), we obtain

$$\inf_{u\in\mathcal{U}}J(u)-J(u^*)\geqq 0.$$

This implies that u^* is an optimal control. Thus, the proof is complete.

Remark 6.2. By examining the arguments given for Theorem 6.1, we note that:

(i) $\mathcal{U}$ is sequentially compact in the weak* topology;

(ii) If a minimizing sequence converges to $u^*\in\mathcal{U}$ in the weak* topology, then u^* is an optimal control.

We now present a necessary and sufficient condition for optimality for the problem (P2) in the next theorem.

THEOREM 6.2. *Consider the problem* (P2). *A necessary and sufficient condition for* $u^*\equiv(u_1^*, u_2^*, u_3^*)\in\mathcal{U}$ *to be an optimal control is that*

$$\begin{aligned}&\sum_{i=1}^n(\bar{F}_i(x,t)\cdot u_1^*(x,t))Z(u^*)_{x_i}(x,t)+(\bar{f}(x,t)\cdot u_1^*(x,t))Z(u^*)(x,t)+H_1(x,t,u_1^*(x,t))\\ &=\min_{v_1\in U_1}\Big\{\sum_{i=1}^n(\bar{F}_i(x,t)\cdot v_1)Z(u^*)_{x_i}(x,t)+(\bar{f}(x,t)\cdot v_1)Z(u^*)(x,t)+H_1(x,t,v_1)\Big\}\end{aligned}\tag{67}$$

for almost all $(x,t)\in Q$;

$$\begin{aligned}&(\bar{\psi}_0(x)\cdot u_2^*(x))Z(u^*)(x,0)+H_2(x,u^*(x))\\ &=\min_{v_2\in U_2}\{(\bar{\psi}_0(x)\cdot v_2)Z(u^*)(x,0)+H_2(x,v_2)\}\end{aligned}\tag{68}$$

for almost all $x \in \Omega$, *and*

$$(\bar{\psi}(s,t)\cdot u_3^*(s,t))Z(u^*)(s,t)+H_3(s,t,u_3^*(s,t)) = \min_{v_3\in U_3}\{(\bar{\psi}(s,t)\cdot v_3)Z(u^*)(s,t)+H_3(s,t,v_3)\} \tag{69}$$

for almost all $(s,t)\in S_T$.

Proof. The sufficient condition follows readily from the inequality (64).

To prove the necessary condition, let $u^*\in\mathcal{U}$ be an optimal control. Then, for any $u\in\mathcal{U}$ and $0<\varepsilon\leqq 1$, we have

$$(J(u^*+\varepsilon(u-u^*))-J(u^*))/\varepsilon \geqq 0,$$

and hence

$$\lim_{\varepsilon\to 0}(J(u^*+\varepsilon(u-u^*))-J(u^*))/\varepsilon = J_{u^*}(u)-J_{u^*}(u^*)\geqq 0 \tag{70}$$

where

$$\begin{aligned} J_{u^*}(u) \equiv & \int_Q\int \{\nabla G_1(u^*)(x,t)\phi(u)(x,t)+\nabla H_1(u_1^*)(x,t)\cdot u_1(x,t)\}\,dx\,dt \\ & +\int_\Omega \{\nabla G_2(u^*)(x)\phi(u)(x,T)+\nabla H_2(u_2^*)(x)\cdot u_2(x)\}\,dx \\ & +\int_{S_T}\int \{\nabla G_3(u^*)(s,t)\phi(u)(s,t)+\nabla H_3(u_3^*)(s,t)\cdot u_3(s,t)\}\,ds\,dt. \end{aligned} \tag{71}$$

By virtue of (49) and the convexity of the functions H_i $(i=1,2,3)$, we have

$$\begin{aligned} & \int_Q\int \Big\{\sum_{i=1}^m (\bar{F}_i(x,t)\cdot(u_1(x,t)-u_1^*(x,t)))Z(u^*)_{x_i}(x,t) \\ & \quad +(\bar{f}(x,t)\cdot(u_1(x,t)-u_1^*(x,t)))Z(u^*)(x,t) \\ & \quad +(H_1(x,t,u_1(x,t))-H_1(x,t,u_1^*(x,t)))\Big\}\,dx\,dt \\ & +\int_\Omega \{(\bar{\psi}_0(x)\cdot(u_2(x)-u_2^*(x)))Z(u^*)(x,0) \\ & \quad +(H_2(x,u_2(x))-H_2(x,u_2^*(x)))\}\,dx \\ & +\int_{S_T}\int \{(\bar{\psi}(s,t)\cdot(u_3(s,t))-u_3^*(s,t)))Z(u^*)(s,t) \\ & \quad +(H_3(s,t,u_3(s,t))-H_3(s,t,u_3^*(s,t)))\}\,ds\,dt \\ & \geqq J_{u^*}(u)-J_{u^*}(u^*)\geqq 0. \end{aligned} \tag{72}$$

Using the same approach as that used in the proof of the necessary condition in Theorem 5.1, the conditions (67) to (69) follow easily from the inequality (72). Thus, the proof is complete.

In what follows, we shall devise an algorithm for solving the problem (P2). This algorithm is an iterative method and can be used to construct a minimizing sequence of controls $\{u^k\}\subset\mathcal{U}$ corresponding to any given initial control $u^1\in\mathcal{U}$. For convenience, this algorithm will be referred to as Algorithm (M). Its detailed statement is now given as follows:

ALGORITHM (M)

Step 1. Choose an initial control $u^1\in\mathcal{U}$ and set $k=1$.

Step 2. Choose $\tilde{u}^k \in \mathcal{U}$ by the requirement that

$$\begin{aligned}\sum_{i=1}^{n} (\bar{F}_i(x,t)\cdot \tilde{u}_1^k(x,t))Z(u^k)_{x_i}(x,t)+(\bar{f}(x,t)\cdot \tilde{u}_1^k(x,t))Z(\tilde{u}^k)(x,t)\\ +\nabla H_1(u_1^k)(x,t)\cdot \tilde{u}_1^k(x,t)\\ =\inf_{v_1\in U_1}\Big\{\sum_{i=1}^{n}(\bar{F}_i(x,t)\cdot v_1)Z(u^k)_{x_i}(x,t)\\ +(\bar{f}(x,t)\cdot v_1)Z(u_1^k)(x,t)+\nabla H_1(u_1^k)(x,t)\cdot v_1\Big\}\end{aligned} \tag{73}$$

for almost all $(x,t)\in Q$;

$$\begin{aligned}(\bar{\psi}_0(x)\cdot \tilde{u}_2^k(x))Z(u^k)(x,0)+\nabla H_2(u_2^k)(x)\cdot \tilde{u}_2^k(x)\\ =\inf_{v_2\in U_2}\{(\bar{\psi}_0(x)\cdot v_2)Z(u^k)(x,0)+\nabla H_2(u_2^k)(x)\cdot v_2\}\end{aligned} \tag{74}$$

for almost all $x\in\Omega$, and

$$\begin{aligned}(\bar{\psi}(s,t)\cdot \tilde{u}_3^k(s,t))Z(u^k)(s,t)+\nabla H_3(u_3^k)(s,t)\cdot \tilde{u}_3^k(s,t)\\ =\inf_{v_3\in U_3}\{(\bar{\psi}(s,t)\cdot v_3)Z(u^k)(s,t)+\nabla H_3(u_3^k)(s,t)\cdot v_3\}\end{aligned} \tag{75}$$

for almost all $(s,t)\in S_T$, where $Z(u^k)$ is the solution of the system (10) corresponding to the control u^k.

Step 3. Let $u^{k+1}=u^k+\alpha^k(\tilde{u}^k-u^k)$, where α_k is such that

$$J(u^k+\alpha^k(\tilde{u}^k-u^k))=\inf_{0\leqq\alpha\leqq 1} J(u^k+\alpha(\tilde{u}^k-u^k)).$$

Step 4. Go to Step 2 with k replaced by $k+1$.

Remark 6.3. Consider an optimal control problem which consists of the system (63) and the following cost functional:

$$\begin{aligned}J_{uk}(u)\equiv\int_Q\int\{\nabla G_1(u^k)(x,t)\phi(u)(x,t)+\nabla H_1(u_1^k)(x,t)\cdot u_1(x,t)\}\,dx\,dt\\ +\int_\Omega\{\nabla G_2(u^k)(x)\phi(u)(x,T)+\nabla H_2(u_2^k)(x)\cdot u_2(x)\}\,dx\\ +\int_{S_T}\int\{\nabla G_3(u^k)(s,t)\phi(u)(s,t)+\nabla H_3(u_3^k)(s,t)\cdot u_3(s,t)\}\,ds\,dt.\end{aligned} \tag{76}$$

By applying Theorem 5.1 to the present case, it is easy to verify that $\tilde{u}^k$ (determined by Step 2 of Algorithm (M)) minimizes the cost functional $J_{uk}(u)$.

For the convergence of the algorithm, we need the following additional assumptions:

(A20) $G_1(x,t,y)$, $G_2(x,y)$, $G_3(s,t,y)$ are twice differentiable with respect to y such that

$$\begin{aligned}\|\nabla^2 G_1(\cdot,\cdot,\phi(u)(\cdot,\cdot))\|_{2,Q}\leqq M_3,\\ \|\nabla^2 G_2(\cdot,\phi(u)(\cdot,0))\|_{2,\Omega}\leqq M_3,\\ \|\nabla^2 G_3(\cdot,\cdot,\phi(u)(\cdot,\cdot))\|_{2,S_T}\leqq M_3\end{aligned}$$

for all $u\equiv(u_1,u_2,u_3)\in\mathcal{U}$, where M_4 is a constant independent of $u\in\mathcal{U}$ and $\nabla^2 H_i$ the Hessians of G_i $(i=1,2,3)$ with respect to y.

(A21) $H_1(x, t, u_1)$, $H_2(x, u_2)$ and $H_3(s, t, u_3)$ are twice differentiable with respect to u_1, u_2 and u_3 respectively such that

$$\|\nabla^2 H_1(\cdot, \cdot, u_1(\cdot, \cdot))\|_{1,Q} \leqq M_4,$$
$$\|\nabla^2 H_2(\cdot, u_2(\cdot))\|_{1,\Omega} \leqq M_4,$$
$$\|\nabla^2 H_3(\cdot, \cdot, u_3(\cdot, \cdot))\|_{1,S_T} \leqq M_4$$

for all $u \equiv (u_1, u_2, u_3) \in \mathcal{U}$, where M_4 is a constant independent of $u \in \mathcal{U}$ and $\nabla^2 H_i$ $(i = 1, 2, 3)$ denote the Hessians of H_i $(i = 1, 2, 3)$ with respect to u_i.

LEMMA 6.1. *The sequence* $\{u^k\}$ *generated by Algorithm* (M) *is a minimizing sequence. That is to say,*

$$\lim_{k\to\infty} J(u^k) = J(u^*).$$

where u^* *is an optimal control.*

Proof. Let

$$C_k(\varepsilon) \equiv J(u^k + \varepsilon(\tilde{u}^k - u^k))$$

where $\varepsilon \in [0, 1]$ and $\tilde{u}^k$ is obtained from u^k as stated in Step 2 of Algorithm (M).

Then, by Taylor's theorem in remainder form, it follows that

$$C_k(s) = C_k(0) + \varepsilon C_k'(0) + \tfrac{1}{2}(\varepsilon)^2 C_k''(\theta\varepsilon) \tag{77}$$

for some $\theta \in (0, 1)$.

Clearly,

$$C_k(0) = J(u^k), \tag{78}$$

$$C_k'(0) = J_{uk}(\tilde{u}^k) - J_{uk}(u^k), \tag{79}$$

where $J_{uk}(u)$ is as defined by (76), and

$$\begin{aligned} C_k''(\theta\varepsilon) = &\int_Q \int \{\nabla^2 G_1(x, t, \phi(\xi)(x, t))(\phi(\tilde{u}^k) - \phi(u^k))^2 \\ &+ (\nabla^2 H_1(x, t, \xi_1(x, t))(\tilde{u}_1^k(x, t) - u_1^k(x, t))) \cdot (\tilde{u}_1^k(x, t) - u_1^k(x, t))\} \, dx \, dt \\ &+ \int_\Omega \{\nabla^2 G_2(x, \phi(\xi)(x, T))(\phi(\tilde{u}^k)(x, T) - \phi(u^k)(x, T))^2 \\ &+ (\nabla^2 H_2(x, \xi_2(x))(\tilde{u}_2^k(x) - \tilde{u}_2^k(x))) \cdot (\tilde{u}_2^k(x) - u_2^k(x))\} \, dx \\ &+ \int_{S_T} \int \{\nabla^2 G_3(s, t, \phi(\xi)(s, t))(\phi(\tilde{u}^k)(s, t) - \phi(u^k)(s, t))^2 \\ &+ (\nabla^2 H_3(s, t, \xi_3(s, t))(\tilde{u}_3^k(s, t) - u_3^k(s, t))) \cdot (\tilde{u}_3^k(s, t) - u_3^k(s, t))\} \, ds \, dt \end{aligned} \tag{80}$$

where $\xi \equiv (\xi_1, \xi_2, \xi_2) = u^k + \theta\varepsilon(\tilde{u}^k - u^k)$.

By virtue of Assumptions (A20), (A21) and Lemma 3.1, it follows from (80) that there exists a positive constant M_5 such that

$$|C_k''(\theta\varepsilon)| \leqq M_5 \tag{81}$$

where M_5 is independent of k and ε.

Then, it follows from (77) to (81) that

$$J(u^k \varepsilon(\tilde{u}^k - u^k)) \leqq J(u^k) + \varepsilon(J_{uk}(\tilde{u}^k) - J_{uk}(u^k)) + \tfrac{1}{2}(\varepsilon)^2 M_5. \tag{82}$$

Let u^* be an optimal control. Then, by Remark 6.3, we have

$$J_{uk}(\tilde{u}^k) \leqq J_{uk}(u^*). \tag{83}$$

Thus, it folows from (82) and (83) that

$$\begin{aligned} &J(\tilde{u}^k + \varepsilon(\tilde{u}^k - \tilde{u}^k)) - J(u^*) \\ &\quad \leqq J(u^k) - J(u^*) + \tfrac{1}{2}\varepsilon(J_{uk}(\tilde{u}^k) - J_{uk}(u_k)) + \tfrac{1}{2}\varepsilon(J_{uk}(u^*) - J_{uk}(u^k)) + \tfrac{1}{2}(\varepsilon)^2 M_5. \end{aligned} \tag{84}$$

Let

$$\varepsilon = \varepsilon^* \equiv \min\left\{\frac{1}{M_5}(J_{uk}(u^k) - J_{uk}(\tilde{u}^k)), 1\right\}. \tag{85}$$

Then, it is clear that $\varepsilon^* \in [0, 1]$ and that (84) reduces to

$$J(v^*) - J(u^*) \leqq J(u^k) - J(u^*) + \tfrac{1}{2}\varepsilon^*(J_{uk}(u^*) - J_{uk}(u^k)), \tag{86}$$

where $v^* = u^k + \varepsilon^*(\tilde{u}^k - u^k)$.

By virtue of the convexity of the functions G_i and H_i $(i = 1, 2, 3)$, we have

$$J(u^*) - J(u^k) \geqq J_{uk}(u^*) - J_{uk}(u^k). \tag{87}$$

In view of Step 3 of Algorithm (M), we note that

$$J(u^{k+1}) \leqq J(v^*). \tag{88}$$

Thus, from (86), (87) and (88), it follows that

$$J(u^{k+1}) - J(u^*) \leqq (1 - \tfrac{1}{2}\varepsilon^*)(J(u^k) - J(u^*)). \tag{89}$$

From (85), (83) and (87), we have

$$\begin{aligned} 1 - \tfrac{1}{2}\varepsilon^* &= \max\left\{1 - \frac{1}{2M_5}(J_{uk}(u^k) - J_{uk}(\tilde{u}^k)), \tfrac{1}{2}\right\} \\ &\leqq \max\left\{1 - \frac{1}{2M_5}(J_{uk}(u^k) - J_{uk}(u^*)), \tfrac{1}{2}\right\} \\ &\leqq \max\left\{1 - \frac{1}{2M_5}(J(u^k) - J(u^*)), \tfrac{1}{2}\right\}. \end{aligned} \tag{90}$$

Let $a^k \equiv (1/2M_5)(J(u^k) - J(u^*))$. Then, it follows from (90) that

$$1 - \tfrac{1}{2}\varepsilon^* \leqq \max\{1 - a^k, \tfrac{1}{2}\}. \tag{91}$$

Combining (89) and (91), we obtain

$$a^{k+1} \leqq a^k \max\{1 - a^k, \tfrac{1}{2}\}. \tag{92}$$

This inequality shows that the nonnegative sequence $\{a^k\}$ is monotonically decreasing. Thus, it follows that $a^* = \lim_{k\to\infty} a^k$ exists and, by (92),

$$a^* \leqq a^* \max\{1 - a^*, \tfrac{1}{2}\}.$$

This, in turn, implies that $a^* = 0$, and hence we have

$$\lim_{k\to\infty} J(u^k) = J(u^*).$$

Thus the proof is complete.

Next, we shall present some results concerning the convergence of the sequence of controls $\{u^k\}$ generated by Algorithm (M).

THEOREM 6.3. *The sequence* $\{u^k\}$, *generated by Algorithm* (M), *has a subsequence which converges to an optimal control in the weak* topology of* $L_\infty(Q, R^{m_1}) \times L_\infty(\Omega, R^{m_2}) \times L_\infty(S_T, R^{m_3})$. *Furthermore, if* $u^* \in \mathcal{U}$ *is an accumulation point of the sequence* $\{u^k\}$ (*with respect to the weak* topology*), *then it is an optimal control.*

Proof. The proof follows readily from Lemma 6.3 and Remark 6.2.

THEOREM 6.4. *Suppose that* $H_1(x, t, \cdot)$, $H_2(x, \cdot)$ *and* $H_3(s, t, \cdot)$ *are, respectively, strictly convex in* u_1, u_2 *and* u_3 *for almost all* $(x, t) \in Q$, $x \in \Omega$ *and* $(s, t) \in S_T$. *Then the sequence* $\{u^k\}$, *generated by Algorithm* (M), *converges to the optimal control in the weak* topology.*

Proof. Since H_i $(i = 1, 2, 3)$ are strictly convex and the system (63) is linear, it is easy to verify that the functional $J(u)$ is also strictly convex on $\mathcal{U}$. (Here, we identify all the elements of $\mathcal{U}$ which are equal almost everywhere.)

Thus, $J(u)$ has a unique minimum on $\mathcal{U}$, and hence the problem (P2) has a unique optimal control. Let the optimal control be denoted by u^*. Then, by virtue of Remark 6.2(i), it follows that every subsequence of the sequence $\{u^k\}$ has a further subsequence which converges to the unique optimal control u^* in the weak* topology. Thus, the whole sequence $\{u^k\}$ converges to u^* in the same topology. The proof is complete.

In fact, under the assumptions of Theorem 6.4 we can also show that the sequence $\{u^k\}$ converges to the optimal control in the almost-everywhere topology. This is a stronger result than that obtained in Theorem 6.4, because, for the class of admissible controls considered in this paper, almost-everywhere convergence implies weak* convergence.

THEOREM 6.5. *Under the assumptions of Theorem* 6.4, *the sequence* $\{u^k\}$, *generated by Algorithm* (M), *converges almost everywhere to the optimal control* u^*.

Proof. By remark 6.1, we have

$$
\begin{aligned}
J(u^k) - J(u^*) \geqq \int_Q \int \Big\{ \sum_{i=1}^{n} & (\bar{F}_i(x,t) \cdot (u_1^k(x,t) - u_1^*(x,t)))Z(u^*)_{x_i}(x,t) \\
& + (\bar{f}(x,t) \cdot (u_1^k(x,t) - u_1^*(x,t)))Z(u^*) \\
& + (H_1(x,t,u_1^k(x)) - H_1(x,t,u_1^*(x,t))) \Big\} \, dx \, dt \\
+ \int_\Omega & \{(\bar{\psi}_0(x) \cdot (u_2^k(x) - u_2^*(x)))Z(u^*)(x,0) \\
& + (H_2(x, u_3^k(x)) - H_2(x, u_3^*(x)))\} \, dx \\
+ \int_{S_T} \int & \{(\bar{\psi}(s,t) \cdot (u_3^k(s,t) - u_3^*(s,t)))Z(u^*)(s,t) \\
& + (H_3(s,t,u_3^k(s,t)) - H_3(s,t,u_3^*(s,t)))\} \, ds \, dt.
\end{aligned} \tag{93}
$$

According to Theorem 6.2, those integrands on the right-hand side of (93) are nonnegative almost everywhere in their domains of definition. Since the left-hand side of (93) tends to zero as $k \to \infty$, it follows that

$$
\begin{aligned}
& \sum_{i=1}^{n} (\bar{F}_i(x,t) \cdot u_1^k(x,t))Z(u^*)_{x_i}(x,t) \\
& \quad + (\bar{f}(x,t) \cdot u_1^k(x,t))Z(u^*)(x,t) + H_1(x,t,u_1^k(x,t)) \\
& \to \sum_{i=1}^{n} (\bar{F}_i(x,t) \cdot u_1^*(x,t))Z(u^*)_{x_i}(x,t) \\
& \quad + (\bar{f}(x,t) \cdot u_1^*(x,t))Z(u^*)(x,t) + H_1(x,t,u_1^*(x,t))
\end{aligned} \tag{94}
$$

for almost all $(x, t) \in Q$;

$$(\bar{\psi}_0(x) \cdot u_2^k(x))Z(u^*)(x, 0) + H_2(x, u_2^k(x)) \to (\bar{\psi}_0(x) \cdot u_2^*(x))Z(u^*)(x, 0) + H_2(x, u_2^*(x)) \tag{95}$$

for almost all $x \in \Omega$, and

$$(\bar{\psi}(s, t) \cdot u_3^k(s, t))Z(u^*)(s, t) + H_3(s, t, u_3^k(s, t)) \to (\bar{\psi}(s, t) \cdot u_3^k(s, t))Z(u^*)(s, t) + H_3(s, t, u_3^*(s, t)) \tag{96}$$

for almost all $(s, t) \in S_T$.

We are now going to prove that $u_1^k(x, t) \to u_1^*(x, t)$ as $k \to \infty$ for almost all $(x, t) \in Q$.

Note that the strict convexity assumption on $H_1(x, t, \cdot)$ ensures that the minimum in (67) is attained at a unique point $u_1^*(x, t)$ for almost all $(x, t) \in Q$. Let $(\hat{x}, \hat{t})$ be a point in Q at which (94) and the unique minimum condition (67) hold. Then, we can show that $u_1^k(x, t) \to u_1^*(\hat{x}, \hat{t})$. If this were false, we could choose a subsequence $\{u_1^{k_i}(\hat{x}, \hat{t})\}$ of the sequence $\{u_1^k(\hat{x}, \hat{t})\}$ and a point $\hat{u}_1 \neq u_1^*(\hat{x}, \hat{t})$ such that

$$\lim_{i \to \infty} u_1^{k_i}(\hat{x}, \hat{t}) = \hat{u}_1. \tag{97}$$

Thus, it follows from (94) and the continuity of $H_1(\hat{x}, \hat{t}, \cdot)$ that

$$\sum_{i=1}^{n} (\bar{F}_i(\hat{x}, \hat{t}) \cdot \hat{u}_1)Z(u^*)_{x_i}(x, t) + (\bar{f}(\hat{x}, \hat{t}) \cdot \hat{u}_1)Z(u^*)(x, t) + H_1(\hat{x}, \hat{t}, \hat{u}_1)$$

$$= \sum_{i=1}^{n} (\bar{F}_i(\hat{x}, \hat{t}) \cdot u_1^*(\hat{x}, \hat{t}))Z(u^*)_{x_i}(\hat{x}, \hat{t}) + (\bar{f}(\hat{x}, \hat{t}) \cdot u_1^*(\hat{x}, \hat{t}))Z(u^*)(\hat{x}, \hat{t}) + H_1(\hat{x}, \hat{t}, u_1^*(\hat{x}, \hat{t})).$$

Clearly, this contradicts the unique minimum condition at $(\hat{x}, \hat{t})$. Thus, $u_1^k(\hat{x}, \hat{t}) \to u_1^*(\hat{x}, \hat{t})$. Since almost all points in Q can be chosen as $(\hat{x}, \hat{t})$, $u_1^k \to u_i^*$ almost everywhere on Q.

Similarly, we can prove that $u_2^k \to u_2^*$ and $u_3^k \to u_3^*$ almost everywhere in their respective domains of definition. This completes the proof.

Remark 6.4. For some readers, it may be helpful to consider the functional $J(u)$ given in (3) as the sum of various terms, each of which involves a composition of several more basic functionals or operators. For example, the term containing G_1 can be thought of as a composition $J^1 = G_1 \circ \mathcal{S} \circ \mathcal{F}e$, where

$$\mathcal{F}e(u_1, u_2, u_3) = (F_1, f, \psi_0, \psi)$$

(the data for the initial bounday value problem), $\mathcal{S}$ is a linear operator which maps that data to the solution $\phi(u)$, and G_1 is the convex functional defined in terms of $\mathcal{S}$. In this situation, G_1 is linear for problem (P1), while $\mathcal{F}e$ is linear for problem (P2). The (necessary and) sufficient conditions for optimality then involve convexity arguments, the chain rule for derivatives and computational of the adjoint of $\mathcal{S}$.

Acknowledgments. The authors wish to express their most sincere appreciation to Dr. K. G. Choo for reading the entire manuscript and giving them many valuable comments and suggestions. Furthermore, they also wish to thank the two referees of this paper for their most valuable comments and suggestions. References [4] and [7] and Remarks 4.1 and 6.4 were added in response to some of their comments and suggestions.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New york, 1975.
[2] N. U. AHMED AND K. L. TEO, *Optimal Control of Distributed Parameter Systems*, North-Holland, New York, 1981.
[3] EARL R. BARNES, *An extension of Gilbert's algorithm for computing optimal controls*, J. Optim. Theory Appl. 7 (1971), pp. 402–443.
[4] V. BARBU, *Boundary control problems with convex criterion*, this Journal, 18 (1980), pp. 227–243.
[5] A. G. BUTKOVSKIY, *Distributed Control Systems*, American Elsevier, New York, 1969.
[6] C. J. HIMMELBERG, M. Q. JACOBS AND F. S. VEN VLECK, *Measurable multiplications, selectors, and Filippov's implicit functions lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.
[7] K. MALANOWSKI, *Convergence of approximations vs. regularity of solutions for convex, control-constrained optimal-control problems*, Appl. Math. Optim., 8 (1981), pp. 69–95.
[8] O. A. LADYZHENSKAYA, V. A. SOLONNIKOV AND N. N. URALČEVA, *Linear and Quasilinear Equations of Parabolic Type*, American Mathematical Society, Providence, RI, 1968.
[9] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, II, Springer-Verlag, Berlin, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0007 $01.25/0

SOME PROPERTIES OF A CLASS OF CONTINUOUS LINEAR PROGRAMS*

E. J. ANDERSON†, P. NASH† AND A. F. PEROLD‡

Abstract. This paper discusses a class of linear programs posed in a function space; a member of this class is called a separated continuous linear program (SCLP). Such problems occur, for example, in the planning of production and inventory. We characterize the L_∞ extreme point solutions of SCLP in a manner analogous to the basic solutions of finite dimensional linear programming and give a sufficient condition for there to exist optimal extreme point solutions with finitely many constant-basis intervals. SCLP is to date the most general continuous linear program for which such strong characterizations have been found.

Introduction. A number of authors have considered the problem

$$\text{maximize} \quad \int_0^T c(t)^T x(t)\,dt$$

$$\text{subject to} \quad Bx(t) + \int_0^t Kx(s)\,ds = b(t),$$

$$x(t) \geqq 0, \qquad t \in [0, T].$$

This problem, called a continuous linear program (CLP), was studied by Bellman (1957) who introduced it in an economic context. Since then Levinson (1966), Tyndall (1967) and Grinold (1970), amongst others, have investigated this problem. These authors have been primarily concerned with establishing strong duality theorems when CLP and an appropriate dual are both posed in $L_\infty[0, T]$.

Lehmann (1954), Drews (1974), Hartberger (1974) and Segers (1974) have investigated the possibility of a solution algorithm for CLP. This would amount to a generalization of the simplex method to a function space setting. Perold (1978) has dealt in detail with the problems involved in the construction of such an algorithm.

In this paper, we consider a subclass of continuous linear programming problems, a member of which is called a *separated continuous linear program* (SCLP). The form of SCLP is

$$\text{maximize} \quad \int_0^T c(t)^T x(t)\,dt$$

$$\text{subject to} \quad \int_0^t Gx(s)\,ds + y(t) = a(t), \tag{1}$$

$$Hx(t) + z(t) = b(t), \tag{2}$$

$$x(t), y(t), z(t) \geqq 0, \qquad t \in [0, T]. \tag{3}$$

Here x, z, b and c are bounded, measurable functions; y and a are absolutely continuous functions. The dimensions of x, y and z are n_1, n_2 and n_3 respectively. The description "separated" refers to the fact that the constraints are in two sets, the integral constraints (1) and the instantaneous constraints (2), (3).

* Received by the editors June 23, 1980, and in revised form September 7, 1982.
† University Engineering Department, Cambridge, England CB2 1RX.
‡ Graduate School of Business Administration, Harvard University, Boston, Massachusetts 02163.

SCLP has an alternative formulation as a linear optimal control problem, with state positivity constraints;

$$\begin{aligned} \text{maximize} \quad & \int_0^T \{c_1(t)^T y(t) + c_2(t)^T u(t)\}\, dt \\ \text{subject to} \quad & \dot{y}(t) = h(t) - Gu(t), \\ & y(0) = a(0), \\ & u(t) \in U(t) = \{u : u \geqq 0,\ Hu \leqq b(t)\}, \\ & y(t) \geqq 0, \qquad t \in [0, T]. \end{aligned}$$

The form of SCLP can be recovered from this by formally integrating the dynamics and removing the state dependence from the objective by integrating by parts.

Problems of SCLP type arise in considering continuous-time multi-commodity network flow problems of various kinds, including production and inventory problems, deterministic reservoir control problems and transportation and storage problems. The integral constraints (1) represent constraints on storage at nodes of the network, for example stockpiles or reservoirs, while the instantaneous constraints (2) typically represent constraints on processing or transportation capacity in the arcs.

The intent of this work is to investigate the nature of the optimal solutions to SCLP with the hope of furthering the development of algorithms for this and more general continuous linear programs. To this end we begin, in § 1, by defining basic solutions, and establish, under fairly weak conditions, that the extreme points of the set of feasible solutions for SCLP can be characterized as basic. Thus optimal basic solutions exist for SCLP whenever an optimal solution can be found at an extreme point. This is a stronger property than that obtained by Perold (1981) for the general CLP. In § 2, we make use of this property to show that the optimal solutions for a subclass of SCLP's can always be chosen to be piecewise linear.

The results of this paper mirror closely the well-known bang-bang theorems of optimal control theory. However, the proofs are very different: on one hand SCLP is harder to solve than the usual linear control problem because of the presence of inequality constraints on the state variables; on the other SCLP is easier to solve because there is no feedback allowed in (1), i.e., y does not appear under the integral sign. CLP is the hardest of all, allowing both state variable inequality constraints and feedback. Whether as strong results can be obtained for CLP remains an open question. Perold (1981) has given counter examples for CLP with time varying coefficient B and K.

1. Basic feasible solutions. Let $n = (n_1 + n_2 + n_3)$ and $m = (n_2 + n_3)$. For any x in $L_\infty^n[0, T]$ define the *support* of x, denoted by S_x, to be the set-valued function on $[0, T]$ such that $S_x(t)$ is the set of indices of nonzero components of $x(t)$; that is

$$S_x(t) = \{k : |x_k(t)| > 0\}, \qquad t \in [0, T].$$

Strictly, of course, this defines an equivalence class of set-valued functions, differing from each other on sets of measure zero. This has no essential bearing on what follows, and we shall ignore it for the sake of brevity in the exposition.

Define

$$\hat{x}(t) = \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix}.$$

Note that if $\hat{x}$ is feasible (i.e., if its components x, y and z satisfy (1), (2) and (3)) then it is determined from x alone.

Define the $n \times m$ matrix K by

$$K = \begin{bmatrix} G & I & 0 \\ H & 0 & I \end{bmatrix}.$$

Note that rank $(K) = m$. Let

$$\bar{x}(t) = \begin{bmatrix} x(t) \\ \dot{y}(t) \\ z(t) \end{bmatrix}.$$

Again, if $\hat{x}$ is feasible then $\bar{x}$ is determined from x alone. Clearly $\bar{x}$ will also determine $\hat{x}$.

Equation (1) can be differentiated, at least almost everywhere, to give

$$Gx(t) + \dot{y}(t) = \dot{a}(t). \tag{4}$$

Thus (1) and (2) are equivalent to

$$K\bar{x}(t) = \begin{bmatrix} \dot{a}(t) \\ b(t) \end{bmatrix}, \qquad y(0) = a(0). \tag{5}$$

If $\hat{x}$ is feasible and the columns of K indexed by $S_{\hat{x}}(t)$ are linearly independent for almost all t in $[0, T]$, then $\hat{x}$ is called a *basic feasible solution* (b.f.s.). Thus if $\hat{x}$ is a b.f.s., $S_{\hat{x}}(t)$ contains no more than m elements for almost all t in $[0, T]$.

If $\hat{x}$ is feasible and there is no feasible $\hat{x}'$ such that $S_{\hat{x}'}(t) \subset S_{\hat{x}}(t)$ almost everywhere on $[0, T]$, with strict inclusion on a set of nonzero measure, then $\hat{x}$ is said to have *minimal support.*

Let F be the set of feasible solutions. We shall require the following assumptions:

A: F is bounded;

B: $a(t)$ is differentiable everywhere and $\dot{a}(t)$ is bounded.

The following theorem is straightforward to establish.

THEOREM 1. *If* A *holds, then F is weak* compact and there is an optimal solution for* SCLP *at an extreme point of F.*

This theorem shows that in looking for an optimal solution for SCLP we need only consider extreme points of F. The next step is to characterize these extreme points, which is done in the theorem below.

THEOREM 2. *Suppose that $\hat{x}$ is feasible for* SCLP *and B holds. Then the following statements are equivalent*:

(i) *$\hat{x}$ is basic.*

(ii) *$\hat{x}$ has minimal support.*

(iii) *$\hat{x}$ is an extreme point of F.*

Proof. This theorem is proved by demonstrating the equivalence of (i) and (ii) and of (i) and (iii).

(a) Suppose that $\hat{x}$ is feasible with minimal support but is not basic. Then there is some set P of nonzero measure on which the columns of K indexed by $S_{\hat{x}}(t)$ are linearly dependent. Choose some subset of P, say P', also of nonzero measure, on which $S_{\hat{x}}(t)$ is constant, say equal to S. If S does not contain the index of any y variable, we can use a standard argument from finite dimensional linear programming at each $t \in P'$ to produce a new feasible solution with strictly smaller support than $\hat{x}$ on P'. The difficulty is that when S does contain the index of some y variables, we have to show that this can be done without any of those variables becoming negative.

This is done as follows.

By definition, the y variables indexed by S are strictly positive at each t in P'. Since the y's are continuous functions, we can choose an $\varepsilon(t)>0$ and an open interval I_t for each t in P', such that

$$\hat{x}_i(t')>\varepsilon(t) \quad \text{for all } i\in S, \qquad n_1<i\leqq n_1+n_2 \quad \text{for all } t'\in I_t.$$

Let

$$J=\bigcup_{t\in P'} I_t.$$

Then $P'\subset J$, and so for at least one of the pairs $\{(I_t, \varepsilon(t))\}$, say (I, ε), $P''=P'\cap I$ has measure greater than zero and

$$\hat{x}_i(t')>\varepsilon \quad \text{for all } i\in S, \qquad n_1<i\leqq n_1+n_2 \quad \text{for all } t'\in I.$$

From the definition of P there is some $d\in R^n$, $d\neq 0$ with $Kd=0$ and $d_i=0$ for all $i\notin S$. If $\bar{x}'$ is defined from $\bar{x}$ by

$$\bar{x}'(t)=\bar{x}(t)+h(t)d, \qquad t\in[0, T], \tag{6}$$

where $h(t)$ is any bounded measurable scalar function, then $\bar{x}'$ satisfies (5). Hence $\hat{x}'$ (defined in the obvious way from $\bar{x}'$) is feasible if it is nonnegative. We construct a function h which is zero outside P'' and such that $\hat{x}'$ is nonnegative and has strictly smaller support than $\hat{x}$.

First define functions f_1 and f_2 by

$$f_1(t)=\begin{cases}\min\limits_{k\in I_1}\{\bar{x}_k(t)/d_k\}, & I_1\neq\varnothing,\\ 1, & I_1=\varnothing,\end{cases}$$

$$f_2(t)=\begin{cases}\min\limits_{k\in I_2}\{-\bar{x}_k(t)/d_k\}, & I_2\neq\varnothing,\\ 1, & I_2=\varnothing\end{cases}$$

for each $t\in P''$, where

$$I_1=\{k: d_k>0,\ k\leqq n_1 \text{ or } k>n_1+n_2\},$$

$$I_2=\{k: d_k<0,\ k\leqq n_1 \text{ or } k>n_1+n_2\}.$$

Then f_1 and f_2 are bounded, measurable functions, nonzero on at least a subset of P'' of nonzero measure; $f_1(f_2)$ gives the instantaneous largest multiple of d which can be subtracted from (added to) $\bar{x}$ without making any component of x or z negative (except perhaps on a set of measure zero). Note that I_1 and I_2 cannot both be empty. We can now choose disjoint sets P_1, P_2 in P'', each with nonzero measure, and such that

$$\int_{P_1} f_1(t)\,dt=\int_{P_2} f_2(t)\,dt \quad \text{and} \tag{7}$$

$$\left|d_i\int_{P_j} f_j(t)\,dt\right|<\varepsilon, \qquad n_1<i\leqq n_1+n_2, \quad j=1, 2. \tag{8}$$

Define $\bar{x}'$ from (6) by setting

$$h(t)=\begin{cases} f_1(t), & t\in P_1,\\ -f_2(t), & t\in P_2,\\ 0, & \text{otherwise}\end{cases}$$

and define $\hat{x}'$ from $\bar{x}'$ in the usual way. From the definition of f_1 and f_2, x' and z' are nonnegative. Also, by (8), $\bar{x}$ is changed on a set which is small enough for y' to remain positive in I and, by (7), outside this interval y' is the same as y. Thus $\hat{x}'$ is nonnegative and hence feasible. Furthermore, $S_{\hat{x}'}(t)\subset S_{\hat{x}}(t)$, for almost all $t\in[0, T]$, with strict inclusion either for almost all t in P_1 or for almost all t in P_2 (and both if neither I_j is empty). This contradicts the assumption that $\hat{x}$ has minimal support.

(b) Suppose that $\hat{x}$ is feasible and basic but does not have minimal support. Then there is some feasible $\hat{x}'$ with $S_{\hat{x}'}(t)\subset S_{\hat{x}}(t)$, $t\in[0, T]$, with strict inclusion on a set on nonzero measure. Define

$$g(t)=\bar{x}(t)-\bar{x}'(t), \qquad t\in[0, T].$$

Then

$$Kg(t)=K\bar{x}(t)-K\bar{x}'(t)=0, \qquad t\in[0, T].$$

But except possibly on a set of measure zero, $g_i(t)=0$ unless i is in $S_{\hat{x}}(t)$, which contradicts the assumption that $\hat{x}$ is basic.

(a) and (b) together prove the equivalence of (i) and (ii). We now show that (i) and (iii) are equivalent.

(c) Suppose that $\hat{x}$ is an extreme point of F but is not basic. Construct a feasible $\hat{x}'$ in the same manner as was done in part (a) above. Then for δ chosen small enough, $\hat{x}+\delta(\hat{x}'-\hat{x})$ and $\hat{x}-\delta(\hat{x}'-\hat{x})$ are both feasible. This contradicts the assumption that $\hat{x}$ is an extreme point.

(d) Suppose that $\hat{x}$ is a basic feasible solution but is not an extreme point. Then there is some $\hat{x}'$ and $\hat{x}''$ in F with $\hat{x}$ an interior point of the line segment joining them. Define.

$$g(t)=\bar{x}'(t)-\bar{x}''(t), \qquad t\in[0, T].$$

Then

$$Kg(t)=K\bar{x}'(t)-K\bar{x}''(t)=0, \qquad t\in[0, T].$$

But

$$S_{\hat{x}}(t)=S_{\hat{x}'}(t)\cup S_{\hat{x}''}(t)$$

so $g_i(t)$ is zero if $i\notin S_{\hat{x}}(t)$ and $g(t)$ is nonzero on a set of measure greater than zero. This contradicts the assumption that $\hat{x}$ is a basic feasible solution.

2. A class of SCLP's with piecewise linear solutions. A common feature of SCLP's which arise in practice is that they have optimal solutions with $S_{\hat{x}}$ piecewise constant and having only a finite number of changes. An SCLP which has an optimal solution with this property is called *regional*. The behavior of an optimal solution depends on the functions a, b and c, and it appears difficult to find general conditions on these functions under which an SCLP is regional. However, a class of SCLP's which can be shown to be regional is discussed below. Moreover, for this class of

problems $\hat{x}$ is linear on intervals where $S_{\hat{x}}$ is constant, giving optimal solutions which are piecewise linear. We assume throughout this section that F is bounded so that there will always exist an optimal extreme point of F and hence a basic optimal solution.

THEOREM 3. *If b is constant, and*

$$a(t) = a_1 + a_2 t, \qquad t \in [0, T],$$

$$c(t) = c_1 + c_2 t, \qquad t \in [0, T],$$

where a_1, a_2, c_1, c_2 are constant vectors, then SCLP *has an optimal solution $\hat{x}$ with*

$$\hat{c}_2^T \hat{x}(t_1) \leqq \hat{c}_2^T \hat{x}(t_2), \tag{9}$$

for all $t_1 < t_2$, $t_1, t_2 \in [0, T] \backslash N$, where N has measure zero and

$$\hat{c}_2 = \begin{bmatrix} c_2 \\ 0 \\ 0 \end{bmatrix}.$$

Proof. Suppose that $\hat{x}$ is an optimal basic solution. Define $q(t)$ by

$$q(t) = \int_0^t \hat{c}_2^T \hat{x}(s)\, ds.$$

The proof proceeds by showing that q is a convex function. Suppose otherwise; then as q is continuous, there is some interval (t_1, t_2) such that

$$q(t) > r(t), \qquad t \in (t_1, t_2),$$

where

$$r(t) = q(t_1) + \frac{(t - t_1)}{(t_2 - t_1)} (q(t_2) - q(t_1)), \qquad t \in (t_1, t_2).$$

Now define x' by

$$x'(t) = \begin{cases} (t_2 - t_1)^{-1} \int_{t_1}^{t_2} x(s)\, ds, & t \in (t_1, t_2), \\ x(t) & \text{otherwise.} \end{cases}$$

Define $\hat{x}'$ from x' in the usual way. Then $\hat{x}'$ is feasible. Comparing objective functional values for $\hat{x}$ and $\hat{x}'$ gives

$$\begin{aligned} \int_0^T \hat{c}(s)^T \hat{x}'(s)\, ds - \int_0^T \hat{c}(s)^T \hat{x}(s)\, ds &= \int_{t_1}^{t_2} s c_2^T (x'(s) - x(s))\, ds \\ &= \int_{t_1}^{t_2} s \frac{d}{ds} (r(s) - q(s))\, ds \\ &= \int_{t_1}^{t_2} (q(s) - r(s))\, ds + t_2 (r(t_2) - r(t_1) - q(t_2) + q(t_1)) \\ &> 0, \end{aligned}$$

since $r(t_1) = q(t_1)$, $r(t_2) = q(t_2)$ and $r(s) < q(s)$ on (t_1, t_2). This contradicts the assumed optimality of x. Hence $q(t)$ is convex and so its derivative is monotonic increasing

(see, for example, Rockafellar (1979)). But its derivative is equal to $\hat{c}_2^T\hat{x}(t)$ almost everywhere, which establishes the theorem.

THEOREM 4. *Under the conditions of Theorem* 3, *there is an optimal solution for which x is a piecewise linear function.*

Proof. We begin by showing that $\hat{c}_2^T\hat{x}(t)$ is a step function. Note firstly that for $\hat{x}$ a basic solution, $\bar{x}(t)$ takes on one of finitely many values, say $\bar{x}^{(1)}, \cdots, \bar{x}^{(L)}$. This is because

(a) there are only finitely many choices of $S_{\hat{x}}(t)$ (at any t, independent of t);

(b) the choice of support for a basic solution uniquely determines $\bar{x}(t)$ from (5); and

(c) the right-hand side of (5) is assumed constant.

Secondly, the form of c_2 implies

$$\hat{c}_2^T\hat{x}(t) = \hat{c}_2^T\bar{x}(t).$$

Hence $\hat{c}_2^T\hat{x}(t)$ is finite valued, taking on possible values $\hat{c}_2^T\bar{x}^{(i)}$, $i = 1, \cdots, L$. That $\hat{c}_2^T\hat{x}(t)$ is a step function now follows from the monotonicity of $\hat{c}_2^T\hat{x}(t)$ established in Theorem 3.

Next, assume (by reordering if necessary) that

$$\hat{c}_2^T\bar{x}^{(i)} \leqq \hat{c}_2^T\bar{x}^{(j)}, \qquad i < j. \tag{10}$$

If the $\hat{c}_2^T\bar{x}^{(i)}$ are all distinct (10) holds with strict inequality, and the result is established since each step of $\hat{c}_2^T\hat{x}(t)$ then corresponds to $\bar{x}(t)$ being held constant at some $\bar{x}^{(i)}$ over an interval of time.

If (10) does not hold with strict inequality, choose a sequence $\{c_2^{(n)}\}$ approaching c_2 and such that (10) holds with strict inequality if c_2 is replaced by any $c_2^{(n)}$ in the definition of $\hat{c}_2$. Let $\hat{x}(t)^{(n)}$ be an optimal solution to the problem with c_2 replaced by $c_2^{(n)}$. From the above, $\hat{x}(t)^{(n)}$ can be chosen to satisfy the conditions of the theorem. Thus $\hat{x}(t)^{(n)}$ can be described by a vector $p^{(n)}$ in R^L, with

$$\bar{x}(t)^{(n)} = \bar{x}^{(i)} \quad \text{for almost all } t \in \left(\sum_{j=0}^{i-1} p_j^{(n)}, \sum_{j=0}^{i} p_j^{(n)}\right]$$

and $\hat{x}^{(n)}$ defined from $\bar{x}^{(n)}$ in the usual way. As each component of $p^{(n)}$ is bounded by T, we may choose a subsequence from $p^{(n)}$ converging to p, say. Define $\bar{x}(t)$ by

$$\bar{x}(t) = \bar{x}^{(i)}, \qquad t \in \left(\sum_{j=0}^{i-1} p_j, \sum_{j=0}^{i} p_j\right].$$

Then it is clear that $\hat{x}$ defined from $\bar{x}$ in the usual way is optimal for the original problem, and the theorem is proved.

Theorems 3 and 4 have extensions to problems in which a and c are piecewise linear, with a finite number of breakpoints. Suppose that $\{t_1, t_2, \cdots, t_{r-1}\}$ is the set of times at which either a or c has a breakpoint, with $t_0 = 0$, $t_r = T$. Then the proofs of Theorems 3 and 4 hold when applied between t_{i-1} and t_i, $i = 1, 2, \cdots, r$.

For a basic $\hat{x}$ define a *region* (t_1, t_2) to be an interval over which the support of $\hat{x}$ is constant. It is worth pointing out that Theorem 4 allows the SCLP to be solved as a quadratic program. This can be done by choosing as variables the lengths of the regions. The theorem implies that there are only a finite number of regions which occur in fixed order. The objective function is quadratic in the region lengths and the constraints are linear. However, the size of this quadratic program grows rapidly with n_2 and n_3; the number of variables is approximately factorial $(n_2 + n_3)$.

REFERENCES

[1] R. BELLMAN, *Dynamic Programming*, Princeton Univ. Press, Princeton, NJ, 1957.
[2] W. P. DREWS, *A simplex-like algorithm for continuous-time linear optimal control problems*, Optimisation Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 309–322.
[3] R. GRINOLD, *Symmetry duality for a class of continuous linear programming problems*, SIAM J. Appl. Math., 18 (1970), pp. 84–97.
[4] R. J. HARTBERGER, *Representation extended to continuous time*, Optimisation Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 297–307.
[5] R. S. LEHMAN, *On the continuous simplex method*, RM-1386, Rand Corporation, 1954.
[6] N. LEVINSON, *A class of continuous linear programming problems*, J. Math. Anal. and Appl., 16 (1966), pp. 73–83.
[7] A. F. PEROLD, *Fundamentals of a continuous time simplex method*, Technical Report SOL 78-26, Stanford Univ., Stanford, CA, 1978.
[8] ———, *Extreme points and basic feasible solutions in continuous time linear programming*, this Journal, 19 (1981), pp. 52–63.
[9] R. J. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
[10] R. G. SEGERS, *A generalised function setting for dynamic optimal control problems*, Optimisation Methods for Resource Allocation, R. W. Cottle and J. Krarup, eds., Crane Russak and Co. Inc., New York, 1974, pp. 279–296.
[11] W. F. TYNDALL, *An extended duality theory for continuous linear programming problems*, SIAM J. Appl. Math., 15 (1967), pp. 1294–1298.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0008 $01.25/0

STABILIZATION AND STRUCTURAL ASSIGNMENT OF DIRICHLET BOUNDARY FEEDBACK PARABOLIC EQUATIONS*

I. LASIECKA† AND R. TRIGGIANI‡

Abstract. A parabolic equation defined on a bounded domain is considered, with input acting on the boundary through the Dirichlet B.C. expressed as a specified finite dimensional feedback of the solution. The free system (zero B.C.) is assumed throughout to be unstable. Two main results are established. First, we provide in § 2 a novel proof that fully solves the corresponding boundary feedback stabilization problem (thereby removing an annoying assumption required by the technique of [Appl. Math. Optim., 6 (1980), pp. 201–220]: under certain natural algebraic conditions based on the finitely many unstable eigenvalues, we establish the existence of general boundary vectors, for which the corresponding feedback semigroup decays exponentially to zero in the $L_2(\Omega)$-operator norm (or, more generally, in the $H^s(\Omega)$-operator norm, $0 \leqq s < \frac{1}{2}$). However, most of the paper is devoted to the second problem, structural or spectral assignment, which is a natural question relevant to the selfadjoint case. Here, under the same algebraic condition plus mild extra conditions, we establish the existence of boundary vectors that yield a more refined and stronger result for the corresponding feedback solutions, in the form of the following desirable structural or spectral property: for positive times, the feedback solutions can be expressed as an infinite linear combination of decaying exponentials. A semigroup approach is employed for both problems, but the corresponding techniques of solution are vastly different.

Key words. stabilization, boundary feedback, parabolic equations

1. Introduction and statement of main results. Let Ω be a bounded open domain in R^ν with boundary Γ, assumed to be an $(\nu-1)$-dimensional variety with Ω locally on one side of Γ. Here, Γ may have finitely many conical points [K4]. Let $A(\xi, \partial)$ be a uniformly strongly elliptic operator of order two in Ω of the form

$$A(\xi, \partial) = \sum_{|\alpha| \leqq 2} a_\alpha(\xi)\partial^\alpha, \tag{1.0}$$

with smooth real coefficients a_α, where the symbol ∂, rather than the traditional D, denotes differentiation. In the present paper, the symbol D will denote the Dirichlet map, as defined below. We consider a diffusion system based on Ω with input applied on Γ; that is,

$$\frac{\partial x}{\partial t}(t, \xi) = -A(\xi, \partial)x(t, \xi) \quad \text{in } (0, T]\times\Omega, \tag{1.1}$$

$$x(0, \xi) = x_0(\xi), \qquad \xi \in \Omega, \tag{1.2}$$

$$x(t, \zeta) = f(t, \zeta) \qquad \text{in } (0, T]\times\Gamma \text{ (Dirichlet B.C.).} \tag{1.3}$$

Here, $f(t, \zeta)$ is the input function or control function (or forcing term) defined on $(0, T]\times\Gamma$ which influences the solution $x(t, \xi)$. Let $-A$ be the operator: $L_2(\Omega) \supset \mathscr{D}(-A) \to L_2(\Omega)$, defined by $(-A\phi)(\xi) = -A(\xi, \partial)\phi(\xi)$, for $\phi \in \mathscr{D}(-A)$, where $\mathscr{D}(-A)$

* Received by the editors March 21, 1980, and in final revised form May 24, 1982. This research was performed while the authors were visiting the Department of System Science, University of California, Los Angeles.

† Department of System Science, University of California, Los Angeles, California 90024, and Department of Mathematics, University of Florida, Gainesville, Florida 32611. The research of this author was supported in part by the Air Force Office of Scientific Research under grant AFOSR-78-3350.

‡ Department of Mathematics, Iowa State University, Ames, Iowa 50011, and Department of Mathematics, University of Florida, Gainesville, Florida 32611. The research of this author was supported in part by the Air Force Office of Scientific Research under grant AFOSR-77-3338 through I.S.U.

consists of the closure in $H^2(\Omega)$ of functions h in $C^2(\bar{\Omega})$ that satisfy the boundary condition $h|_\Gamma = 0$. (For a C^2-boundary Γ, we have $\mathscr{D}(-A) = H_0^1(\Omega) \cap H^2(\Omega)$.) The operator $-A$ generates an *analytic* (*holomorphic*) semigroup on $X = L_2(\Omega)$ [F1, Ex., p. 101], [P1, Thm. 2.9, p. 139], [D3, p. 1740] conveniently denoted by e^{-At}, $t \geqq 0$.

The feedback system. As in [T3], [T4], the distinctive new feature of the present paper is that we demand that the input function $f(t, \zeta)$ be expressed in a feedback form (see Fig. 1.1), as a linear operator (of finite-dimensional range) acting on the solution vector $x(t, \xi)$. For the present paper, we choose the feedback operator F to be a continuous operator from $L_2(\Omega)$ to a J-dimensional subspace of $L_2(\Gamma)$ of the form:

$$f(t, \zeta) = \sum_{j=1}^{J} \langle x(t, \cdot), w_j(\cdot) \rangle g_j(\zeta) \overset{\text{df}}{=} Fx(t, \cdot) \quad \text{on } (0, T] \times \Gamma. \tag{1.4}$$

Here, $w_j(\cdot)$ and $g_j(\cdot)$ are fixed vectors in $L_2(\Omega)$ and $L_2(\Gamma)$, respectively, and the symbol $\langle \cdot, \cdot \rangle$ denotes the inner product in $L_2(\Omega)$. The vectors g_j are assumed to be linearly independent.[1] For $J = 1$, we shall write w and g instead of w_1 and g_1.

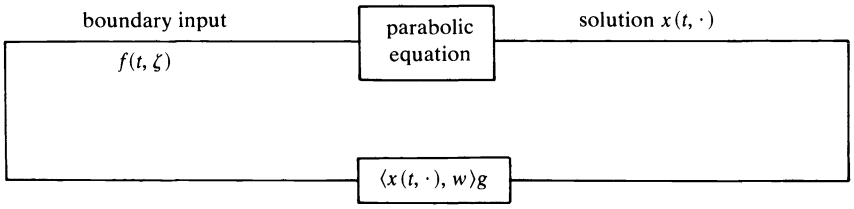


FIG. 1.1. *The feedback system.*

Since Ω is a bounded domain, the resolvent operator $R(\lambda, -A)$ is compact [D3, p. 1740]. Hence the spectrum $\sigma(-A)$ of $-A$ is only a point spectrum and consists of a sequence of isolated distinct eigenvalues $\{\lambda_k\}$, $k = 1, 2, \cdots, |\lambda_k| \to \infty$, with corresponding normalized linearly independent eigenvectors $\{\Phi_{km}\}$, $k = 1, 2, \cdots, M_k$ (M_k being the geometric multiplicity of λ_k). As is well known, since e^{-At} is analytic, the $\{\lambda_k\}$ are contained in a triangular sector delimited by the rays: $a + \rho e^{\pm i\theta}$, $0 \leqq \rho < \infty$, $\pi/2 < \theta < \pi$, a real, with no finite accumulation point. Thus, at the right of any vertical line in the complex plane, there are at most finitely many of them. Our standing assumption—for the problems considered in this paper to be significant—is that: there are $(K-1)$ eigenvalues $\lambda_1, \cdots, \lambda_{K-1}$ at the right of the imaginary axis ordered, say, by decreasing real parts

$$\operatorname{Re} \lambda_K < 0 \leqq \operatorname{Re} \lambda_{K-1} \leqq \cdots \leqq \operatorname{Re} \lambda_2 \leqq \operatorname{Re} \lambda_1. \tag{1.5}$$

Thus, the generator $-A$ is *unstable*, in the sense that there are *free* solutions (corresponding to $f(t, \zeta) \equiv 0$), that blow up in time, in fact exponentially.

[1] With minor technical changes, the results of this paper extend to cover the feedback

$$f(t, \zeta) = \sum_{j=1}^{J} ((\gamma x)(t, \cdot), w_j(\cdot))_\Gamma g_j(\zeta)$$

where γ denotes *any* continuous operator: $H^\sigma(\Omega) \to L_2(\Gamma)$ for any $\sigma < \frac{1}{2}$, and where now $w_j \in L_2(\Gamma)$ and $(\cdot, \cdot)_\Gamma$ is the $L_2(\Gamma)$-inner product. (The limitation on σ is sharp, as since for $\gamma x = x|_\Gamma$ the feedback problem is not well posed: see [T3, Remark 2.3].)

We then pose two problems, the second of which under stronger assumptions gives a stronger conclusion. A qualitative statement thereof is as follows:

(i) *Stabilization or stabilizability* (Theorem 1.2 for general A): Find (if possible) appropriate vectors $g_j \in L_2(\Gamma)$—as well as their minimum number—and the *least* conditions on the vectors $w_j \in L_2(\Omega)$ so as to guarantee that all corresponding feedback closed-loop solutions $x(t, x_0)$ decay exponentially as $t \to \infty$ in the strongest possible norm.

(ii) *Structural or spectral assignment* (Theorem 2.3 for selfadjoint A): Under the additional condition that the unstable $-A$ is selfadjoint (say, $-A(\xi, \partial) = \Delta + c^2$, for c sufficiently large), find, if possible, appropriate vectors g_j and the *least* conditions on the vectors $w_j \in L_2(\Omega)$, to guarantee that *all* corresponding feedback closed-loop solutions $x(t, x_0)$, due to initial data in naturally "large" subspaces, are in fact functions of the special class IDE for $t \geqq 0$.

Requirement (ii) means (see Definition 1.2) that such solutions can be expressed for $t \geqq 0$ as an infinite sum (in the strongest possible space topology) of countably many decaying exponentials. Thus, a fortiori, they decay as in (i).[2]

This structural problem is the major object of the present paper.

A solution to the stabilization problem was already presented in [T4], under an additional annoying assumption on the subspace of $L_2(\Omega)$ from which the vectors w_j may be drawn. Here, however, we present a new proof, which, while relying on the stabilization of the "finite dimensional part" given by [T4, Lemma 2.2], is radically different in its general conception, and thus manages to eliminate this unnecessary assumption. This new proof is essentially obtained as a quick consequence of the following recent result—unavailable at the time of writing of [T4]—on the existence of a feedback semigroup on $L_2(\Omega)$ for the closed-loop problem (1.1)–(1.4), which we quote from our paper [L7]. Actually, [L7] treats a technically more difficult case, while the treatment here is to be found in [L8, § 2].

THEOREM 1.1 [L7]. *The feedback closed-loop solutions $x(t, x_0)$ can be expressed simply as: $x(t, x_0) = S_F(t)x_0, x_0 \in L_2(\Omega), t \geqq 0$ where $S_F(t)$ defines a (feedback) semigroup on $L_2(\Omega)$, which is analytic and compact for $t > 0$ and whose generator A_F has a compact resolvent on $L_2(\Omega)$. Actually, the feedback semigroup extends/restricts to an analytic, compact semigroup $S_{F,\theta}(t)$ for positive times on each (fixed) interpolation space,*

$$[\mathscr{D}(A^{1/4-\rho}), [\mathscr{D}(A^{3/4+\rho})]']_\theta = \mathscr{D}(A^{1/4-\rho-\theta}), \qquad 0 \leqq \theta \leqq 1,$$

between $\mathscr{D}(A^{1/4-\rho}) = H^{1/2-2\rho}(\Omega)$ and $[\mathscr{D}(A^{3/4+\rho})]'$ (cf. identifications (2.3), and (2.3') in footnote 13, below). The corresponding $A_{F,\theta}$ has compact resolvent.[3]

In order to state our theorems on stabilization and structural assignment, we need to introduce a few quantities. Henceforth, we shall let X_u be the (unstable) subspace of $L_2(\Omega)$ corresponding to $\{\lambda_k\}$, $1 \leqq k \leqq K-1$, and the (stable) subspace X_s be its

[2] The structural assignment problem amounts to this requirement: should the sought-after vectors g_j exist, then the *original* system—which is unstable as a *free* system—once operating as a boundary feedback system with such g_j's in the feedback (1.4), has the same qualitative behavior as that of a free stable system.

[3] In the statement of this theorem, we may assume that the fractional powers A^θ, $0 \leqq \theta \leqq 1$, are well defined, for otherwise we simply replace A with a suitable translation, without affecting the local regularity in time. Moreover, the conventional notation $\mathscr{D}(A^{-s}) = [\mathscr{D}(A^s)]'$, $s > 0$, is used, along with $S_{F,\theta}(t) = S_F(t)$ and $A_{F,\theta} = A_F$ for $\theta = 1/4 - \rho$, i.e. on $L_2(\Omega)$. Finally, $\mathscr{D}(A^{1/4-\rho-\theta})$ is topologized by

$$|x|_{\mathscr{D}(A^{1/4-\rho-\theta})} = |A^{1/4-\rho-\theta}x|_{L_2(\Omega)}, \qquad 0 \leqq \theta \leqq 1.$$

Here and hereafter, ρ is a positive number, introduced in connection with (2.3) ff. that will be kept fixed throughout. Thus, dependence on ρ is not explicitly indicated.

orthogonal complement. We shall write $x_u = Px$ and $x_s = Qx$ for the orthogonal projectors P and Q onto X_u and X_s, respectively (see § 2).

First, the following important number:

DEFINITION 1.1.[4] Let the integer l_T $(1 \leqq l_T \leqq \dim X_u)$ denote the number of *linearly independent* Neumann traces $\{(\partial\Phi_{km}/\partial\eta)|_\Gamma\}$, $k = 1, \cdots, K-1, m = 1, \cdots, M_K$, of the normalized eigenfunctions associated with the *unstable* eigenvalues (1.5).

We next introduce the $J \times M_k$ matrix W_k defined in terms of $w_{ju} = Pw_j \in X_u$ by

$$W_k = \begin{vmatrix} \langle w_{1u}, \Phi_{k1}\rangle & \langle w_{1u}, \Phi_{k2}\rangle & \cdots & \langle w_{1u}, \Phi_{kM_k}\rangle \\ \langle w_{2u}, \Phi_{k1}\rangle & \langle w_{2u}, \Phi_{k2}\rangle & \cdots & \langle w_{2u}, \Phi_{kM_k}\rangle \\ \vdots & \vdots & & \vdots \\ \langle w_{Ju}, \Phi_{k1}\rangle & \langle w_{Ju}, \Phi_{k2}\rangle & \cdots & \langle w_{Ju}, \Phi_{kM_k}\rangle \end{vmatrix}, \qquad k = 1, \cdots, K-1, \tag{1.6}$$

and associated with each unstable eigenvalue of A, and moreover the $J \times (\dim X_u)$ matrix $W = [W_1, W_2, \cdots, W_{k-1}]$.

As to our results for the stabilization and for the structural assignment problems, they require, in an essential way, that the domain Ω be of dimension $\nu \geqq 2$. The one-dimensional case ($\nu = 1$) can be seen a priori to be hopeless in general (see [T4, Final Remark]). The stabilization result is:

THEOREM 1.2. *Let $\nu = \dim \Omega \geqq 2$ and let Ω either have*[5] *C^∞-boundary Γ, or else be a parallelepiped. Let the operator $-A$ have its (necessarily) point spectrum satisfy the instability condition* (1.5). *Let the restriction $-A_u$ of $-A$ on the unstable subspace X_u generated by the eigenvectors of $\{\lambda_k\}_{K=1}^{K-1}$ be diagonalizable*[6] *on X_u.*

Assume that the vectors $w_{ju} = Pw_j \in X_u$ are chosen so as to satisfy the full rank conditions

$$\text{rank } W_k = M_k, \qquad k = 1, \cdots, K-1, \tag{1.7a}$$

at the unstable eigenvalues and, moreover,

$$\dim X_u \leqq l_T + l_w - 1 \tag{1.7b}$$

where l_w is defined by

$$\text{rank } W = l_w, \qquad (\max\{M_k, k = 1, \cdots, K-1\} \leqq l_w).$$

Let $\varepsilon > 0$ be preassigned.

Then, there exist suitable vectors $g_j \in L_2(\Gamma)$, whose minimal number is discussed in Remark 2.2, *such that the feedback semigroup $S_{F,\theta}(t)$, claimed in Theorem* 1.1, *satisfies the following exponential decay*:

$$|S_{F,\theta}(t)|_\theta \leqq M_{\theta,\varepsilon}\, e^{(\operatorname{Re}\lambda_K + \varepsilon)t}, \qquad t \geqq 0, \tag{1.8}$$

with $|\cdot|_\theta$ the uniform operator norm from $\mathcal{D}(A^{1/4-\rho-\theta}) = H^{1/2-2\rho-2\theta}(\Omega)$ into itself, $0 \leqq \theta \leqq 1/4-\rho$, for all vectors w_{js} in a suitably small sphere of X_s, depending on ε and the g_j's. The vectors g_j are given by $g_j = \bar{g}_j + g_j^*$, where the $\bar{g}_j$'s are the solution of [L6, *moment*

[4] When Ω is a parallelepiped, $l_T = \dim X_u$, and conditions (1.7) below can always be satisfied. When Ω is a sphere, $l_T < \dim X_u$, but conditions (1.7) can still be satisfied: see [L6, Remark 3.3] for more details.

[5] This assumption on Γ is needed *only* to invoke [S1, Cor. 2.2] to guarantee [L6, L8, (3B.11), App.].

[6] The assumption that A_u is diagonalizable is retained only for the convenience of having "clean", easy-to-check tests such as (1.7)(a)–(b), expressed in terms of (not necessarily orthogonal) normalized eigenvectors. Otherwise, resort to the Jordon canonical form is necessary.

problem (3B.7), *Appendix* 3B], *unique in the space*

$$\mathcal{F} = \operatorname{sp}\left\{\left.\frac{\partial \Phi_{km}}{\partial \eta}\right|_{\Gamma}, k = 1, \cdots K-1, m = 1, \cdots M_k\right\}, \tag{1.9}$$

and the g_j^**'s are any vectors orthogonal to* $\mathcal{F}$.

The structural assignment problem demands for $t > 0$ a much fuller description of the feedback solutions with regard to their structure (and not merely to their norm upper bound, as in the stabilization problem). Thus, a technically and conceptually more elaborate tour de force is needed for its solution: this will be aimed at preserving the *form* of the closed-loop solution throughout, in order to force it to possess the desired structural property. All this counts for a major portion of its proof as compared, say, with [T4] or with § 2.

We now state our structural results by first singling out the technically simpler situation where all eigenvalues have geometric multiplicity equal to one.

DEFINITION 1.2. A (scalar or vector valued) function $z(t)$ of the form

$$z(t) = \sum_{k=1}^{\infty} a_k e^{\alpha_k t}, \qquad t \in R^+,$$

with α_k *negative* real numbers, where $\alpha_k \to -\infty$ as $k \to \infty$ and where $\sum_{k=1}^{\infty} |a_k| < \infty$, will, in this paper, be called a function of the special class IDE (infinite linear combinations of decaying exponentials), where $|\cdot|$ denotes either the absolute value or the appropriate space norm.

For instance, the functions $t^n e^{-t}$, $n = 1, 2, \cdots$, although exponentially decaying as $t \to +\infty$, are not of the special class IDE.

THEOREM 1.3 (feedback solutions of the special class IDE in a weak topology when $M_k \equiv 1$). *Let the same assumptions as made in Theorem* 1.2 *on* Ω, A, *and* w_{ju} *through* (1.7) *still apply, except that now* A *is specialized to be selfadjoint. For any* $\rho > 0$, *let*

$$0 \neq \langle w_{js}, \Phi_k \rangle \leqq \text{const}/k^{1+(2/\nu)(3/4+\rho)}, \qquad k = K, K+1, \cdots, \quad j = 1, \cdots J, \tag{1.10}^7$$

so that the vectors $\bar{w}_{js} = A_s^{3/4+\rho} w_{js} \in X_s$ *defined in the subsequent* (4.7) *satisfy*

$$0 \neq \langle \bar{w}_{js}, \Phi_k \rangle \leqq \text{const}/k, \qquad k = K, K+1, \cdots, \quad j = 1, \cdots J, \tag{1.10}'$$

(*as one sees via* (4.21)). *Then*

for all such vectors w_{js} *in a sufficiently small sphere of* X_s (1.11)

there exist boundary vectors $g_j \in L_2(\Gamma)$, *whose minimal number is discussed in Remark* 2.2, *such that the corresponding feedback solution* $x(t, x_0)$ *of the feedback system* (1.1), (1.2), (1.3) *and* (1.4), *with initial datum*[8]

$$x_0 \in \mathcal{D}((cI + A)^{1/4-\rho}) = H^{1/2-2\rho}(\Omega) \qquad \text{(see (2.3))} \tag{1.12}$$

has the following property: the scalar function

$$\langle (cI + A)^{1/4-\rho} x(t, x_0), y \rangle$$

is of the special class IDE *for any* $y \in L_2(\Omega)$. *Equivalently stated, the feedback solution* $x(t, x_0)$ *is of the special class* IDE *in the weak graph topology of* $\mathcal{D}((cI + A)^{1/4-\rho})$,

---

[7] Therefore, $w_j \in \mathcal{D}((cI + A)^{3/4+\rho})$ *by* (4.2.1).

[8] In the more general case of footnote 1, ρ is a positive number such that $\sigma < 1/2 - \rho$. In (1.12), c makes the fractional powers well defined.

equivalent to $H^{1/2-2\rho}(\Omega)$ *(see* (2.3)). *The vectors* g_j *are given as in the conclusion of Theorem* 1.2.

Actually, the proof will show the following more precise result, which displays a more subtle structure of the feedback solutions.

COROLLARY 1.4. *Under the assumptions of Theorem* 1.3, *the suitable feedback solutions, which are claimed in the conclusion, can be written for* $t \geqq 0$ *as*:[9]

$$\langle (cI+A)^{1/4-\rho} x(t, x_0), y\rangle = \sum_{i=1}^{K-1} b_i e^{c_i t} + \sum_{r=1}^{\infty} \gamma_r e^{\alpha_r t} \tag{1.13}$$

where[10] *the* $\{c_i\}_{i=1}^{K-1}$ *are negative constants, which can be preassigned in any chosen interval (in particular, at the left of* λ_K), *that replace the unstable eigenvalues* $\lambda_1, \cdots, \lambda_{K-1}$; *the* $\{\alpha_r\}$ *are a suitable sequence of negative constants having the same asymptotic behavior as the* $\{\lambda_r\}$: $[\alpha_r - \lambda_r] \to 0$ *as* $k \to \infty$ *(see* (4.38)). *Moreover, the coefficients* $\{\gamma_r\}$ *are in* l_1 *and along with the coefficients* $\{b_i\}$ *are explicitly exhibited in the proof as dependent on y, the initial datum in* $H^{1/2-2\rho}(\Omega)$, *and on the system parameters, including the sought-after vectors* $g_j \in L_2(\Gamma)$: *see* (4.59) *and also* (4.58′) *which depend on* $\{d_r\}$. *The sequence* $\{d_r\}$ *is related to the sequence* $\{n_r\}$ *by* (4.46)–(4.47) *which, in turn, is related to the initial point and the system's parameters via* (4.16).

An expansion similar to (1.13) *holds, this time in the weak topology of* $L_2(\Omega)$, *if the initial datum is only assumed in* $L_2(\Omega)$.

A reformulation of expansion (1.13) in terms of spectral properties of the feedback generator is given next. To appreciate it, one should note that the feedback generator A_F, corresponding to those special vectors w_j and g_j, as in Theorem 1.3, that produce feedback solutions as in (1.13), *cannot be a selfadjoint operator*,[11] so that an orthonormal basis in $L_2(\Omega)$ of eigenvectors of A_F is out of the question (with nonzero feedback). On the other hand, for any vectors w_j and g_j, the generalized eigenvectors of the corresponding feedback generator A_F always span all of $L_2(\Omega)$ (cf. subsequent Remark 1.1). With the special vectors w_j and g_j of Theorem 1.3, the situation achieved for the corresponding A_F falls in between these two cases.

COROLLARY 1.5. *The following spectral properties hold for the feedback generator* A_F *corresponding to the vectors* w_j *and* g_j *of Theorem* 1.3 (*or Corollary* 1.4):

(i) *The distinct constants* $\{c_i\}_{i=1}^{K-1}$ *and* $\{\alpha_r\}_{r=1}^{\infty}$ *are the eigenvalues of such feedback generator* A_F.

(ii) *The corresponding (normalized) eigenvectors* $\{e'_{F,i}\}_{i=1}^{K-1}$ *and* $\{e_{F,r}\}_{r=1}^{\infty}$ *form a (Schauder) basis in* $L_2(\Omega)$ (*nonorthogonal, when the* g_j*'s or the* w_j*'s are not all zero*) *so that the following expansions apply*:

$$x = \sum_{i=1}^{K-1} \eta'_i(x)\, e'_{F,i} + \sum_{r=1}^{\infty} \eta_r(x)\, e_{F,r}, \qquad x \in L_2(\Omega), \tag{1.14}$$

$$A_F x = \sum_{i=1}^{K-1} c_i \eta'_i(x)\, e'_{F,i} + \sum_{r=1}^{\infty} \alpha_r \eta_r(x)\, e_{F,r}, \qquad x \in \mathscr{D}(A_F), \tag{1.15}$$

---

[9] Note that the right-hand side of (1.13) is analytic in $\operatorname{Re} t > 0$ consistently with Theorem 1.1 for $\theta = 1/2$.

[10] As the proof shows, the constants $\{c_1\}_{i=1}^{K-1}$ and $\{\alpha_r\}_{r=1}^{\infty}$ are distinct from each other: for definiteness we preassign the c_i's as in (4.0).

[11] Since, in this case, A_F would be dissipative, contrary to Appendix 1A. By Green's second theorem for all $x, y \in \mathscr{D}(A_F)$ (and $J = 1$), $\langle A_F x, y\rangle - \langle x, A_F y\rangle = \langle x, w\rangle(g, \partial y/\partial n|_\Gamma)_\Gamma - \langle y, w\rangle(g, \partial x/\partial n|_\Gamma)_\Gamma$ and the identical vanishing of the right-hand side implies either $g = 0$ or else $w = 0$.

where the bounded linear functionals $\{\eta_r\}$ *and the eigenvectors* $\{e_{F,r}\}$ *are biorthogonal sequences*:

$$\eta_n(e_{F,m}) = \begin{cases} 1, & n = m, \\ 0, & n \neq m, \end{cases}$$

and, similarly, $\{\eta_i'\}$ *and* $\{e_{F,i}'\}$.
Thus

$$S_F(t)x_0 = \sum_{i=1}^{K-1} \eta_i'(x)\, e^{c_i t} e_{F,i}' + \sum_{r=1}^{\infty} \eta_r(x)\, e^{\alpha_r t} e_{F,r}.$$

The general case, where $M_k \not\equiv 1$, can be treated similarly. However, its detailed treatment would have considerably overloaded the presentation, particularly at the notational level. It is therefore analyzed only in the first part of the proof (§§ 2 and 3), while the more technical part of the proof (§ 4) is restricted to the case $M_k \equiv 1$.

Remark 1.1. Let A_F denote the generator of the feedback semigroup $S_F(t)$ on $L_2(\Omega)$, claimed by Theorem 1.1. Thus, for $J = 1$:

$$\mathscr{D}(A_F) = \{x \in L_2(\Omega)\colon A_F x \in L_2(\Omega) \text{ and } x|_\Gamma = \langle x, w\rangle g\}.$$

On the other hand, on the basis of the *differential* version (cf. [L8, Eq. (2.7)]) of the boundary feedback equation, one readily deduces that A_F can be characterized more explicitly as:

$$\text{(1.16)} \quad \begin{aligned} &\text{(i)} && A_F = -A(I - Dg\langle \cdot, w\rangle), \\ &\text{(ii)} && \mathscr{D}(A_F) = \{x \in L_2(\Omega)\colon x - Dg\langle x, w\rangle \in \mathscr{D}(A)\}. \end{aligned}$$

One important consequence of the explicit factorization of A_F in (1.4)(i) is the following:

CLAIM. *For all vectors* $g \in L_2(\Gamma)$ *and all vectors* $w \in L_2(\Omega)$, *the operator* A_F *in* (1.16) *has the following spectral property*:

$$\overline{\text{sp}}\{\textit{generalized eigenvectors of } A_F\} = L_2(\Omega)$$

(*the closure of the span is in the* $L_2(\Omega)$*-topology*).

The above claim simply follows from [D3, Vol III, general Theorem, p. 2374], which is applicable to the operator A_F in (1.14) since:

(i) A is selfadjoint in $L_2(\Omega)$.

(ii)[12] A^{-m} is of Hilbert–Schmidt type for sufficiently large positive integer m, as follows from the known asymptotic estimates of the eigenvalues $\{\lambda_n\}$ of A (cf. (4.21)):

$$\|A^{-m}\|_{H-S}^2 = \sum_{n=1}^{\infty} \frac{1}{\lambda_n^{2m}} \sim \sum_{n=1}^{\infty} \frac{1}{n^{4m/\nu}} < \infty \qquad (\text{for } 4m > \nu).$$

(iii) The perturbation $Dg\langle \cdot, w\rangle$ is compact in $L_2(\Omega)$.

The enlightening fact, that the feedback semigroup which arises in the stabilization and structural assignment problems is generally *not* a contraction, is illustrated in Appendix 1A.

2. Preliminaries and proof of the stabilization Theorem 1.2.

2.1. Preliminaries. The starting point of the approach taken in this paper, in studying the boundary feedback closed-loop system (1.1)–(1.3), (1.4), is based on a recently developed operator theoretic model that aims at describing, through a variation of parameter type formulas, nonsmooth boundary input open-loop systems like

[12] Without loss of generality, we assume here that A^{-1} is well defined as a bounded operator on $L_2(\Omega)$.

(1.1)–(1.3). Such a model is a semigroup rooted abstract version of (1.1)–(1.3), which takes the following input-solution integral form:

$$x(t, x_0) = e^{-At}x_0 + \int_0^t A\, e^{-A(t-\tau)}Df(\tau)\, d\tau. \tag{2.1}$$

Here f is given in a certain "time-space" function space and determines x as an element of another function space (say $f \in L_2(0, T; L_2(\Gamma)) \to x \in L_2(0, T; H^{1/2}(\Omega))$ (see the original account in [B1], [B2], [W1] and the very general treatment in [L1]). Here the operator D is the "Dirichlet map", that is (see [N1, Thm. 1.2, p. 250]), the continuous linear map: $L_2(\Gamma) \to H^{1/2}(\Omega)$ which solves the homogeneous elliptic problem corresponding to (1.1)–(1.3). It is defined by $y = Dg$, where $-A(\xi, \partial)y = 0$ in Ω; $y|_\Gamma = g$. Thus, in case of the feedback input (1.4), the abstract semigroup version of the closed loop system (1.1)–(1.4) under study becomes the following integral equation:

$$x(t) = e^{-At}x_0 + \int_0^t A\, e^{-A(t-\tau)}D \sum_{j=1}^{J} \langle x(\tau), w_j\rangle g_j\, d\tau. \tag{2.2}$$

As a consequence of the statement: $Df \notin \mathcal{D}(-A)$, $f \in L_2(\Gamma)$ unless $f = 0$, any attempt to provide a *differential* version of (2.1), or (2.2), in the form $\dot{x} = (-A + \Pi_1)x$ in the space $L_2(\Omega)$ is bound to fail. For a differential version in factor form $\dot{x} = -A(I + \Pi_2)x$, see Remark 1.1. What is needed is an extension of (2.2) to a space larger than $L_2(\Omega)$, which we will identify below. To do this in our present context, we find it expedient first to decompose the space $L_2(\Omega)$. Following a procedure introduced in [T2], we let $X = L_2(\Omega)$ be decomposed into two orthogonal subspaces X_u and X_s corresponding, respectively, to the subsets $\{\lambda_1, \cdots, \lambda_{K-1}\}$ and $\{\lambda_K, k \geqq K\}$ of the spectrum of $-A$ that satisfies assumption (1.5). (The subscripts u and s stand for "stable" and "unstable", respectively.) Here, we appeal to the standard decomposition theorem as in [K1, Thm. 6.17, p. 178]. With P denoting the orthogonal projection of $L_2(\Omega)$ onto X_u and $Q = I - P$, then $Q\mathcal{D}(-A) \supset \mathcal{D}(-A)$, X_u and X_s are invariant under $-A$ and hence under the semigroup e^{-At}. As for the spectra, we have $\sigma(-A_u) = \{\lambda_1, \cdots, \lambda_{K-1}\}$, $\sigma(-A_s) = \{\lambda_k, k \geqq K\}$, where $-A_\cdot$ is the restriction of $-A$ on $X_\cdot$, $-A_u$ is bounded, in fact, finite-dimensional. Finally, P and Q commute with $-A$, hence with the semigroup e^{-At}. We shall henceforth use the notation $x_u = Px$ and $x_s = Qx$. Notice that the fractional powers of A_s are well defined. As observed above, for $g \neq 0$, we always have $QDg \notin \mathcal{D}(-A_s)$. However, the following relations, which we shall apply crucially, hold:[13]

$$\mathcal{D}(A_s^{1/4-\rho}) = QH^{1/2-2\rho}(\Omega), \qquad 0 < \rho \leqq \tfrac{1}{4}, \tag{2.3}$$

the identification being set theoretical and topological, with norm

$$|x|_{H^{1/2-\rho}(\Omega)} = |A_s^{1/4-\rho}x|, \qquad x \in QH^{1/2-\rho}(\Omega).$$

Relations (2.3) are contained in the literature of fractional powers [F2], [L5], [L1, App. B], [M2, p. 189]. Now, elliptic theory [L4, pp. 187–188, N1] shows that:

$$\text{range of } D \subset H^{1/2}(\Omega), \tag{2.4}$$

[13] Moreover, one also has

$$\mathcal{D}(A_s^{3/4-\rho}) = QH_0^{3/2-2\rho}(\Omega), \qquad 0 < \rho \leqq \tfrac{3}{4}, \quad \rho \neq \tfrac{1}{2} \tag{2.3'}$$

while of course $H_o^s(\Omega) = H^s(\Omega)$, $0 \leqq s \leqq \frac{1}{2}$[L4].

and from (2.3) we then obtain

$$Q[\text{range of } D] \subset QH^{1/2-2\rho}(\Omega) = \mathscr{D}(A_s^{1/4-\rho}), \qquad 0 < \rho \leqq \tfrac{1}{4}. \tag{2.5}$$

2.2. Proof of the stabilization Theorem 1.2. Having introduced the relevant machinery, we project (2.2) onto X_u and X_s. By virtue of (2.5), the projections of the solution $x(t)$ in (2.2) onto X_u and X_s are, respectively,

$$x_u(t) = e^{-A_u t}x_{0u} + \int_0^t e^{-A_u(t-\tau)} \sum_{j=1}^{J} A_u PDg_j[\langle x_u(\tau), w_{ju}\rangle + \langle x_s(\tau), w_{js}\rangle]\, d\tau, \tag{2.6}$$

(2.7)

$$x_s(t) = e^{-A_s t}x_{0s} + \int_0^t A_s^{3/4+\rho} e^{-A_s(t-\tau)} \sum_{j=1}^{J} A_s^{1/4-\rho} QDg_j[\langle x_s(\tau), w_{js}\rangle + \langle x_u(\tau), w_{ju}\rangle]\, d\tau.$$

Remark 2.1. These projections are generally coupled. They become decoupled if the vectors w_j are chosen in X_u, so that $w_j = w_{ju}$ and $w_{js} = 0$. This is the case that corresponds to the *simplest* solution of the problems under study and presumes unlimited freedom in choosing the vectors w_j. Throughout this paper, we shall, however, consider only the more challenging situation where the projections x_u and x_s are *coupled* with no further mention of the decoupled situation, where in fact proofs simplify considerably.

We can now provide the anticipated *differential* version of (2.7). It is known [P1, Thm. 2, 5, p. 135] that $-A_s$ generates a contraction (analytic) semigroup and thus, by Lumer–Phillips's theorem, both $-A_s$ and $-A_s^*$ are maximal dissipative. It follows (see e.g. [L7] for details) that A_s and A_s^* can be extended via isomorphism techniques. Thus, A_s extends as an isomorphism from $\mathscr{D}(A_s^{\alpha})$ onto $[\mathscr{D}(A_s^{1-\alpha})]'$ for all $\alpha: 0 \leqq \alpha \leqq 1$. Let henceforth α be frozen as $\alpha = \frac{1}{4} - \rho$ (cf. (2.3)) and let $\tilde{A}_s$ denote the *extension of* A_s *acting as an isomorphism from* $\mathscr{D}(A_s^{1/4-\rho})$ *onto* $[\mathscr{D}(A_s^{3/4+\rho})]'$ *and viewed as an unbounded operator on the basic space* $[\mathscr{D}(A_s^{3/4+\rho})]'$. Similarly, $\tilde{A}_s^{3/4+\rho}$ is an isomorphism from $X_s = QL_2(\Omega)$ onto $[\mathscr{D}(A_s^{3/4+\rho})]'$. The relevant topologies are:

$$|x|_{\mathscr{D}(A_s^{1/4-\rho})} = |A_s^{1/4-\rho}x|_{L_2(\Omega)}, \qquad |x|_{[\mathscr{D}(A_s^{3/4+\rho})]'} = |A_s^{-3/4-\rho}x|_{L_2(\Omega)}. \tag{2.8}$$

Similarly, one can extend the semigroup $e^{-A_s t}$ originally defined as an analytic, contraction semigroup on X_s, to an analytic, contraction semigroup on $[\mathscr{D}(A_s^{1-\alpha})]'$, $0 \leqq \alpha \leqq 1$. Since the resolvent $R(\lambda, A_s)$ is compact as an operator on $L_2(\Omega)$, the resolvent $R(\lambda, \tilde{A}_s)$ is compact as an operator on all of $[\mathscr{D}(A_s^{3/4+\rho})]'$: see [L7]. The conclusion of all this that matters here is that: the *differential* version of the infinite dimensional projection of (2.7) is:

$$\begin{aligned} \dot{x}_s &= -\tilde{A}_s x_s + \tilde{A}_s^{3/4+\rho} \sum_{j=1}^{J} A_s^{1/4-\rho} QDg_j[\langle x_s, w_{js}\rangle + \langle x_u, w_{ju}\rangle], \\ x_s(0) &= x_{s0} \in [\mathscr{D}(A_s^{3/4+\rho})]' \end{aligned} \tag{2.9}$$

with $x_s \in \mathscr{D}(A_s^{1/4-\rho})$, $\dot{x}_s \in [\mathscr{D}(A_s^{3/4+\rho})]'$, $A_s^{1/4-\rho}QDg_j \in X_s$. As we are seeking suitable boundary vectors $g_j \in L_2(\Gamma)$, which produce the desired behavior of the closed-loop solutions, we find it convenient to consider the projections (2.6) and (2.9) after setting

$$p_j \equiv A_u PDg_j \in X_u, \qquad g_j \equiv A_s^{1/4-\rho}QDg_j \in X_s, \qquad j = 1, \cdots, J, \tag{2.10}$$

and to think of the vectors p_j and q_j as, for the time being, just vectors in X_u and X_s, respectively, without any connection with the vectors g_j which generate them. The

question of synthesizing p_j and q_j through an appropriate g_j will be taken up later on. We set further

$$\text{(2.11)}\quad \begin{aligned} &\text{(a)} \qquad P_1 x_u \equiv \sum_{j=1}^{J} p_j \langle x_u, w_{ju}\rangle, \qquad P_2 x_s \equiv \sum_{j=1}^{J} p_j \langle x_s, w_{js}\rangle, \\ &\text{(b)} \quad Q_1 x_u \equiv \tilde{A}_s^{3/4+\rho} \sum_{j=1}^{J} q_j \langle x_u, w_{ju}\rangle, \qquad Q_2 x_s \equiv \tilde{A}_s^{3/4+\rho} \sum_{j=1}^{J} q_j \langle x_s, w_{js}\rangle, \end{aligned}$$

where P_i, $i=1, 2$, are linear bounded operators from X_u, respectively X_s, into X_u, and Q_i, $i=1, 2$, are linear bounded operators from X_u, respectively X_s, into $[\mathscr{D}(A_s^{3/4+\rho})]'$. By means of (2.10)–(2.11), we thus combine the projections (2.6), (2.9) into the following differential operator equation:

$$\text{(2.12)}\quad \frac{d}{dt}\begin{vmatrix} x_u \\ x_s \end{vmatrix} = \begin{vmatrix} -A_u & 0 \\ 0 & -\tilde{A}_s \end{vmatrix} \begin{vmatrix} x_u \\ x_s \end{vmatrix} + \begin{vmatrix} P_1 & P_2 \\ Q_1 & Q_2 \end{vmatrix} \begin{vmatrix} x_u \\ x_s \end{vmatrix}$$

with $(x_u, x_s) \in X_u \times \mathscr{D}(A_s^{1/4-\rho})$ and $(\dot{x}_u, \dot{x}_s) \in X_u \times [\mathscr{D}(A_s^{3/4+\rho})]'$, and finally into the differential equation

$$\text{(2.13)}\quad \frac{d}{dt}\begin{vmatrix} x_u \\ x_s \end{vmatrix} = \tilde{A}_F \begin{vmatrix} x_u \\ x_s \end{vmatrix},$$

where we have introduced the (feedback) operator

$$\text{(2.14)}\quad \tilde{A}_F = \begin{vmatrix} -A_u + P_1, & P_2 \\ Q_1, & -\tilde{A}_s + Q_2 \end{vmatrix}$$

generating an *analytic* semigroup on $X_u \times [\mathscr{D}(A_s^{3/4+\rho})]'$. (Essentially (see [L7] for details) this is so since the operator $\begin{vmatrix} P_1 & P_2 \\ Q_1 & Q_2 \end{vmatrix}$ is bounded with respect to $\begin{vmatrix} -A_u & 0 \\ 0 & -\tilde{A}_s \end{vmatrix}$ with relative bound equal to zero, for the entries P_i, Q_i have finite dimensional range. Thus, the standard perturbation result as in [K1, p. 497] applies.) Tedious but straightforward computations then yield the resolvent operator $R(\lambda, \tilde{A}_F)$ as given by

$$\text{(2.15)}\quad R(\lambda, \tilde{A}_F) = \begin{vmatrix} ① & ② \\ ③ & ④ \end{vmatrix}$$

where, with

$$\text{(2.16)}\quad \Pi_\lambda \equiv (\lambda I_u + A_u - P_1) - P_2(\lambda I_s + \tilde{A}_s - Q_2)^{-1} Q_1,$$

which is a bounded operator $X_u \to$ itself, the entries are

$$\text{(2.17)}\quad \begin{aligned} &① = \Pi_\lambda^{-1} : X_u \to X_u, \\ &② = \Pi_\lambda^{-1} P_2 (\lambda I_s + \tilde{A}_s - Q_2)^{-1} : [\mathscr{D}(A_s^{3/4+\rho})]' \to X_u, \\ &③ = (\lambda I_s + \tilde{A}_s - Q_2)^{-1} Q_1 \Pi_\lambda^{-1} : X_u \to [\mathscr{D}(A_s^{3/4+\rho})]', \\ &④ = (\lambda I_s + \tilde{A}_s - Q_2)^{-1} \{I + Q_1 \Pi_\lambda^{-1} P_2 (\lambda I_s + \tilde{A}_s - Q_2)^{-1}\} : [\mathscr{D}(A_s^{3/4+\rho})]' \to \text{itself}. \end{aligned}$$

Thus, $R(\lambda, \tilde{A}_F)$ exists as a well-defined bounded operator on all of $X_u \times [\mathscr{D}(A_s^{3/4+\rho})]'$ for those λ and only for those λ for which its four entries ①, $\cdots$, ④ are simultaneously well defined; i.e., by (2.17), for which

$$\text{(2.18)}\quad [\Pi_\lambda^{-1} \text{ and } (\lambda I_s + \tilde{A}_s - Q_2)^{-1} \text{ are simultaneously well defined}].$$

Our analysis of Π_λ^{-1} will be based, by (2.16), on the study of the inverse $(\lambda I_u + A_u - P_1)^{-1}$ combined with a perturbation argument when P_2 is "small" (i.e. w_{js} is small). The next lemma describes the situation for $(\lambda I_u + A_u - P_1)$ in a form that will also be used later on in § 4.

LEMMA 2.1. *With the generator $-A$ satisfying the instability condition* (1.5), *let the restriction $-A_u$ of $-A$ on the unstable subspace X_u be diagonalizable on X_u. Assume that the vectors w_j satisfy the algebraic conditions* (1.7) (a)–(b) *at the unstable eigenvalues. Then*

(i) *There exist vectors p_j, $j = 1, \cdots, J$, in X_u such that the corresponding matrix*[14] $\bar{A}_u \equiv -A_u + P_1$, *with P_1 given by* (2.11) (a), *has a set of eigenvalues arbitrarily close to any preassigned set of* $(\dim X_u)$ *complex numbers (appearing in complex conjugate pairs, if A_u and P_j are all real).*

In particular, these eigenvalues of $\bar{A}_u$ may be preassigned to be all distinct and equal to negative constants c_i, $i = 1, \cdots, \dim X_u$, strictly on the left of the vertical line through $\operatorname{Re} \lambda_K$, *in which case the solution to the unperturbed part of* (2.6), *i.e.,*

$$\dot{z} = -A_u z + \sum_{j=1}^{J} A_u P D g_j \langle z, w_{ju} \rangle, \qquad z \in X_u, \quad z(0) = x_{0u}, \tag{2.19}$$

or $\dot{z} = \bar{A}_u z$, is

$$z(t) = e^{\bar{A}_u t} x_{0u} = \sum_{i=1}^{\dim X_u} e^{c_i t} \langle x_{0u}, \psi_i \rangle \psi_i. \tag{2.20}$$

Here, ψ_i is the normalized eigenvector of $\bar{A}_u$ corresponding to the simple eigenvalue c_i and the $\{\psi_i\}$, $i = 1, \cdots, \dim X_u$, form a basis on X_u.

(ii) *Moreover, when* $\dim \Omega = \nu \geqq 2$, *each vector p_j, $j = 1, \cdots, J$, can indeed be synthesized, as required by the left equality of* (2.10), *by any of the infinitely many vectors $g_j \in L_2(\Gamma)$, that satisfy* [L6, *App.* 3B, *moment problem* (3B.7)]. *The minimal number J required is discussed in Remark* 2.2 *below. The case* $\dim \Omega = \nu = 1$ *is also included, provided* $\dim X_u \leqq 3$.

Proof of Lemma 2.1. See [L6, App. 3B] or [T4, App.][15] where the proof is given in the t-domain. □

Remark 2.2. Conditions (1.7) (a) and the definition of l_w in particular imply

$$J \geqq \max \{M_k, k = 1, \cdots, K-1\} \quad \text{and} \quad J \geqq l_w.$$

Moreover, the proof given in [L6, App. 3B] shows that the largest multiplicity of the unstable eigenvalues is indeed the minimum number J_{min} of boundary vectors g_j required for the conclusion of Lemma 2.1, provided that the traces

$$\left\{ \frac{\partial \Phi_{km}}{\partial \eta} \bigg|_\Gamma \right\}, \qquad k = 1, \cdots, K-1, \quad m = 1, \cdots, M_k, \tag{2.21}$$

of the eigenfunctions are linearly independent.[16] Otherwise, more vectors $g_j \in L_2(\Gamma)$ are needed. For instance, if $M_k \equiv 1$, $1 \leqq k \leqq K-1$ and[17] $l_T < \dim X_u = K-1$, then J

[14] $\bar{A}_u$ is a square matrix of size equal to $\dim X_u$, depending on the vectors $A_u P D g_j$ and w_{ju} (besides A_u), as can be seen by using in X_u the (not necessarily orthogonal) basis of normalized eigenvectors Φ_{km}, $k = 1, \cdots, K-1$, which makes the matrix corresponding to the operator A_u diagonal.

[15] The proof in the Appendix of [T4] tacitly makes use of assumption (1.7b) which, however, was inadvertently omitted from the text.

[16] This is the case when Ω is a parallelepiped.

[17] This is the case when Ω is a sphere.

suitable vectors $g_j \in L_2(\Gamma)$ that satisfy the moment problem [L6, App. 3B, (3B.7)], where $J \geqq \dim X_u - l_T + 1$, will suffice.[18] A full analysis of the situation, amounting to a certain output stabilizability problem in X_u, is contained in the proof of [L6, App. 3B, Lemma 3.1], or of [T4].

The desired answer to (2.18) is now:

PROPOSITION 2.2. *Let the assumptions of Lemma* 2.1 *hold, and let* $\{g_j\}_{j=1}^J$ *be any J-tuple among the infinitely many possible guaranteed in the conclusion of Lemma* 2.1. *Let* $\varepsilon > 0$ *be preassigned. Then all vectors* w_{js} *in a suitably dependent small sphere of* X_s *have the property that the corresponding bounded operators* P_2 *and* Q_2 *in* (2.11) *are suitably small, so that both* Π_λ *and* $(\lambda I_s + \tilde{A}_s - Q_2)$ *are boundedly invertible for all* λ *in the right half plane*: $\operatorname{Re}\lambda \geqq \operatorname{Re}\lambda_K + \varepsilon$.

Hence, by (2.18), *the resolvent* $R(\lambda, \tilde{A}_F)$ *is well defined as a (bounded) operator on all of* $[\mathscr{D}(A_s^{3/4+\rho})]'$ *for all* λ *in the right half plane*: $\operatorname{Re}\lambda \geqq \operatorname{Re}\lambda_K + \varepsilon$.

Proof. The proof is a standard computation and therefore is omitted.

COROLLARY 2.3. *In the situation described by Proposition* 2.2, *we have*

$$\|e^{\tilde{A}_F t}\| \leqq M_\varepsilon e^{(\operatorname{Re}\lambda_K + \varepsilon)t}, \qquad t \geqq 0, \tag{2.22}$$

where $\|\ \|$ *is the uniform operator norm on* $[\mathscr{D}(A_s^{3/4+\rho})]'$.

Proof. The semigroup $e^{\tilde{A}_F t}$ is analytic on $[\mathscr{D}(A_s^{3/4+\rho})]'$ and hence obeys here the spectrum determined growth condition [T2]. Then the corollary follows from the conclusion of Proposition 2.2. □

We now restrict $e^{\tilde{A}_F t}$ to $L_2(\Omega)$, or more generally, to $X_u \otimes \mathscr{D}(A_s^{1/4-\rho-\theta}) \equiv H^{1/2-2\rho-2\theta}(\Omega)$ (see (2.3)), with ρ fixed by (2.5) and $0 \leqq \theta \leqq \frac{1}{4} - \rho$, the case of interest. We thus obtain a still-analytic semigroup, whose generator we denote, as in Theorem 1.1, by A_F on $L_2(\Omega)$, or more generally by $A_{F,\theta}$ on $H^{1/2-2\rho-2\theta}(\Omega)$: cf. [L7] for details.

PROPOSITION 2.4. *In the conditions described by Proposition* 2.2, *we have for any* $x \in H^{1/2-2\rho-2\theta}(\Omega)$, $0 \leqq \theta \leqq 1/4 - \rho$:

$$R(\lambda, \tilde{A}_F)x = R(\lambda, A_{F,\theta})x \in H^{1/2-2\rho-2\theta}(\Omega), \qquad \operatorname{Re}\lambda \geqq \operatorname{Re}\lambda_K + \varepsilon,$$

i.e., the restriction of $R(\lambda, \tilde{A}_F)$ *over* $H^{1/2-2\rho-2\theta}(\Omega)$ *coincides with* $R(\lambda, A_{F,\theta})$. *Therefore,* $R(\lambda, A_{F,\theta})$ *is well defined for all* λ *with* $\operatorname{Re}\lambda \geqq \operatorname{Re}\lambda_K + \varepsilon$.

Proof. We write down the proof for $\theta = 0$, the other cases being similar. By inspection of $R(\lambda, \tilde{A}_F)$ in (2.15)–(2.17), we see that we need only examine entry ④; i.e.,

$$(\lambda I_s + \tilde{A}_s - Q_2)^{-1}\{I + Q_1 \Pi_\lambda^{-1} P_2 (\lambda I_s + \tilde{A}_s - Q_2)^{-1}\}|_{X_s}$$

or

$$(\lambda I_s + \tilde{A}_s - Q_2)^{-1} y, \quad y \in X_s, \quad \text{and} \quad (\lambda I_s + \tilde{A}_s - Q_2)^{-1} Q_1 z, \quad z \in X_u. \tag{2.23}$$

Using the identity

$$(\lambda I_s + \tilde{A}_s - Q_2)^{-1} = [I_s - R(\lambda, -\tilde{A}_s)Q_2]^{-1} R(\lambda, -\tilde{A}_s), \tag{2.24}$$

we are led to study

$$R(\lambda, -\tilde{A}_s)Q_2 x_s = (\lambda I_s + \tilde{A}_s)^{-1} \tilde{A}_s^{3/4+\rho} A^{1/4-\rho} \sum_{j=1}^{J} QDg_j \langle x_s, w_{js} \rangle$$

[18] In the case of one-dimensional Ω, where J and l_T are at most equal to two, the unstable eigenspace cannot be of dimensions more than three.

(from (2.10)–(2.11) (b)) and deduce that

$$R(\lambda, -\tilde{A}_s)Q_2 x_2 \in \mathscr{D}(A_s^{1/4-\rho}) = QH^{1/2-2\rho}(\Omega),$$

since

$$A_s^{1/4-\rho}R(\lambda, -\tilde{A}_s)Q_2 x = (\lambda I_s + \tilde{A}_s)^{-1}\tilde{A}_s A^{1/4-\rho} \sum_{j=1}^{J} QDg_j\langle x_s, w_{js}\rangle \in QL_2(\Omega) = X_s.$$

Thus $[I_s - R(\lambda, \tilde{A}_s)Q_2]$: $\mathscr{D}(A_s^{1/4-\rho}) \to$ itself, and since $R(\lambda, -\tilde{A}_s)y \in \mathscr{D}(-\tilde{A}_s) = \mathscr{D}(A_s^{1/4-\rho})$ (see above (2.8)), we conclude from (2.24) that

$$(\lambda I_s + \tilde{A}_s - Q_2)^{-1} y \in \mathscr{D}(A_s^{1/4-\rho}) = QH^{1/2-2\rho}(\Omega),$$

as desired. As to the right-hand term in (2.23), we use again identity (2.24) and (2.10)–(2.11) (b) to deduce similarly

$$(\lambda I_s + \tilde{A}_s - Q_2)^{-1}Q_1 z = [I_s - R(\lambda, -\tilde{A}_s)Q_2]^{-1}R(\lambda, -\tilde{A}_s)\tilde{A}_s^{3/4+\rho}A_s^{1/4-\rho} \sum_{j=1}^{J} QDg_j\langle z, w_{ju}\rangle$$

$$\in \mathscr{D}(A_s^{1/4-\rho}) = QH^{1/2-2\rho}(\Omega),$$

and thus, as desired,

$$\text{range of } [\text{entry } ④|_{X_s}] \in \mathscr{D}(A_s^{1/4-\rho}) = QH^{1/2-2\rho}(\Omega). \qquad \square$$

As a consequence of Propositions 2.2 and 2.4, we then obtain

COROLLARY 2.5. *In the conditions described by Proposition* 2.2, *one has*

$$|S_{F,\theta}(t)|_\theta = |e^{A_{F,\theta}t}|_\theta \leqq M_{\theta,\varepsilon}\, e^{(\operatorname{Re}\lambda_K + \varepsilon)t}, \qquad t \geqq 0,$$

where $|\cdot|_\theta$ *is the uniform operator norm on* $H^{1/2-2\rho-2\theta}(\Omega)$, $0 \leqq \theta \leqq \frac{1}{4} - \rho$.

Proof. Theorem 1.1 guarantees that $e^{A_F t}$ is analytic on $H^{1/2-2\rho-2\theta}(\Omega)$ and thus obeys the spectrum determined growth condition [T2]. Corollary 2.5 then follows from Proposition 2.4. $\square$

To conclude the proof of Theorem 1.2, it remains to tackle the *synthesis* problem of the vectors p_j and q_j in the case where $\dim \Omega = \nu \geqq 2$.

First, Lemma 2.1 (ii) provides vectors $\bar{g}_j$, uniquely in the space $\mathscr{F}$ (see (1.9)). Then with

$$g_j = \bar{g}_j + g_j^*, \qquad g_j^* \perp \mathscr{F}, \tag{2.25}$$

we compute, as in [L6, App. 3B above Eq. (3B.3)]:

$$\begin{aligned} A_u PDg_j^* &= -\sum_{k=1}^{K-1} \lambda_k \langle Dg_j^*, \Phi_{km}\rangle \Phi_{km} \\ &= -\sum_{k=1}^{K-1} \lambda_k (g_j^*, D^*\Phi_{km})_{L_2(\Gamma)}\Phi_{km} \\ &= -\sum_{k=1}^{K-1} \left(g_j^*, \frac{\partial \Phi_{km}}{\partial \eta}\bigg|_\Gamma\right)_{L_2(\Gamma)} \Phi_{km} \\ &= 0. \end{aligned}$$

Thus, any such selection as in (2.25) produces vectors $p_j = A_u PDg_j = A_u PD\bar{g}_j$ which stabilize on X_u, as well as vectors $q_j \in X_s$ according to (2.10). With $\varepsilon > 0$ preassigned, then Corollary 2.3 guarantees the desired exponential decay for all vectors w_{js} in a suitably small sphere in X_s, with radius depending, among other things, on ε and the g_j's. The proof of Theorem 1.2 is now complete. $\square$

3. Structural assignment problem when $-A$ is selfadjoint. In this and the following section, we are concerned with Theorem 1.3 (and its generalizations) when the generator $-A$ is selfadjoint. This assumption is automatically guaranteed when the double-index coefficients a_α in (1.0) are real and symmetric [D3]. In this case, the point spectrum $\sigma(-A)$ of $-A$ consists of a sequence of *real* isolated eigenvalues $\{\lambda_k\}$ with no finite accumulation point: $\lambda_k \to -\infty$ and corresponding orthonormal eigenvectors $\{\Phi_{km}\}$, $k = 1, \cdots, m = 1, \cdots, M_k$, M_k being the geometric multiplicity of λ_k, that form a basis on $L_2(\Omega)$. The semigroup e^{-At} is then explicitly given by

$$e^{-At}x = \sum_{k=1}^{\infty} e^{\lambda_k t} \sum_{m=1}^{M_k} \langle x, \Phi_{km}\rangle \Phi_{km}, \qquad x \in L_2(\Omega), \quad t \geqq 0. \tag{3.1}$$

When all the $\{\lambda_k\}$ are negative, then the free (that is, $f(t, \zeta) \equiv 0$) open loop solutions $x(t, x_0) = e^{-At}x_0$ are functions of class IDE in the weak topology of $L_2(\Omega)$; i.e., the function $\langle x(t, x_0), y\rangle$, $t \in \overline{\mathbb{R}^+}$, $x_0 \in L_2(\Omega)$ is of the class IDE for all y in $L_2(\Omega)$. This remark motivated the structural assignment problem formulated in the introduction. In the spirit of Remark 2.1, we attack directly the more general situation where the projections (2.6)–(2.7) are coupled and leave off the decoupled case, which is obtained when all w_j can be constrained initially to be in X_u (simplest solution).

We collect below facts to be used in the sequel.

Remark 3.1. With reference to Lemma 2.1, a vector in X_u will always be referred to the basis $\{\psi_i\}_{i=1}^{\dim X_u}$. On the other hand, a vector in X_s will always be referred to the basis $\{\Phi_{km}\}_{k=K}^{\infty}$. We shall also adopt, henceforth, the following short notation:

$$\text{if } v \in X_u, \quad \text{we set } (v)_i = \langle v, \psi_i\rangle, \quad i = 1, \cdots, \dim X_u,$$

$$\text{if } v \in X_s, \quad \text{we set } [v]_{km} = \langle v, \Phi_{km}\rangle, \quad k = K, K+1, \cdots.$$

Remark 3.2. For handy reference below, we collect here the following results and observations. Let $\lambda_k \neq c_i$. Then,

$$\int_0^t e^{\lambda_k(t-\tau)} e^{c_i\tau}\, d\tau = \frac{e^{c_i t} - e^{\lambda_k t}}{c_i - \lambda_k}, \tag{3.2a}$$

and

$$\int_\sigma^t e^{\lambda_k(t-\tau)} e^{c_i(\tau-\sigma)}\, d\tau = \frac{e^{c_i(t-\sigma)} - e^{\lambda_k(t-\sigma)}}{c_i - \lambda_k}. \tag{3.2b}$$

In other words, convolving two different exponentials preserves the exponential character. By contrast, we have

$$\int_0^t e^{c(t-\tau)} e^{c\tau}\, d\tau = t e^{ct}.$$

Thus, convolving the *same* exponentials destroys its exponential character. These remarks will play a crucial role in the development given below in proving the desired structural properties of the feedback solutions.

4. Continuation of proof of Theorem 1.3 when $M_K \equiv 1$. The proof of Theorem 1.3 proceeds through a lengthy sequence of intermediate results. To simplify the technical exposition, we shall streamline the notation at the outset. As the geometric

multiplicity M_k of all eigenvalues is assumed identically one: $M_k \equiv 1$, we then consistently write Φ_k for Φ_{k1} throughout. With the constant ρ in (2.5) fixed once and for all, we also set[19]

$$A_s^{1/4-\rho} QDg = q = \{q_k\}_{k=K}^{\infty} \in X_s = QL_2(\Omega) \tag{4.1}$$

for the sought-after vector in X_s and

$$A_u PDg = p = \{p_i\}_{i=1}^{K-1} \in X_u = PL_2(\Omega) \tag{4.2}$$

for the corresponding vector in X_u, provided by Lemma 2.1. Here, according to the convention of Remark 3.2, we mean explicitly:

$$q_k = \langle q, \Phi_k \rangle, \quad k \geqq K, \quad \text{but} \quad p_i = \langle p, \Psi_i \rangle, \quad i = 1, \cdots, K-1. \tag{4.3}$$

We first consider the projections (2.6), (2.7) for $J = 1$, thinking at first of p and q in (4.1), (4.2) as, for the time being, just vectors in X_u and X_s, respectively, without any connection with the vector $g \in L_2(\Gamma)$ which generates them. The question of synthesizing *p and q* through an appropriate $g \in L_2(\Gamma)$ is then handled as at the end of the proof of Theorem 1.2, the stabilization theorem. Finally, throughout this section, the initial point x_{0s} is assumed to lie in $\mathscr{D}(A_s^{1/4-\rho})$, with no further explicit mention made; see (1.12).

4.1. Reduction to a Volterra integral equation in $\mathscr{A}(t) = \langle x_s(t), w_s \rangle$. Our starting point is the pair of projections (2.6), (2.7), with the notation of (4.1), (4.2), but now we rewrite the unperturbed part (2.19) of (2.6) in a more convenient form, as provided by (2.20):

$$x_u(t) = e^{\bar{A}_u t} x_{0u} + \int_0^t e^{\bar{A}_u (t-\tau)} p \langle x_s(\tau), w_s \rangle \, d\tau, \tag{4.4}$$

$$x_s(t) = e^{A_s t} x_{0s} + \int_0^t A_s^{3/4+\rho} e^{A_s (t-\tau)} q [\langle x_s(\tau), w_s \rangle + \langle x_u(\tau), w_u \rangle] \, d\tau. \tag{4.5}$$

Now define a vector $\bar{w}_s \in \mathscr{D}(A_s^{*-3/4-\rho}) \subset X_s$ by

$$\langle x_s(\tau), w_s \rangle = \langle A_s^{-3/4-\rho} x_s(\tau), \bar{w}_s \rangle = \langle x_s(\tau), A_s^{*-3/4-\rho} \bar{w}_s \rangle. \tag{4.6}$$

In other words, we use assumption (1.10)[20] on w, which implies $w_s \in \mathscr{D}(A^{*3/4+\rho})$, and we write

$$w_s = A_s^{*-3/4-\rho} \bar{w}_s \tag{4.7a}$$

or

$$\bar{w}_s = A_s^{*3/4+\rho} w_s. \tag{4.7b}$$

We then introduce the unknown function $\mathscr{A}(t)$:

$$\mathscr{A}(t) = \langle x_s(t), w_s \rangle = \langle A_s^{-3/4-\rho} x_s(t), \bar{w}_s \rangle. \tag{4.8}$$

Applying $A_s^{-3/4-\rho}$ to (4.5) and taking the inner product with $\bar{w}_s$ yields, by virtue of (4.8),

[19] To avoid unnecessary notational complications, we assume henceforth that Lemma 2.1 holds with only one vector, i.e., with $J = 1$. Moreover, for definiteness, and for the technical necessity to avoid "resonance" (cf. Remark 3.2), we shall henceforth assume that the constants $\{c_i\}_{i=1}^{K-1}$ provided by Lemma 2.1 are preassigned to lie in the interval $(\lambda_{K+1}, \lambda_K)$:

$$\lambda_{K+1} < c_{K-1} < \cdots < c_2 < c_1 < \lambda_K < 0. \tag{4.0}$$

[20] We would have $\mathscr{D}(A) = \mathscr{D}(A^*)$ even in the nonselfadjoint case [L4], and hence by interpolation $\mathscr{D}((cI + A)^\theta) = \mathscr{D}((cI + A^*)^\theta)$, $0 \leqq \theta \leqq 1$, since $-(cI + A)$ is maximal dissipative.

$$\mathscr{d}(t) = \langle e^{A_s t} A_s^{-3/4-\rho} x_{0s}, \bar{w}_s \rangle + \int_0^t \langle e^{A_s(t-\tau)} q, \bar{w}_s \rangle [\mathscr{d}(\tau) + \langle x_u(\tau), w_u \rangle] \, d\tau. \tag{4.9}$$

We next compute, by means of (4.4) and a change in the order of integration,

$$\int_0^t e^{A_s(t-\tau)} q \langle x_u(\tau), w_u \rangle \, d\tau = \int_0^t e^{A_s(t-\tau)} q \langle e^{\bar{A}_u \tau} x_{0u}, w_u \rangle \, d\tau + \int_0^t \int_\sigma^t e^{A_s(t-\tau)} q \langle e^{\bar{A}_u(\tau-\sigma)} p, w_u \rangle \, \mathscr{d}(\sigma) \, d\tau \, d\sigma.$$

By (3.1), (2.22) and the notational convention in Remark 3.2,

$$\begin{aligned}\int_0^t e^{A_s(t-\tau)} q \langle x_u(\tau), w_u \rangle \, d\tau &= \sum_{k=K}^{\infty} \sum_{i=1}^{K-1} \left\{ \int_0^t e^{\lambda_k(t-\tau)} e^{c_i \tau} (x_{0u})_i (w_u)_i \, d\tau \right. \\ &\qquad \left. + \int_0^t \int_\sigma^t e^{\lambda_k(t-\tau)} e^{c_i(\tau-\sigma)} (w_u)_i p_i \mathscr{d}(\sigma) \, d\tau \, d\sigma \right\} q_k \Phi_k \\ &= \sum_{k=K}^{\infty} \sum_{i=1}^{K-1} \left\{ \frac{e^{c_i t} - e^{\lambda_k t}}{c_i - \lambda_k} (x_{0u})_i (w_u)_i \quad \text{by (3.2)} \right. \\ &\qquad \left. + \int_0^t \frac{e^{c_i(t-\sigma)} - e^{\lambda_k(t-\sigma)}}{c_i - \lambda_k} \mathscr{d}(\sigma) \, d\sigma (w_u)_i p_i \right\} q_k \Phi_k.\end{aligned}$$

Hence, (4.9) becomes

$$\mathscr{d}(t) = \mathscr{n}(t) + \int_0^t \mathscr{h}(t-\tau) \mathscr{d}(\tau) \, d\tau, \qquad 0 \leqq t < \infty, \tag{4.10}$$

where, in the notational convention of Remark 3.2,

$$\begin{aligned}\mathscr{n}(t) &= \langle e^{A_s t} A_s^{-3/4-\rho} x_{0s}, \bar{w}_s \rangle + \sum_{k=K}^{\infty} \sum_{i=1}^{K-1} \frac{e^{c_i t} - e^{\lambda_k t}}{c_i - \lambda_k} (x_{0u})_i (w_u)_i [\bar{w}_s]_k q_k \\ &= - \sum_{i=1}^{K-1} e^{c_i t} (x_{0u})_i (w_u)_i \sum_{k=K}^{\infty} \frac{[\bar{w}_s]_k q_k}{\lambda_k - c_i} \qquad (4.11) \\ &\quad + \sum_{k=K}^{\infty} e^{\lambda_k t} [\bar{w}_s]_k \left\{ [A_s^{-3/4-\rho} x_{0s}]_k + q_k \sum_{i=1}^{K-1} \frac{(x_{0u})_i (w_u)_i}{\lambda_k - c_i} \right\},\end{aligned}$$

$$\begin{aligned}\mathscr{h}(t) &= \langle e^{A_s t} q, \bar{w}_s \rangle + \sum_{k=K}^{\infty} \sum_{i=1}^{K-1} \frac{e^{c_i t} - e^{\lambda_k t}}{c_i - \lambda_k} p_i (w_u)_i q_k [\bar{w}_s]_k \\ &= - \sum_{i=1}^{K-1} e^{c_i t} p_i (w_u)_i \sum_{k=K}^{\infty} \frac{q_k [\bar{w}_s]_k}{\lambda_k - c_i} + \sum_{k=K}^{\infty} e^{\lambda_k t} q_k [\bar{w}_s]_k \left\{ 1 + \sum_{i=1}^{K-1} \frac{p_i (w_u)_i}{\lambda_k - c_i} \right\}.\end{aligned} \tag{4.12}$$

We can rewrite (4.11) and (4.12) in a simplified manner as

$$\mathscr{n}(t) = \sum_{r=1}^{\infty} n_r e^{\beta_r t}, \qquad 0 \leqq t < \infty, \tag{4.13}$$

$$\mathscr{h}(t) = \sum_{r=1}^{\infty} h_r e^{\beta_r t}, \tag{4.14}$$

where

$$\beta_r = c_r, \quad r = 1, \cdots, K-1, \qquad \beta_r = \lambda_r, \quad r = K, K+1, \cdots, \tag{4.15}$$

$$n_r = -(x_{0u})_r (w_u)_r \sum_{k=K}^{\infty} \frac{q_k [\bar{w}_s]_k}{\lambda_k - c_r}, \qquad r = 1, \cdots, K-1, \tag{4.16a}$$

$$n_r = \frac{[A_s^{1/4-\rho} x_{0s}]_r [\bar{w}_s]_r}{\lambda_r} + q_r [\bar{w}_s]_r \sum_{i=1}^{K-1} \frac{(x_{0u})_i (w_u)_i}{\lambda_r - c_i}, \qquad r = K, K+1, \cdots. \tag{4.16b}$$

(Here, x_{0s} is written in the largest fractional power of A_s compatible with the assumption (1.12).) Furthermore,

$$h_r = -p_r(w_u)_r \sum_{k=K}^{\infty} \frac{q_k[\bar{w}_s]_k}{\lambda_k - c_r}, \qquad r = 1, \cdots, K-1, \tag{4.17a}$$

$$h_r = q_r[\bar{w}_s]_r \left\{1 + \sum_{i=1}^{K-1} \frac{p_i(w_u)_i}{\lambda_r - c_i}\right\}, \qquad r = K, K+1, \cdots. \tag{4.17b}$$

Notice that the existence of the solution $\mathscr{d}(t)$ in (4.10), as an analytic function for $t>0$, is already known through the existence result of Theorem 1.1. One may alternatively invoke the standard theory of linear Volterra integral equations [M1]. It should be kept in mind that the functions $\mathscr{n}(t)$, $\mathscr{h}(t)$, and hence $\mathscr{d}(t)$, depend upon q.

The above expressions will play a crucial role in the analysis, given below, of the Volterra equation (4.10). Notice that each coefficient n_r and h_r, for $r \geqq K$, depends only on the corresponding coordinate q_r, while for $1 \leqq r \leqq K-1$ it depends in a cumulative way on all $\{q_k\}_{k=K}^{\infty}$. This fact will be a source of difficulties. We also remark that we shall henceforth borrow freely from (4.15), both the notation $\{\beta_r\}$ in place of $\{c_r\}$ and $\{\lambda_r\}$ and the other way around.

4.2. Existence of admissible vectors q generating Volterra solutions $\mathscr{d}(t)$ of class IDE with negative exponents $\{\alpha_r\}_{r=1}^{\infty}$ all distinct from all $\{\beta_r\}_r^{\infty} = 1$. In order to establish that the solution $\mathscr{d}(t)$ of (4.10) is of class IDE for a suitable vector q, we find it convenient to associate with (4.10) the following sequence of auxiliary Volterra equations:

$$\mathscr{d}_N(t) = \mathscr{n}_N(t) + \int_0^t \mathscr{h}_N(t-\tau)\mathscr{d}_N(\tau)\, d\tau, \tag{4.18}$$

where $N = 1, 2, \cdots$, and

$$\mathscr{n}_N(t) = \sum_{r=1}^{N} n_r e^{\beta_r t}, \tag{4.19a}$$

$$\mathscr{h}_N(t) = \sum_{r=1}^{N} h_r e^{\beta_r t}, \tag{4.19b}$$

$$0 \leqq t < \infty.$$

LEMMA 4.1. *Let the initial point x_{0s} be in $\mathscr{D}(A_s^{1/4-\rho})$ and also let $q \in QL_2(\Omega)$. Then:*

(i) *The corresponding sequences*

$$(-A_s)\{n_r\}_{r=K}^{\infty} \overset{\text{df}}{=} \{\lambda_r n_r\}_{r=K}^{\infty} \quad \textit{and} \quad \{h_r\}_{r=K}^{\infty},$$

defined by (4.16), (4.17), *all belong to the space l_1; moreover,*

$$\lambda_r n_r \leqq \text{const}/r \quad \textit{and} \quad h_r \leqq \text{const}/r.$$

(ii) *The corresponding functions $\mathscr{n}(t)$ and $\mathscr{h}(t)$ are functions of class* IDE.

(iii) *The corresponding functions $\mathscr{n}_N(t)$ and $\mathscr{h}_N(t)$ in* (4.19) *converge uniformly over $\overline{\mathbb{R}^+}$ to the functions $\mathscr{n}(t)$ and $\mathscr{h}(t)$, respectively, in* (4.13) *and* (4.14).

Proof. Conclusion (i) is immediate from the explicit expressions (4.16) and (4.17) of the coefficients via (1.10′). As a consequence, $\mathscr{n}(t)$ and $\mathscr{h}(t)$ are the uniform limits over $\overline{\mathbb{R}^+}$ of the decaying exponentials thus establishing conclusion (ii), as required by Definition 1.2. Conclusion (i) also clearly implies (iii). □

We start with a general result which will be refined and complemented below in Theorem 4.9.

PROPOSITION 4.2. *For any vector* $q \in QL_2(\Omega)$, *the corresponding solutions* $d_N(t)$ *to the Volterra equation* (4.18) *converge uniformly over* $\mathbb{R}^+$ *to the corresponding solution* $d(t)$ *of the Volterra equation* (4.10).

Proof. Let $\hat{k}(s) = \sum_{r=1}^{\infty} h_r/(s-\beta_r)$ be the Laplace transform of $k(t)$ of (4.14) for $\operatorname{Re} s > 0$. By Lemma 4.1(i) on $\{h_r\}$, we can achieve $|1-\hat{k}(s)| > \rho_u > 0$ for $\operatorname{Re} s > u$, for a suitably large u. Then the (one-sided) Laplace transform $\hat{d}_N(s)$ of $\hat{d}_N(t)$ exists here by

$$\hat{d}_N(s) = \frac{\hat{n}_N(s)}{1-\hat{k}_N(s)}, \qquad N = K, K+1, \cdots,$$

and the uniqueness of the solution $d_N(t)$ to (4.18). Then, Lemma 4.1(iii) and the definition of the (one-sided) Laplace transform imply that, as $N \to \infty$, the functions $\hat{n}_N(s)$ and $\hat{k}_N(s)$ are uniformly convergent to $\hat{n}(s)$ and $\hat{k}(s)$, respectively, over $\operatorname{Re} s$ suitably large. Again, by the uniqueness property of the solution $d(t)$ to (4.10), the functions $\hat{d}_N(s)$ also converge uniformly over $\operatorname{Re} s$ suitably large to the function $\hat{d}(s)$. But then, the inverse Laplace integral (see [D.2, Thm. 24.4, p. 157]) implies that, as $N \to \infty$, $d_N(t)$ converges uniformly to $d(t)$ over $[0, \infty)$. □

We next establish some properties enjoyed by the Laplace transform of solutions $d(t)$.

PROPOSITION 4.3. *For any vector* $q \in QL_2(\Omega)$, *the Laplace transform* $\hat{d}(s)$ *of the corresponding solution* $d(t)$, *extended over the entire complex plane* $\mathbb{C}$ *in a natural way by the right-hand side of* (4.20) *below, is a meromorphic function over* $\mathbb{C}$.

Moreover, if, for a suitable q, *the corresponding* (*continuous over* $\overline{\mathbb{R}^+}$) *solution* $d(t)$ *of* (4.10) *is of class* IDE, *then* $\hat{d}(s)$ *has countably many simple poles* $\{\alpha_r\}$, α_r *real and negative, which are simple zeros for* $[1-\hat{k}(s)]$: $\hat{k}(\alpha_r) \equiv 1$. *Such poles are either finitely many or else, if infinitely many, have moduli tending to infinity*: $|\alpha_r| \to \infty$ *as* $r \to \infty$.

Proof. As in the preceding proof, we have explicitly, from (4.13), (4.14),

$$\hat{d}(s) = \frac{\hat{n}(s)}{1-\hat{k}(s)} = \frac{\sum_{r=1}^{\infty} n_r/(s-\beta_r)}{1-\sum_{r=1}^{\infty} h_r/(s-\beta_r)}, \tag{4.20}$$

which is the Laplace transform of $d(t)$ for $\operatorname{Re} s > 0$ and is extended to $\mathbb{C}$ by the expression on the right-hand side. As the ratio of two meromorphic functions over $\mathbb{C}$ (with common poles $\{\beta_r\}_{r=1}^{\infty}$, in fact), $\hat{d}(s)$ is meromorphic and hence its poles are either finitely many or else their moduli tend to infinity [K3]. In addition, $\hat{d}(s)$ admits an expansion as the sum of its principal parts plus an entire function [K3, Mittag–Leffler theorem]; [L3]. The poles of $\hat{d}(s)$ are zeros of the denominator $[1-\hat{k}(s)]$. If the real s_0 is such a zero with multiplicity m, then the term $t^{m-1}e^{s_0 t}$ occurs in the antitransform of the Mittag–Leffler expansion of $\hat{d}(s)$. Hence, the statement on the $\{\alpha_r\}$ is a consequence of the assumed IDE character of $d(t)$. □

Remark 4.1. The poles $\{\alpha_r\}$ cannot in general be finitely many.

With the above remark in mind, we proceed now to characterize the admissible vectors q (recall (4.1)) whose corresponding solutions $d(t)$ are functions of the class IDE with the additional requirement that the exponents be *all* different from the set $\{\beta_r\}_{r=1}^{\infty}$ in (4.15) (recall Remark 3.2).

Remark 4.2. We refer here to a basic known result on the asymptotic behavior of the eigenvalues of second-order elliptic selfadjoint differential operators, which will play a crucial role below. If ν denotes, as in the Introduction, the dimension of the euclidean space containing the domain Ω, then (see [T1, pp. 392–395], [C1, Ch. VI. §§ 3.3–3.4]) the estimate

$$\beta_k = \lambda_k \sim k^{2/\nu}, \qquad k = K, K+1, \cdots, \tag{4.21}$$

holds. Here, and hereafter, the symbol $\sim$ means that the left-hand side can be estimated by the right-hand side from below and from above with the aid of constants independent of the variable in question (k, in this case) going to infinity.

We now let m be the smallest (nonnegative) integer strictly greater than $((\nu/2)-1)$. Then, (4.21) implies

$$\sum_{k=1}^{\infty} \frac{1}{\beta_k^{m+1}} \sim \sum_{k=1}^{\infty} \frac{1}{k^{2(m+1)/\nu}} < \infty. \tag{4.22}$$

Therefore, by virtue of the Weistrass factorization theorem [L3, p. 390], the function $\mathscr{B}(s)$, defined by

$$\mathscr{B}(s) = \prod_{k=1}^{\infty} \left(1 - \frac{s}{\beta_k}\right) E(s, \beta_k, m) \tag{4.23}$$

where

$$E(s, \beta_k, m) \overset{\text{df}}{=} \exp\left(\frac{s}{\beta_k} + \frac{1}{2}\left(\frac{s}{\beta_k}\right)^2 + \cdots + \frac{1}{m}\left(\frac{s}{\beta_k}\right)^m\right), \tag{4.24}$$

is an entire function with zeros of multiplicity one precisely at the points $\{\beta_k\}_{k=1}^{\infty}$ and no other zero (the integer m is the genus of $\mathscr{B}(s)$). Then, the meromorphic function $\hat{\mathscr{I}}(s)$ in (4.20) can be rewritten as the ratio of two entire functions, in fact,

$$\hat{\mathscr{I}}(s) = \frac{\hat{n}(s)\mathscr{B}(s)}{\mathscr{D}(s)}$$

where

$$\mathscr{D}(s) = (1 - \hat{h}(s))\mathscr{B}(s), \tag{4.25}$$

while $\mathscr{D}(s)$ and $(1-\hat{h}(s))$ have precisely the same zeros. To motivate our further analysis, let us now assume, in the light of Proposition 4.3, that there exists a vector $q \in QL_2(\Omega)$ (this assumption will be shown later to be nonvoid) such that the corresponding function $(1-\hat{h}(s))$, obtained through (4.17), has countably many negative zeros, all simple, of the form $\{\alpha_k\}_{k=1}^{\infty}$, any α_k being different from all the $\{\beta_j\}_{j=1}^{\infty}$, but with a similar asymptotic behavior: $\alpha_k \sim \beta_k$. Then, the function

$$\prod_{k=1}^{\infty} \left(1 - \frac{s}{\alpha_k}\right) E(s, \beta_k, m)$$

is well defined and vanishes precisely at $\{\alpha_k\}_{k=1}^{\infty}$. By standard complex analysis theory [K3, p. 6], such a function differs from $\mathscr{D}(s)$ at most by a factor $e^{z(s)}$, where $z(s)$ is an entire function; that is,

$$\mathscr{D}(s) = e^{z(s)} \prod_{k=1}^{\infty} \left(1 - \frac{s}{\alpha_k}\right) E(s, \beta_k, m).$$

As a matter of fact, $e^{z(s)}$ must be a constant, and in fact equal to $A_\infty = \prod_{k=1}^{\infty} (\alpha_k/\beta_k)$, provided this infinite product is well defined, i.e. provided $(\alpha_k - \beta_k)/\beta_k \in l_1$. To see this, one writes

$$e^{z(s)} = \left(1 - \sum_{k=1}^{\infty} \frac{h_k}{s - \beta_k}\right)\left(\prod_{k=1}^{\infty} \frac{\alpha_k}{\beta_k}\right) \frac{\prod_{k=1}^{\infty} (1 - \beta_k/s)}{\prod_{k=1}^{\infty} (1 - \alpha_k/s)},$$

from which one obtains the limit value A_∞ by letting s go to infinity in any way except along the negative real axis; this leads to $e^{z(s)}$ as being $O(1)$ and hence, by Liouville's theorem, as being the constant A_∞. We have thus proved the first part of the following claim, whose assumption, as already remarked, will be shown later to be nonvoid.

PROPOSITION 4.4. *Let there exist a vector* $q = \{q_k\}_{k=K}^{\infty} \in QL_2(\Omega)$ *whose corresponding function* $\hat{k}(t)$ *in* (4.14) *obtained through the constants* $\{h_k\}_{k=1}^{\infty}$ *of* (4.17) *satisfies*

$$\hat{k}(\alpha_k) \equiv 1, \qquad k = 1, 2, \cdots,$$

with multiplicity one, for a negative sequence $\{\alpha_k\}_{k=1}^{\infty}$ *with*

$$\alpha_k \neq \beta_j, \quad \alpha_k \sim \beta_k \quad \text{and} \quad (\alpha_k - \beta_k)/\beta_k \in l_1, \qquad k = j = 1, 2, \cdots. \tag{4.26}$$

Then, the following identity over $\mathbb{C}$ *holds*:

$$(1 - \hat{k}(s))\mathcal{B}(s) = \mathcal{D}(s), \tag{4.27}$$

where $\mathcal{B}(s)$ *is defined by* (4.23) *and*

$$\mathcal{D}(s) = A_\infty \prod_{k=1}^{\infty} \left(1 - \frac{s}{\alpha_k}\right) E(s, \beta_k, m), \tag{4.28}$$

where

$$A_\infty = \prod_{k=1}^{\infty} \left(\frac{\alpha_k}{\beta_k}\right).$$

Moreover, the corresponding sequence h_r *is expressed by*

$$h_r = \frac{A_\infty \beta_r \prod_{k=1}^{\infty} (1 - \beta_r/\alpha_k) E(\beta_r, \beta_k, m)}{\prod_{k=1, k\neq r}^{\infty} (1 - \beta_r/\beta_k) E(\beta_r, \beta_k, m)}, \qquad r = 1, 2, \cdots. \tag{4.29}$$

Proof. The entire proposition was proved above, following (4.25), except for the expressions (4.29), which we now derive as a consequence of (4.27). For the assumed q, rewrite (4.27) explicitly as

$$\left(1 - \frac{h_r}{s - \beta_r} - \sum_{\substack{j=1 \\ j\neq r}}^{\infty} \frac{h_j}{s - \beta_j}\right) \mathcal{B}(s) = \mathcal{D}(s).$$

In other words, by (4.23), for $r = 1, 2, \cdots$,

$$\left(1 - \sum_{\substack{j=1 \\ j\neq r}}^{\infty} \frac{h_j}{s - \beta_j}\right) \mathcal{B}(s) + \frac{h_r}{\beta_r} \prod_{\substack{k=1 \\ k\neq r}}^{\infty} \left(1 - \frac{s}{\beta_k}\right) E(s, \beta_k, m) = \mathcal{D}(s).$$

We now set $s = \beta_r$ in the above expression. Using $\mathcal{B}(\beta_r) = 0$ and (4.28), we obtain the sired formulas (4.29). □

The following lemma will be needed.

LEMMA 4.5. *For* $\alpha_k \sim \beta_k$, *we have*

$$\text{(i)} \qquad \frac{\prod_{k=1, k\neq r}^{\infty} (1 - \beta_r/\alpha_k) E(\beta_r, \beta_k, m)}{\prod_{k=1, k\neq r}^{\infty} (1 - \beta_r/\beta_k) E(\beta_r, \beta_k, m)} \sim 1,$$

$$\text{(ii)} \qquad \frac{\prod_{k=1, k\neq r}^{\infty} (1 - \alpha_r/\beta_k) E(\alpha_r, \beta_k, m)}{\prod_{k=1, k\neq r}^{\infty} (1 - \alpha_r/\alpha_k) E(\alpha_r, \beta_k, m)} \sim 1.$$

Proof of Lemma 4.5. We have, for $r, k = 1, 2, \cdots$,

$$\frac{\beta_r}{\alpha_k} \sim \frac{\beta_r}{\beta_k} \quad \text{and also} \quad \frac{\alpha_r}{\beta_k} \sim \frac{\alpha_r}{\alpha_k}, \tag{4.30}$$

and the conclusion follows. □

COROLLARY 4.6. *Under the assumptions of Proposition* 4.4, *the following asymptotic estimate holds for the sequence* h_r *in* (4.29) *generated by the assumed vector*

q, *as* $r \to \infty$,

$$h_r \sim \alpha_r - \beta_r \sim q_r[\bar{w}_s]_r \to 0 \quad \text{as } r \to \infty. \tag{4.31}$$

Proof. The first $\sim$ on the left stems from (4.29) via Lemma 4.5(i). The second $\sim$ on the right then follows from (4.17b), as the term in { } in that equation is obviously ~ 1. □

Remark 4.3. The proof of Proposition 4.4 and Corollary 4.6 actually refines the initial estimate $\alpha_k \sim \beta_k$ of (4.26) by leading to conclusion (4.31), which is *stronger*. In fact, (4.31) yields $\alpha_k - \beta_k \to 0$ as $k \to \infty$ and hence, by virtue also of (4.21) we have, as $k \to \infty$,

$$\frac{\alpha_k}{\beta_k} \sim \frac{q_k[\bar{w}_s]_k}{k^{2/\nu}} + 1 \sim 1 \tag{4.32}$$

and the estimate at the left of (4.26) follows. Note that estimate (4.31) relates the assumed q and $\{\alpha_k\}_{k=1}^{\infty}$. Reference to Fig. 4.1 below will greatly help in following the rest of the proof.

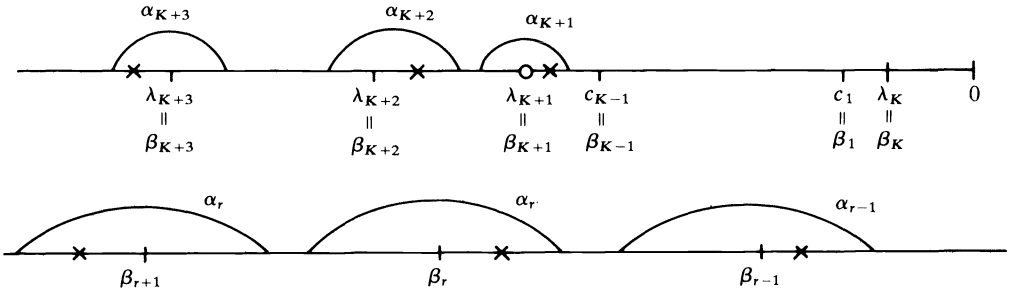


FIG. 4.1. *Asymptotic behavior of the constants* $\{\alpha_r\}$ *with respect to the constants* $\{\beta_r\}$ *(see* (4.31)).

We now tackle the problem of the existence of vectors q, as postulated in Proposition 4.4 and Corollary 4.6. To this end, it is convenient to introduce the following definition which is motivated by (4.17) and the paragraph below (4.17).

DEFINITION 4.1. An l_1-sequence $\{\tilde{h}_k\}_{k=1}^{\infty}$ will be said to *satisfy the realizability conditions* for the problem under study if it satisfies the conditions, for $r = 1, \cdots, K-1$,

$$\tilde{h}_r = p_r(w_u)_r \sum_{k=K}^{\infty} \tilde{h}_k \Big/ (\lambda_k - c_r)\left\{1 + \sum_{i=1}^{K-1} \frac{p_i(w_u)_i}{\lambda_k - c_i}\right\}, \tag{4.33}$$

which are crucial for the realizability of such $\{\tilde{h}_k\}_{k=1}^{\infty}$ through a vector q as demanded by (4.17).

Reversing the procedure of Proposition 4.4 and Corollary 4.6, we now first assign a sequence $\{\alpha_k\}_{k=1}^{\infty}$, with appropriate asymptotic behavior as suggested by (4.31), and then solve (4.27) for a suitable sequence $\{\tilde{h}_k\}_{k=1}^{\infty}$ (see Proposition 4.7). We then study in Theorem 4.8 how to force such a solution $\{\tilde{h}_k\}_{k=1}^{\infty}$ to satisfy the realizability conditions as well. Actually, even more is accomplished by the following two results:

PROPOSITION 4.7. *Fix an arbitrary vector* $v = \{v_k\}_{k=K}^{\infty} \in l_\infty$, *with* $v_k \neq 0$.

Next, assign a negative sequence $\{\alpha_k\}_{k=1}^{\infty}$, which satisfies

$$\alpha_k \neq \beta_j, \qquad k = j = 1, 2, \cdots, \tag{4.34}$$

and the asymptotic estimate

$$\alpha_k - \beta_k \sim [\bar{w}_s]_k v_k, \tag{4.35}$$

which implies (4.26) (*see Remark* 4.3 *and* (1.10′)).

Then, there exists a sequence $\{\tilde{h}_j\}_{j=1}^{\infty}$ *such that*

$$\left(1-\sum_{j=1}^{\infty}\frac{\tilde{h}_j}{s-\beta_j}\right)\mathscr{B}(s)=\mathscr{D}(s), \tag{4.36}$$

with $\mathscr{B}(s)$ *and* $\mathscr{D}(s)$ *as in* (4.23) *and* (4.28), *respectively. Identity* (4.36) *then implies*

$$\hat{\tilde{h}}(\alpha_k)\overset{\text{df}}{=}\sum_{j=1}^{\infty}\frac{\tilde{h}_j}{\alpha_k-\beta_j}\equiv 1, \qquad k=1,2,\cdots,$$

and, moreover, uniquely determines the $\{\tilde{h}_j\}_{j=1}^{\infty}$, *according to formula* (4.29) *written for* $\tilde{h}_j$.

Proof. See our detailed report [L8, Appendix 4B] or the analogous [L6, II, Proposition 4.7]. □

What follows is a main result that affirms the existence of admissible vectors q as postulated in Proposition 4.4 and Corollary 4.6.

THEOREM 4.8. *Let a vector* $q\in X_s$ *as in* (4.1) *be given; i.e.* $q=A_s^{1/4-\rho}QDg=\{q_k\}_{k=K}^{\infty}$ *for some* $g\in L_2(\Gamma)$, *where now all its coordinates*[21] $q_k=(-\lambda_k)^{-3/4-\rho}(g,\partial\Phi_k/\partial\eta|_\Gamma)_\Gamma$ *will be required*[22] *to be different from zero. Then, for all vectors* $w_s=Qw$, *with* $|w_s|_{l_\infty}$ *sufficiently small, which satisfy* (1.10), *one can construct*:

(i) *a vector* $\bar{v}=\{\bar{v}_k\}_{k=K}^{\infty}\in l_2$, *with*

$$0\neq|\bar{v}_k|\leqq C_v, \qquad k=K,K+1,\cdots, \tag{4.37}$$

where C_v *is sufficiently small so that the corresponding sequence* $\{\bar{\alpha}_k\}_{k=K}^{\infty}$ *defined by*

$$\bar{\alpha}_k-\beta_k=[\bar{w}_s]_k\bar{v}_k, \qquad k=K,K+1,\cdots, \tag{4.38}$$

has its terms $\bar{\alpha}_k$ *negative, distinct, and satisfying*

$$\bar{\alpha}_k\neq\beta_j, \qquad k=K,K+1,\cdots, \quad j=1,2,\cdots; \tag{4.39}$$

(ii) *a sequence* $\{\bar{\alpha}_i\}_{i=1}^{K-1}$ *of negative, distinct constants with*

$$\bar{\alpha}_i\neq\bar{\alpha}_k \quad \text{and} \quad \bar{\alpha}_i\neq\beta_j, \quad i=1,\cdots K-1, \quad k=K,K+1,\cdots, \quad j=1,2,\cdots, \tag{4.40}$$

such that the corresponding sequence $\{\bar{h}_r\}_{r=1}^{\infty}\in l_1$, *defined according to proposition* 4.7, *by*

$$\bar{h}_r=\frac{\beta_r A_\infty\prod_{k=1}^{\infty}(1-\beta_r/\bar{\alpha}_k)E(\beta_r,\beta_k,m)}{\prod_{k=1,k\neq r}^{\infty}(1-\beta_r/\beta_k)E(\beta_r,\beta_k,m)}, \qquad r=1,2,\cdots, \tag{4.41}$$

where the exponential function $E(\cdot,\cdot,m)$ *is defined in* (4.24) *and* A_∞ *is specified following* (4.28), *satisfies the realizability condition* (4.33);

(iii) *moreover, the function* $\hat{\bar{h}}(s)\overset{\text{df}}{=}\sum_{r=1}^{\infty}\bar{h}_r/(s-\beta_r)$, *corresponding to such a sequence* $\{\bar{h}_r\}_{r=1}^{\infty}$, *satisfies* $1-\hat{h}(\bar{\alpha}_k)=0$, $k=1,2,\cdots$, *with multiplicity one, while* $\hat{h}(s)\neq 1$ *for* $s\neq\bar{\alpha}_k$.

Therefore, according to Proposition 4.4, *the function* $\hat{h}(s)$ *defined above satisfies* (4.27).

Proof. Conclusions (i) and (ii) are proved in Appendices 4A and 4B. These appendices construct, ultimately by means of a fixed point technique, the sequence $\{\bar{\alpha}_i\}_{i=1}^{K-1}$ for which the realizability conditions hold, as well as the claimed vector $\bar{v}$ and the sequence $\{\bar{\alpha}_k\}_{k=K}^{\infty}$. Moreover, the continuity of the map $\{v_k\}_{k=K}^{\infty}\to\{\alpha_i\}_{i=1}^{K-1}: l_\infty\to\mathbb{R}^{K-1}$, needed in Appendix 4A, is proved in Appendix 4B.

[21] See Appendix 4A below (4A.4).

[22] This is possible since $\partial\Phi_k/\partial\eta|_\Gamma\neq 0$ for C^∞-boundaries Γ [S1, Cor. 2.2] as well as for parallelepipeds.

To show that such $\{\bar{h}_r\}_{r=1}^{\infty}$ is in l_1, we need only invoke part (i) of Lemma 4.5 to obtain the $\sim$ on

$$\bar{h}_r \sim \bar{\alpha}_r - \beta_r = [\bar{w}_s]_r \bar{v}_r, \tag{4.42}$$

while the equality on the right stems from (4.38). The Schwarz inequality applied to (4.42) ensures that $\{\bar{h}_r\}_{r=1}^{\infty} \in l_1$.

As to the claim for the corresponding function $\hat{\boldsymbol{h}}(s)$, this stems from (4.27), which holds by virtue of Proposition 4.7. □

The next result establishes that any admissible vector q that fulfils (4.27) also furnishes the desired solution to (4.10): $\mathscr{d}(t)$, of the special class IDE, with exponents α_r all different from the constants $\{\beta_r\}$.

THEOREM 4.9. *Let q be an admissible vector with corresponding function $\hat{\boldsymbol{h}}(s)$, obtained through* (4.17), *that satisfies* (4.27) *for negative constants $\{\alpha_k\}$ obeying the estimate* (4.31). (*Such vectors q are provided by the proof of Theorem* 4.8.)

Then, the following properties hold for the corresponding solution $\mathscr{d}(t)$ to the Volterra equation (4.10).

(i) *The solution $\mathscr{d}(t)$ has the form*

$$\mathscr{d}(t) = \sum_{r=1}^{\infty} d_r e^{\alpha_r t}, \qquad t \in \overline{\mathbb{R}^+}, \tag{4.43}$$

that is,

$$\hat{\mathscr{d}}(s) = \sum_{r=1}^{\infty} \frac{d_r}{s-\alpha_r}. \tag{4.44}$$

(*For* $\operatorname{Re} s > 0$, $\hat{\mathscr{d}}(s)$ *is the one-sided Laplace transform of $\mathscr{d}(t)$ and is extended over $\mathbb{C}$ through the right-hand side of* (4.44).)

(ii) *The coefficients $\{d_r\}$ satisfy the condition*

$$\sum_{r=1}^{\infty} |d_r \alpha_r| < \infty \quad \textit{and} \quad d_r \alpha_r \leqq \text{const}/r, \tag{4.45}$$

analogous to the property of $\{n_r \beta_r\}$ of Lemma 4.1(i).

(iii) *The coefficients d_r are the residues,* $\operatorname{res} \hat{\mathscr{d}}(\alpha_r)$, *of $\hat{\mathscr{d}}(s)$ at $\{\alpha_r\}$:*

$$\mathscr{d}_r = \operatorname{res} \hat{\mathscr{d}}(\alpha_r). \tag{4.46}$$

Consequently, $\mathscr{d}(t)$ is a function of the special class IDE.

Proof. To prove (4.44), we apply the Mittag–Leffler expansion to $\hat{\mathscr{d}}(s)$. To this end, the following proposition is crucial.

PROPOSITION 4.10. *Under the hypothesis of Theorem* 4.9, *the residues* $\operatorname{res} \hat{\mathscr{d}}(\alpha_r)$ *of the function $\hat{\mathscr{d}}(s)$ at the points α_r satisfy the estimate*

$$\operatorname{res} \hat{\mathscr{d}}(\alpha_r) = O\left(\left(\frac{1}{r^{1+2/\nu}}\right) + |n_r|\right) \quad \textit{as } r \to \infty.$$

Proof of Proposition 4.10. With $\hat{\mathscr{B}}(s)$ and $\hat{\mathscr{D}}(s)$ the entire functions given by (4.23) and (4.28), we have from (4.20) and (4.27)

$$\begin{aligned} \operatorname{res} \hat{\mathscr{d}}(\alpha_r) &= \lim_{s \to \alpha_r} (s-\alpha_r) \frac{\hat{n}(s)\hat{\mathscr{B}}(s)}{\hat{\mathscr{D}}(s)} \\ &= -\frac{\hat{n}(\alpha_r)\alpha_r(\beta_r - \alpha_r)\prod_{k=1,k\neq r}^{\infty}(1-\alpha_r/\beta_k)E(\alpha_r, \beta_k, m)}{A_\infty \beta_r \prod_{k=1,k\neq r}^{\infty}(1-\alpha_r/\alpha_k)E(\alpha_r, \beta_k, m)} \end{aligned} \tag{4.47}$$

$$\sim \hat{n}(\alpha_r)(\alpha_r - \beta_r), \tag{4.48}$$

by Lemma 4.5(ii), and $\alpha_r \sim \beta_r$, where

$$\hat{n}(\alpha_r) = \sum_{i=1}^{\infty} \frac{n_i}{\alpha_r - \beta_i}. \tag{4.49}$$

As $\{n_i\beta_i\}_{i=1}^{\infty}$ belongs to l_1, according to Lemma 4.1(i), we now need to invoke part (i) of the following lemma. Part (ii) will be needed later on in Lemma 4.12 and Theorem 4.14.

LEMMA 4.11. *For any vector* $b = \{b_j\}_{j=1}^{\infty}$ *such that* $b_j \leqq \text{const}/j$, *the following estimates hold for* $\alpha_i \sim \beta_i \overset{\text{def}}{=} \lambda_i$:

(i)

$$(1) \quad \sum_{i=r+1}^{\infty} \frac{b_i}{\lambda_i(\lambda_i - \alpha_r)} = O\left(\frac{1}{\alpha_r^{2-\varepsilon}}\right),$$

$$(2) \quad \sum_{i=1}^{r-1} \frac{b_i}{\lambda_i(\lambda_i - \alpha_r)} = O\left(\frac{1}{\alpha_r}\right),$$

where ε *is an arbitrary positive number. In addition, by symmetry with the above,*

(ii)

$$(1) \quad \sum_{i=k+1}^{\infty} \frac{b_i}{\alpha_i(\lambda_k - \alpha_i)} = O\left(\frac{1}{\lambda_k^{2-\varepsilon}}\right),$$

$$(2) \quad \sum_{i=1}^{k-1} \frac{b_i}{\alpha_i(\lambda_k - \alpha_i)} = O\left(\frac{1}{\lambda_k}\right).$$

Proof. The proof is relegated to our report [L8, App. 4E]. See also the analogous [L6, II, Lemma 4.11]. □

Continuing with the proof of Proposition 4.10, we apply Lemma 4.11(i) to the sum in (4.49) (in this last case after multiplying the numerator and denominator by λ_i as allowed by Lemma 4.1(i)), to obtain

$$\hat{n}(\alpha_r) = O\left(\frac{1}{\alpha_r}\right) + \frac{n_r}{\alpha_r - \beta_r}. \tag{4.50}$$

Inserting (4.50) into (4.48) yields

$$\operatorname{res} \hat{d}(\alpha_r) = O\left(\frac{\alpha_r - \beta_r}{\alpha_r}\right) + O(n_r). \tag{4.51}$$

The desired conclusion of Proposition 4.10 then follows from (4.51) via

$$\frac{\alpha_r - \beta_r}{\alpha_r} = O\left(\frac{1}{r} \cdot \frac{1}{r^{2/\nu}}\right), \tag{4.52}$$

which is the result of (4.37), (4.38) with $[\bar{w}_s]_r \leqq \text{const}/r$ (see (1.10′)) and of $\alpha_r \sim \beta_r \sim r^{2/\nu}$, from (4.21). □

Returning to the proof of Theorem 4.9, we see from Proposition 4.10 and Lemma 4.1(i) that we can apply the Mittag–Leffler theorem [L3, p. 394], [K3, p. 37ff.] to obtain

$$\hat{d}(s) = \sum_{r=1}^{\infty} \frac{\operatorname{res} \hat{d}(\alpha_r)}{s - \alpha_r} + e(s), \tag{4.53}$$

where $e(s)$ is an entire function. But from (4.20), we see that $\hat{d}(s)$ goes to zero for $s \to \infty$ in any way except along the negative real axis and this leads to $e(s) = O(1)$ and hence, by Liouville's theorem, to $e(s) \equiv 0$. We can then rewrite (4.53) as in (4.44),

from which relations (4.46) follow immediately. We now prove (4.45). To this end, we invoke (4.48) and (4.50) to obtain

$$d_r\alpha_r \sim \alpha_r - \beta_r + n_r\alpha_r \sim h_r + n_r\alpha_r \tag{4.54}$$

where the right-hand side estimate makes use of (4.42). Then (4.45) follows from Theorem 4.8(ii) and Lemma 4.1(i). The proof of Theorem 4.9 is thus complete. □

4.3. For $\mathscr{d}(t)$ as in § 4.2, the projections $x_u(t)$ and $x_s(t)$ are of the special class IDE. With the existence of representation (4.43) for some admissible vector q guaranteed by Theorems 4.8 and 4.9, the following Lemma will be useful in later arguments. The proof of this lemma, however, takes place in the t-domain.

LEMMA 4.12. *The solution $\mathscr{d}(t)$ of the integral equation* (4.10) *has the form* (4.43) *as a function of class* IDE, *where*

$$\alpha_r \neq \beta_j, \qquad r, j = 1, 2, \cdots,$$

if and only if the following conditions are satisfied:

$$\sum_{i=1}^{\infty} \frac{h_i}{c_r - \beta_i} \equiv 1, \qquad r = 1, 2, \cdots, \tag{4.55}$$

$$n_r = -h_r \sum_{i=1}^{\infty} \frac{d_i}{\alpha_i - \lambda_r}, \qquad r = K, K+1, \cdots, \tag{4.56a}$$

$$\frac{(x_{0u})_r}{p_r} = \sum_{i=1}^{\infty} \frac{d_i}{\alpha_i - c_r}, \qquad r = 1, \cdots, K-1. \tag{4.56b}$$

Proof. Let $\mathscr{d}(t)$ be given by (4.43). Then, inserting (4.43), (4.13) and (4.14) into (4.10) and equating to zero, by linear independence argument, all the coefficients of exponentials result, after straightforward computations, in (4.55), as well as in

$$n_r \equiv -h_r \sum_{i=1}^{\infty} \frac{d_i}{\alpha_i - \beta_r}, \qquad r = 1, 2, \cdots. \tag{4.57}$$

Relation (4.55) means, of course, that the $\{\alpha_r\}_{r=1}^{\infty}$ are zeros of the denominator $(1 - \hat{k}(s))$ of $\hat{\mathscr{d}}(s)$. For $r = K, K+1, \cdots$, (4.57) leads to (4.56a) via (4.15). For $r = 1, \cdots, K-1$, however, the ratios n_r/h_r in (4.57) are computed via (4.16a) and (4.17a), thus leading to (4.56b).

Reversing the steps of the above procedure proves the opposite direction. □

A final lemma is needed.

LEMMA 4.13. *For any admissible vector q provided by Theorems* 4.8 *and* 4.9, *the corresponding solution $\mathscr{d}(t)$ of the form* (4.43) *to equation* (4.10) *satisfies the following asymptotic estimates*:

$$\frac{d_r}{\lambda_r - \alpha_r} = O\left(\frac{[A_s^{1/4-\rho} x_{os}]_r}{q_r \lambda_r}\right) + O\left(\frac{1}{\lambda_r}\right).$$

Proof. By (4.56a), we can write for $r = K, K+1, \cdots$,

$$\frac{d_r}{\lambda_r - \alpha_r} = \frac{n_r}{h_r} - \sum_{\substack{j=1 \\ j \neq r}}^{\infty} \frac{d_j}{\lambda_r - \alpha_j}.$$

Next, by virtue of (4.45) and Lemma 4.11 part (ii), we have

$$\sum_{\substack{j=1 \\ j \neq r}}^{\infty} \frac{d_j}{\lambda_r - \alpha_r} = O\left(\frac{1}{\lambda_r}\right).$$

Finally, (4.16b) and (4.17b) provide

$$\frac{n_r}{h_r} = O\left(\frac{[A_s^{1/4-\rho}x_{os}]_r}{q_r\lambda_r}\right) + O\left(\frac{1}{\lambda_r}\right)$$

and the lemma is proved. □

We are finally ready now to draw the desired conclusions to the solution $x(t)$.

THEOREM 4.14. *For any admissible vector q provided by Theorems* 4.8 *and* 4.9, *the projection* $x_u(t)$ *of the solution* $x(t)$ *is an* X_u*-function of the special class* IDE.[23]

Proof. In view of (2.20), it is enough to show the desired conclusion for the integral term of (4.4). By (2.20), (4.8) and (4.43), this term can be rewritten for $t \geqq 0$ as

$$\begin{aligned}
\bar{x}_u(t) &\overset{\text{df}}{=} \int_0^t e^{\bar{A}_u(t-\tau)} p\, d(\tau)\, d\tau \\
&= \sum_{i=1}^{K-1} \left\{ \int_0^t e^{c_i(t-\tau)} p_i \sum_{r=1}^{\infty} d_r e^{\alpha_r \tau}\, d\tau \right\} \psi_i \\
&= \sum_{i=1}^{K-1} \left\{ p_i \sum_{r=1}^{\infty} d_r \frac{e^{\alpha_r t} - e^{c_i t}}{\alpha_r - c_i} \right\} \psi_i \qquad \text{(by (3.2))} \\
&= \sum_{i=1}^{K-1} \left\{ -p_i \left(\sum_{r=1}^{\infty} \frac{d_r}{\alpha_r - c_i} \right) e^{c_i t} + p_i \sum_{r=1}^{\infty} \frac{d_r}{\alpha_r - c_i} e^{\alpha_r t} \right\} \psi_i \\
&= -e^{\bar{A}_u t} x_o + \sum_{i=1}^{K-1} p_i \sum_{r=1}^{\infty} \frac{d_r}{\alpha_r - c_i} e^{\alpha_r t} \psi_i, \qquad t > 0 \qquad \text{(4.58')}
\end{aligned}$$

(by (4.56b) and (2.20)).

But, since the $\{c_i\}$ and the $\{\alpha_r\}$ are chosen, by (4.0), so that

$$\inf_{\substack{r=1,2,\cdots, \\ i=1,\cdots,K-1,}} |\alpha_r - c_i| = \gamma > 0$$

(see Fig. 4.1), relations (4.45) in Theorem 4.9 imply a fortiori that the infinite sum in r in (4.58′) is a function of the special class IDE. The desired conclusion is then contained in (4.58′). □

The proof for the relevant result for $x_s(t)$ on X_s passes through the following theorem.

THEOREM 4.15. *For any admissible vector q provided by Theorems* 4.8 *and* 4.9, *the function* $A_s^{1/4-\rho}x_s(t)$ *is of the special class* IDE *in the* X_s*-weak topology. More precisely, we have for* $y \in X_s = QL_2(\Omega)$

$$\langle (-A_s)^{1/4-\rho} x_s(t), y \rangle = \sum_{r=1}^{\infty} e^{\alpha_r t} d_r \left[1 + \sum_{i=1}^{K-1} \frac{(w_u)_i p_i}{\alpha_r - c_i} \right] \left[\sum_{k=K}^{\infty} \frac{\lambda_k q_k y_k}{\alpha_r - \lambda_k} \right]. \tag{4.59}$$

Proof. We have from (4.5), (4.4):

$$\begin{aligned}
\langle A_s^{1/4-\rho} x_s(t), y \rangle &= \langle e^{-A_s t} A_s^{1/4-\rho} x_{0s}, y \rangle \\
&+ \left\langle \int_0^t A_s e^{-A_s(t-\tau)} q \langle e^{\bar{A}_u \tau} x_{0u}, w_u \rangle\, d\tau, y \right\rangle \qquad \text{(4.60)} \\
&+ \left\langle \int_0^t A_s e^{-A_s(t-\tau)} q [\langle \bar{x}_u(\tau), w_u \rangle + d(\tau)]\, d\tau, y \right\rangle
\end{aligned}$$

[23] Since X_u is finite-dimensional and A_u an operator on it, then $A_u^r x_u(t)$ for any power r is also of the special class IDE.

with $\mathscr{A}(\cdot)$ and $\bar{x}_u(\cdot)$ defined by (4.8) and (4.58), respectively. Since by (4.58′) and (4.43),

$$\langle \bar{x}_u(\tau), w_u\rangle + \mathscr{A}(\tau) = -\langle e^{\bar{A}_u\tau}x_{0u}, w_u\rangle + \sum_{r=1}^{\infty} d_r\left\{1+\sum_{i=1}^{K-1}\frac{(w_u)_i p_i}{\alpha_r - c_i}\right\}e^{\alpha_r\tau}, \tag{4.61}$$

we see that the first term in (4.61), once inserted in (4.60), cancels the second term in (4.60). Hence, substituting (4.61) into (4.60) and using the convolution relations (3.2) yields

$$\begin{aligned}\langle A_s^{1/4-\rho}x_s(t), y\rangle &= \langle e^{-A_s t}A_s^{1/4-\rho}x_{0s}, y\rangle \\ &\quad + \sum_{k=K}^{\infty}\sum_{r=1}^{\infty}\lambda_k q_k y_k d_r\left\{1+\sum_{i=1}^{K-1}\frac{(w_u)_i p_i}{\alpha_r - c_i}\right\}\frac{e^{\alpha_r t}-e^{\lambda_k t}}{\alpha_r-\lambda_k} \\ &= \sum_{k=K}^{\infty} e^{\lambda_k t}\left\{\lambda_k q_k y_k\left(\sum_{r=1}^{\infty}\frac{d_r(1+\sum_{i=1}^{K-1}(w_u)_i p_i/\alpha_r - c_i))}{\lambda_k-\alpha_r}\right)+\lambda_k^{1/4-\rho}\{x_{0s}\}_k y_k\right\} \\ &\quad + \sum_{r=1}^{\infty}\left\{e^{\alpha_r t}d_r\left(1+\sum_{i=1}^{K-1}\frac{(w_u)_i p_i}{\alpha_r - c_i}\right)\left(\sum_{k=K}^{\infty}\frac{\lambda_k q_k y_k}{\alpha_r-\lambda_k}\right)\right\}.\end{aligned} \tag{4.62}$$

To ascertain that all the infinite sums in (4.62) are well defined, we observe preliminarily that

$$\sum_{r=1}^{\infty}\left|\frac{d_r}{\alpha_r-\lambda_k}\right| = \sum_{r=1}^{k-1}\left|\frac{d_r\alpha_r}{\alpha_r(\alpha_r-\lambda_k)}\right| + \sum_{r=k+1}^{\infty}\left|\frac{d_r\alpha_r}{\alpha_r(\alpha_r-\lambda_k)}\right| + \left|\frac{d_k}{\alpha_k-\lambda_k}\right| \tag{4.63a}$$

$$= O\left(\frac{1}{\lambda_k}\right) + O\left(\frac{[A_s^{1/4-\rho}x_{0s}]_k}{\lambda_k q_k}\right), \tag{4.63b}$$

which follows when Lemma 4.11(ii) is applied to the first two sums in (4.63a) (a valid procedure when (4.45) is invoked) and Lemma 4.13 is applied to the last term in (4.63a). Hence, (4.63b) gives

$$\begin{aligned}\sum_{k=K}^{\infty}|\lambda_k q_k y_k|\left(\sum_{r=1}^{\infty}\left|\frac{d_r}{\alpha_r-\lambda_k}\right|\right) &\leqq \text{const}\sum_{k=K}^{\infty}|q_k y_k| + |(A_s^{1/4-\rho}x_{0s})_k y_k| \\ &\leqq \text{const}\,(|q| + |A_s^{1/4-\rho}x_{0s}|)|y|.\end{aligned} \tag{4.64}$$

Therefore, the following interchange of order of summation is allowed

$$\sum_{k=K}^{\infty}|\lambda_k q_k y_k|\sum_{r=1}^{\infty}\left|\frac{d_r}{\alpha_r-\lambda_k}\right| = \sum_{r=1}^{\infty}|d_r|\sum_{k=K}^{\infty}\left|\frac{\lambda_k q_k y_k}{\alpha_r-\lambda_k}\right|, \tag{4.65}$$

showing by (4.64) that, as $|\alpha_r|\to\infty$, the following sequence in r is in l_1:

$$\left\{d_r\left(1+\sum_{i=1}^{K-1}\frac{(w_u)_i p_i}{\alpha_r-c_i}\right)\left(\sum_{k=K}^{\infty}\frac{\lambda_k q_k y_k}{\alpha_r-\lambda_k}\right)\right\}_{r=1}^{\infty}\in l_1. \tag{4.66}$$

Since the term $\sum_{i=1}^{K-1}$ goes to zero as $r\to\infty$, we conclude from (4.65) and (4.66) that (4.62) is well defined as a function of class IDE. To complete the proof of Theorem 4.15, it remains to show that the first sum $\sum_{k=K}^{\infty}$ in (4.62) is, in fact, identically zero. To this end, we recall the definition of $\mathscr{A}(t)$ (Eq. (4.8)) and its expansion (4.43) for q admissible as assumed. We then deduce that for $y=\bar{y}$ with

$$\bar{y} = A_s^{-1/4+\rho}w_s \in X_s$$

the first sum $\sum_{k=K}^{\infty}$ in (4.62) vanishes identically. This implies[24] that all its coefficients are identically zero,

$$\bar{y}_k\left[\lambda_k q_k\left(\sum_{r=1}^{\infty}\frac{d_r(1+\sum_{i=1}^{K-1}(w_u)_i p_i/(\alpha_r-c_i))}{\lambda_k-\alpha_r}\right)+\lambda_k^{1/4-\rho}[x_{0s}]\right]\equiv 0, \qquad k=K, K+1, \cdots.$$

We then divide by the nonzero coefficient $\bar{y}_k$ [see (1.10'): $[\bar{w}_s]_k \neq 0$] and obtain the desired conclusion. Theorem 4.15 is fully proved. □

To finish off the proof of Theorem 1.3, we need to tackle the *synthesis* problem of the vectors $p \in X_u$ and $q \in X_s$ by a suitable boundary vector $g \in L_2(\Gamma)$, as dictated by (4.1) and (4.2). This is done exactly as at the end of Theorem 1.2 in § 2 ((2.25) ff.) □

Proof of Corollary 1.4. The claimed expansion (1.13) is obtained by simply combining the expansions (2.20), (4.59), (3.1) and (4.58'). □

Proof of Corollary 1.5. We write more conveniently $\{\gamma_n\}_{n=1}^{\infty}$ for the sequences $\{\lambda_k\}_{k=K}^{\infty}$, $\{c_i\}_{i=1}^{K-1}$, and $\{\alpha_r\}_{r=1}^{\infty}$. As noted in Corollary 1.4, the expansion (1.13) holds in the weak topology of $L_2(\Omega)$, when $x_0 \in L_2(\Omega)$, in which case, we can write

$$\langle S_F(t)x_0, y\rangle = \sum_{n=1}^{\infty} u_n(x_0, y)e^{\gamma_n t}, \qquad t \geqq 0 \tag{4.67}$$

for the desired feedback semigroup $S_F(t)$ on $L_2(\Omega)$; where the $u_n(x_0, y)$'s are constants depending on x_0 and y, which form an l_1-sequence; moreover, u_n is a bounded linear functional on x_0 for y fixed, and similarly on y for x_0 fixed. Thus, $u_n(x_0, y) = \langle B_n x_0, y\rangle$ for bounded operators B_n on $L_2(\Omega)$. Application of the Laplace transform to (4.67) (term-by-term application is legal) yields for the resolvent of A_F:

$$\langle R(\mu, A_F)x_0, y\rangle = \sum_{n=1}^{\infty}\frac{\langle B_n x_0, y\rangle}{\mu-\gamma_n}, \qquad \mu \neq \{\gamma_n\}, \tag{4.68}$$

after extension by analytic continuation. The constants $\{\gamma_n\}$ are then simple poles of the resolvent and thus simple eigenvalues of A_F [T5, Thm. 5.8–A, p. 306]. Next compute around a small circle Γ_n, centered at a fixed γ_n and containing no other point of the sequence $\{\gamma_n\}$

$$\int_{\Gamma_n} R(\mu, A_F)x_0\, d\mu = \int_{\Gamma_n}\frac{B_n x_0}{\mu-\gamma_n}\, d\mu = 2\pi i B_n x_0$$

by the Cauchy theorem. Thus, B_n is the projection from $L_2(\Omega)$ onto the one-dimensional eigenspace of A_F spanned by the normalized eigenvector $e_{F,n}$, along $(I-B_n)L_2(\Omega)$: $B_n x = \eta_n(x)e_{F,n}$, $\eta_n(x) =$ scalar. Then, $\eta_n(e_{F,n}) = 1$ and $\eta_n(e_{F,m}) = 0$, $n \neq m$. From (4.67) with $t = 0$,

$$x = \sum_{n=1}^{\infty} B_n x = \sum_{n=1}^{\infty}\eta_n(x)e_{F,n}, \qquad x \in L_2(\Omega) \tag{4.69}$$

so that $\{e_{F,n}\}_{n=1}^{\infty}$ is a basis on $L_2(\Omega)$. Since B_n commutes with A_F, we also obtain

$$A_F x = \sum_{n=1}^{\infty} B_n A_F x = \sum_{n=1}^{\infty} A_F B_n x = \sum_{n=1}^{\infty}\gamma_n\eta_n(x)e_{F,n}, \qquad x \in \mathcal{D}(A_F) \tag{4.70}$$

as desired. Expansions (4.69)–(4.70) can be written out explicitly as in (1.14)–(1.15) respectively. □

[24] If $\sum_{k=K}^{\infty} z_k e^{\lambda_k t} \equiv 0$, $t \geqq 0$, with $\{z_k\} \in l_1$, then term-by-term Laplace transforming gives $z_k/(\lambda-\lambda_k) + \sum_{j=1, j\neq k}^{\infty} z_j/(\lambda-\lambda_j) \equiv 0$ for $\lambda \neq \{\lambda_k\}_{k=K}^{\infty}$ by analytic continuation. Integrating along a small circle centered in λ_k yields by Cauchy's theorem $z_k \equiv 0$ as desired.

Appendix 1A. A peculiar property of the feedback semigroup. We close this section by illustrating a property of considerable interest possessed by the feedback semigroup arising in boundary feedback parabolic equations, which will cast further light on the problems here under study. It will suffice to specialize to the canonical situation where $-A(\xi, \partial) = \Delta + c^2$(*).

CLAIM 1A.1. *For any choice of the stabilizing vectors* $\{w_j\}$ *and* $\{g_j\}$ *guaranteed by Theorem* 1.2, *the corresponding feedback semigroup* $S_F(t)$ *is not a contraction on* $L_2(\Omega)$, *and thus the constant* $M_{\theta,\varepsilon}$ *in* (1.8) *of Theorem* 1.2 (*with* $\theta = \frac{1}{4} - \rho$, *i.e. on* $L_2(\Omega)$) *cannot be less than or equal to one.*

We are thus exhibiting a not-so-common analytic semigroup that is not a contraction, and yet decays exponentially to zero in the uniform norm of $L_2(\Omega)$ as $t \to \infty$.

Proof. To substantiate our claim, we consider the simplest case with $K = 2$, where there is only one unstable eigenvalue λ_1: $\cdots < \lambda_2 < 0 < \lambda_1$, with λ_1 and λ_2, say, *simple* eigenvalues. If $\{\phi_i\}$ are the normalized eigenvectors of $-A$, we define vectors h and w as follows:

$$\text{(i)} \quad h = h_1\phi_1 + h_2\phi_2, \qquad \text{(ii)} \quad w = \sum_{i=1}^{\infty} w_i\phi_i,$$

where we impose:

$$\text{(iii)} \quad w_1 \text{ arbitrary}, \qquad \text{(iv)} \quad (h, w) = h_1w_1 + h_2w_2 = 0.$$

In the case with $J = 1$, we obviously have

$$\{x \in H^2(\Omega): x|_\Gamma = \langle x, w\rangle g\} \subset \mathcal{D}(A_F).$$

Thus, since $h|_\Gamma = 0$ by (i), we deduce from (iv) that: *for any* $w \in L_2(\Omega)$ *and for any* $g \in L_2(\Gamma)$, the vector h defined above satisfies: $h \in \mathcal{D}(A_F)$.

From (i), we then obtain $\langle A_Fh, h\rangle = \lambda_1 h_1^2 + \lambda_2 h_2^2$. Now, with λ_1 (positive) and λ_2 (negative) assigned, we choose h_1 and h_2 so that

$$\text{(v)} \quad \langle A_Fh, h\rangle \text{ is positive.}$$

Next, with arbitrary nonzero w_1 given, we determine w_2 from (iv). The other coordinates of w are irrelevant. Then, by (v), we obtain:

(vi) The feedback generator A_F corresponding to any vector $g \in L_2(\Gamma)$ and to any such $w \in L_2(\Omega)$ is *not* dissipative.

In particular, specialize w and g to be "stabilizing" vectors, as guaranteed by the major Theorem 1.2: in doing this, we make use of $w_1 \neq 0$ to satisfy conditions (1.7). The corresponding feedback semigroup is then *not a contraction*, by (vi) and the Lumer–Phillips theorem, yet decays exponentially to zero by (1.8), as claimed. □

Note that, as a consequence, any attempt to stabilize boundary control systems via the sufficient condition

$$\langle A_Fx, x\rangle \leqq -\mu^2\langle x, x\rangle, \qquad x \in \mathcal{D}(A_F),$$

is bound to fail. Therefore, a more sophisticated approach is needed.

Appendix 4A. Proof of Theorem 4.8(i)–(ii). (1) If $v = \{v_k\}_{k=K}^{\infty}$ is any vector satisfying

$$0 \neq |v_k| \leqq C_v, \qquad k = K, K+1, \cdots, \tag{4A.1}$$

we define a corresponding sequence $\{\alpha_k\}_{k=K}^{\infty}$ of scalars by setting

$$\alpha_k - \beta_k = [\bar{w}_s]_k v_k, \qquad k = K, K+1, \cdots. \tag{4A.2}$$

By (1.10′), and (4A.1),

$$[\bar{w}_s]_k v_k \leqq \text{const}/k \to 0 \quad \text{as } k \to \infty,$$

and so $\alpha_k \sim \beta_k$ and $(\alpha_k - \beta_k)/\beta_k \in l_1$. Thus, we conclude that: if C_v in (4A.1) is sufficiently small, then the constants α_k, $k = K, K+1, \cdots$, in (4A.2) are all *real* and *negative*, like the corresponding β_k's, as desired.

(2) If $\{c_i\}_{i=1}^{K-1}$ are the distinct negative constants obtained through Lemma 2.1 and are, e.g., required to be

$$\begin{array}{ccccccc} \lambda_{K+1} & < \lambda_K & < c_{K-1} & < \cdots < & c_1 & < 0, \\ \| & \| & \| & & \| & \\ \beta_{K+1} & \beta_K & \beta_{K-1} & & \beta_1 & \end{array}$$

we consider vectors $\{a_i\}_{i=1}^{K-1}$ in the $\mathbb{R}^{K-1}$ sphere $\mathcal{S}_c$

$$\mathcal{S}_c = \{\{a_i\}_{i=1}^{K-1} : |a_i - c_i| \leqq \rho_c\}, \qquad i = 1, \cdots, K-1, \tag{4A.3}$$

with ρ_c sufficiently small, so that all coordinates $a_1, \cdots, a_{K-1}$ are negative and distinct.

(3) We now let q be a vector of the form

$$q = A_s^{1/4-\rho} QDg \in QL_2(\Omega) \quad \text{for some } g \in L_2(\Gamma). \tag{4A.4}$$

Since

$$q = \sum_{r=K}^{\infty} (-\lambda_r)^{1/4-\rho} < Dg, \Phi_r > \Phi_r = \sum_{r=K}^{\infty} (-\lambda_r)^{1/4-\rho} (g, D^*\Phi_r)_\Gamma \Phi_r$$

and $D^*\Phi_r = (-1/\lambda_r)(\partial\Phi_r/\partial\eta)|_\Gamma$ [L2], [W1], we deduce that the coordinates q_r of q are

$$q = \{q_r\}_{r=K}^{\infty} = \left\{ (-\lambda_r)^{-3/4-\rho} \left(g, \frac{\partial \Phi_r}{\partial \eta}\bigg|_\Gamma \right)_\Gamma \right\}_{r=K}^{\infty}$$

and therefore, by [S1, Cor. 2.2] for C^∞-boundary Γ and for parallelepipeds, they can and will all be required to be different from zero: $q_r \neq 0$, $r = K, K+1, \cdots$.

(4) Next, with q as in (4A.4) *fixed*, and for each $\{a_i\}_{i=1}^{K-1}$ in the sphere $\mathcal{S}_c$ we define a nonlinear operator T_y, depending on the vector $y = \{q_r[\bar{w}_s]_r\}_{r=K}^{\infty}$ from $\{a_i\}_{i=1}^{K-1} \to \{q_r'\}_{r=K}^{\infty}$, by setting: $\{q_r'\}_{r=K}^{\infty} = T_y\{a_i\}_{i=1}^{K-1}$ with

$$q_r' \equiv q_r[\bar{w}_s]_r \frac{\prod_{k=1}^{K-1}(1-\beta_r/\beta_k)E(\beta_r, \beta_k, m)}{\prod_{k=1}^{K-1}(1-\beta_r/a_k)E(\beta_r, \beta_k, m)} \qquad (\beta_k \equiv c_k), \quad r = K, K+1, \cdots, \tag{4A.5}$$

which we shall consider as acting from the (closed) sphere $\mathcal{S}_c$ in R^{K-1} into l_∞.

For the operator T_y the following claim is easily verified:

CLAIM 4A.1. *With the radius ρ_c fixed in advance, and for a given vector q as in* (4A.4) *(so that such q is in l_2, hence in l_∞), one can select a sufficiently small sphere for the vectors $\bar{w}_s \in X_s$ in* (1.10′)*—as assumed in* (1.11)*—such that all the corresponding operators T_y which are defined through* (4A.5) *map the sphere $\mathcal{S}_c$ into an arbitrarily small neighborhood of the origin in l_∞.*

This assertion follows from the definition of T_y in (4A.5) and the fact that for points $\{a_i\}_{i=1}^{K-1}$ in $\mathcal{S}_c$ we have: $\inf_{r,i} \{|\beta_r - a_i| : \{a_i\} \in \mathcal{S}_c\} > 0$, where the inf is taken over all $r = K, K+1, \cdots$, and $i = 1, \cdots, K-1$.

(5) Next, motivated by (4.17b) and (4.29), we define a *nonlinear operator* $F: \{v_k\}_{k=K}^{\infty} \to \{(Fv)_r\}_{r=K}^{\infty} = \{f_r\}_{r=K}^{\infty}$ by

$$f_r = (Fv)_r \equiv \gamma_r \prod_{k=K}^{\infty} \left(1 - \frac{\beta_r}{\alpha_k}\right) E(\beta_r, \beta_k, m), \qquad r = K, K+1, \cdots, \tag{4A.6) (i)}$$

where the constants

$$\gamma_r = \frac{\beta_r A_\infty}{\left\{1+\sum_{i=1}^{K-1}\frac{p_i'(w_u)_i}{\beta_r-c_i}\right\}\prod_{k=K,k\neq r}^{\infty}\left(1-\frac{\beta_r}{\beta_k}\right)E(\beta_r,\beta_k,m)} \tag{4A.6)(ii}$$

depend only on data $\lambda_k = \beta_k$ and on parameters obtained through Lemma 2.1: $(w_u)_i$, p_i and c_i. By virtue of (4A.2), (4A.6)(i) gives $\{(Fv)_r\}_{r=K}^{\infty}$ explicitly in terms of $\{v_k\}_{k=K}^{\infty}$:

$$f_r = (Fv)_r = \gamma_r \prod_{k=K}^{\infty}\left(1-\frac{\beta_r}{\beta_k+[\bar{w}_s]_k v_k}\right)E(\beta_r,\beta_k,m), \quad r=K,K+1,\cdots. \tag{4A.6)(iii}$$

We shall consider F as *acting from a neighborhood of the origin in* l_∞ (see (4A.1)) *into* l_∞. Notice that F maps $\{v_k \equiv 0\}_{k=K}^{\infty}$ into $\{(Fv)_r \equiv 0\}_{r=K}^{\infty}$, more generally, if one coordinate $v_{\bar{k}} = 0$, then by (4A.6) (iii) the corresponding coordinate $(Fv)_{\bar{k}} = 0$ as well. Notice that, because of Lemma 4.5, it follows from (4A.6) that

$$(Fv)_r \sim \frac{\beta_r}{\alpha_r}(\alpha_r-\beta_r) \sim \alpha_r-\beta_r = v_r[\bar{w}_s]_r \quad \text{as } r\to\infty, \tag{4A.7}$$

where, in the last step, we have used (4A.2).

(6) The following proposition, to be proved at the end of the present appendix, will be paramount in our treatment.

PROPOSITION 4A.1. *The inverse mapping theorem applies to the operator F defined above; i.e. there is a neighborhood $\mathcal{N}_v$ of $v=0$ in l_∞ such that F is one-to-one in $\mathcal{N}_v$ with F^{-1} continuous in the $l_\infty \to l_\infty$ topology.*

An important consequence of both Claim 4A.1 and Proposition 4A.1 is:

CLAIM 4A.2. *With the radius ρ_c of $\mathcal{S}_c$ fixed in advance, and for a given vector q as in* (4A.4), *one can select a sufficiently small sphere for the vectors $\bar{w}_s \in X_s$ in* (1.10′)—*as assumed in* (1.11)—*such that the corresponding composite map $F^{-1}T_y$ of T_y followed by F^{-1} is well defined and maps the sphere $\mathcal{S}_c$ into an arbitrarily small neighborhood of the origin in l_∞.*

(7) It will be shown in Appendix 4B that the map $G:\{v_k\}_{k=K}^{\infty}\to\{\alpha_i\}_{i=1}^{K-1}$ from a neighborhood of the origin of l_∞ into R^{K-1}, which produces the constants $\alpha_1,\cdots,\alpha_{K-1}$ for which the realizability conditions (R.C.) (4.33) hold, is in fact continuous:

$$\begin{array}{ccc} \{a_i\}_{i=1}^{K-1}\in\mathcal{S}_c & & \left\{\{\alpha_i\}_{i=1}^{K-1} \text{ for which the R.C. hold}\right\}\in\mathcal{S}_c. \\ T_y\downarrow & & \uparrow G \\ l_\infty \ni \{q_r'\}_{r=K}^{\infty} & \xrightarrow{F^{-1}} & \{v_k\}_{k=K}^{\infty}\in l_\infty. \end{array}$$

Notice that, if we apply F^{-1} on the vector $\{q_r'\}$ given by (4A.5) for a preassigned q, we get a vector $\{v_k\}$, whose corresponding $\{\alpha_k\}_{k=K}^{\infty}$ via (4A.2) are such that

$$q_r' = \gamma_r \prod_{k=K}^{\infty}\left(1-\frac{\beta_r}{\alpha_k}\right)E(\beta_r,\beta_k,m), \qquad r=K,K+1,\cdots,$$

so that, by (4A.7),

$$q_r' \sim v_r[\bar{w}_s]_r, \tag{4A.7′}$$

while (4A.5) gives $q_r' \sim q_r[\bar{w}_s]_r$. Thus the preassigned vector q and the obtained vector $v = F^{-1}q'$, with $q' = T_y a = T_y\{a_i\}_{i=1}^{K-1}$, satisfy

$$q_r \sim v_r. \tag{4A.7″}$$

Thus, if q is only in l_2, so is v.

(8) As a consequence of the last two statements, we obtain a conclusive result, which we state formally:

PROPOSITION 4A.2. *With a radius ρ_c of $\mathscr{S}_c$ fixed in advance, and a given vector q as in* (4A.4), *one can select a suitably small sphere for the vectors $\bar{w}_s \in X_s$ in* (1.10′)—*as assumed in* (1.11)—*such that the corresponding composite map $GF^{-1}T_y$ of T_y followed by F^{-1} and by G is well defined and maps the (closed) sphere $\mathscr{S}_c$ into itself.*

Since $GF^{-1}T_y$ is continuous, being the composition of continuous maps, Brower's fixed-point theorem applies and produces (at least) a fixed point $\{\bar{a}_i\}_{i=1}^{K-1} \in \mathscr{S}_c$, with all coordinates distinct and negative. The corresponding vector $\{\bar{q}'_r\}_{r=K}^{\infty} = T_y\{\bar{a}_i\}_{i=1}^{K-1}$ has all its coordinates different from zero, by (1.10′) and also since all coordinates q_r were taken $\neq 0$; hence (by the observation above Prop. 4A.1), the corresponding vector $\{\bar{v}_k\}_{k=K}^{\infty} = F^{-1}T_y\{\bar{a}_i\}_{i=1}^{K-1}$ is $\sim \bar{q}_r$ and also has all its coordinates different from zero. As to the sequence $\{\bar{\alpha}_i\}_{i=1}^{K-1}$ in the conclusion (ii) of Theorem 3.8, we then take a fixed point $\bar{\alpha}_i = \bar{a}_i$, $i = 1, \cdots K-1$. As to the sequence $\{\bar{\alpha}_k\}_{k=K}^{\infty}$ in the conclusion (i) of Theorem 4.8, we take instead

$$\bar{\alpha}_k = \beta_k + [\bar{w}_s]_k \bar{v}_k, \qquad k = K, K+1, \cdots.$$

With this choice,

$$\bar{\alpha}_k \neq \beta_j, \quad k = K, K+1, \cdots, \quad j = 1, 2, \cdots \quad \text{and} \quad k = 1, \cdots K-1, \quad \text{but} \quad k \neq j,$$

from Appendix 4B. It remains to show that

$$\bar{\alpha}_i \neq c_i \equiv \beta_i, \qquad i = 1, \cdots, K-1.$$

In fact, if—say—$\bar{\alpha}_1 = c_1$, then by (4B.1) in Appendix 4B with $a_i = \bar{\alpha}_i$, $i = 1, \cdots K-1$, we would have that the corresponding $\tilde{h}_1(a) = 0$. Since $\{\bar{\alpha}_i\}_{i=1}^{K-1}$ makes the realizability conditions (4.33) hold, it follows that (4.17a–b) apply to the corresponding sequence $\{\tilde{h}_r(a) = h_r\}_{r=1}^{\infty}$ with p and $\{c_i\}_{i=1}^{K-1}$ coming from Lemma 2.1 and with q the vector as in (4A.4), for which Brower's fixed point theorem holds. But then from (4.17a–b), we see that the condition $h_r = 0$, $r = 1, \cdots, K-1$, can always be avoided by slightly changing, if necessary, say, just one c_i.

(9) To conclude the proof of Theorem 4.8 it remains to establish Proposition 4A.1.

Proof of Proposition 4A.1. We need to verify that the operator F defined in (4A.6) satisfies the following properties [L9, p. 226], [M3, p. 116]:

(a) F admits a well-defined Fréchet derivative $F'(v)$ in a neighborhood $\mathscr{N}_v$ of the origin in l_∞, and, moreover, the map $v \to F'(v)$ is continuous in $\mathscr{N}_v$ in the topology of $l_\infty \to l_\infty$;

(b) The operator $F'(v=0)$ is invertible; i.e. $[F'(v=0)]^{-1}$ exists in l_∞. The validity of (a) will follow a fortiori once we show the following:

ASSERTION. *The second Fréchet derivative $F''(v)$ is well defined as a continuous operator $l_\infty \to l_\infty$.*

In fact, $F''(v)$ is an infinite matrix with the following structure:

$$F''(v) = \begin{vmatrix} E_k \\ E_{K+1} \\ E_{K+2} \\ \vdots \end{vmatrix} \quad \text{where } E_r = \begin{vmatrix} \frac{\partial}{\partial v_1}\frac{\partial f_r}{\partial v_1} & \frac{\partial}{\partial v_2}\frac{\partial f_r}{\partial v_1} & \frac{\partial}{\partial v_3}\frac{\partial f_r}{\partial v_1} & \cdots \\ \frac{\partial}{\partial v_1}\frac{\partial f_r}{\partial v_2} & \frac{\partial}{\partial v_2}\frac{\partial f_r}{\partial v_2} & \frac{\partial}{\partial v_3}\frac{\partial f_r}{\partial v_2} & \cdots \\ \frac{\partial}{\partial v_1}\frac{\partial f_r}{\partial v_3} & \frac{\partial}{\partial v_2}\frac{\partial f_r}{\partial v_3} & \frac{\partial}{\partial v_3}\frac{\partial f_r}{\partial v_3} & \cdots \\ \cdot & \cdot & \cdot & \cdot \end{vmatrix}$$

for $r = K, K+1, \cdots$. According to well-known results [T5, p. 220], $F''(v)$ defines a continuous operator: $l_\infty \to l_\infty$, provided

$$\sup_{\substack{\text{over all}\\ \text{rows}}} \{l_1\text{-norm of a row}\} = \sup_r \left\{ \sup_l \sum_{j=1}^{\infty} \left| \frac{\partial^2 f_r}{\partial v_j \, \partial v_l} \right| \right\} < \infty. \tag{4A.8}$$

From (4A.6) we compute, after setting $Z_k = [\bar{w}_s]_k$, for $j \neq l$:

$$\begin{aligned}
\frac{\partial^2 f_r}{\partial v_j \, \partial v_l} &= \frac{\partial}{\partial v_j} \left[\gamma_r \prod_{\substack{k=K \\ k \neq l}}^{\infty} \left(1 - \frac{\beta_r}{\beta_k + Z_k v_k}\right) \frac{E(\beta_r, \beta_k, m)\beta_r Z_l E(\beta_r, \beta_l, m)}{(\beta_l + Z_l v_l)^2} \right] \\
&= \gamma_r \left\{ \prod_{\substack{k=K \\ k \neq l \\ k \neq j}}^{\infty} \left(1 - \frac{\beta_r}{\beta_k + Z_k v_k}\right) E(\beta_r, \beta_k, m) \right\} \frac{\beta_r Z_j E(\beta_r, \beta_j, m)\beta_r Z_l E(\beta_r, Z_l, m)}{(\beta_j + Z_j v_j)^2 (\beta_l + Z_l v_l)^2} \\
&= \frac{f_r \beta_r Z_j \beta_r Z_l}{\left(1 - \frac{\beta_r}{\beta_j + Z_j v_j}\right)\left(1 - \frac{\beta_r}{\beta_l + Z_l v_l}\right)(\beta_j + Z_j v_j)^2 (\beta_l + Z_l v_l)^2} \qquad \text{(4A.9)} \\
&= \frac{f_r \beta_r^2 Z_j Z_l}{(\beta_j - \beta_r + Z_j v_j)(\beta_l - \beta_r + Z_l v_l)(\beta_j + Z_j v_j)(\beta_l + Z_l v_l)}.
\end{aligned}$$

Similarly from (4A.6) we obtain for $j = l$:

$$\begin{aligned}
\frac{\partial^2 f_r}{\partial v_l^2} &= \gamma_r \left\{ \prod_{\substack{k=K \\ k \neq l}}^{\infty} \left(1 - \frac{\beta_r}{\beta_k + Z_k v_k}\right) E(\beta_r, \beta_k, m) \right\} \frac{-2\beta_r Z_l^2 E(\beta_r, \beta_l, m)}{(\beta_l + Z_l v_l)^3} \\
&= \frac{-2 f_r \beta_r Z_l^2}{(\beta_l - \beta_r + Z_l v_l)(\beta_l + Z_l v_l)^2}.
\end{aligned} \tag{4A.10}$$

To verify (4A.8), we need, according to (4A.9) and (4A.10), to check that the following two quantities Σ_1 and Σ_2 be finite:

$$\begin{aligned}
\Sigma_1 &= \sup_r \left\{ \sup_l \sum_{\substack{j=1 \\ j \neq l}}^{\infty} \left| \frac{\partial^2 f_r}{\partial v_j \, \partial v_l} \right| \right\} \\
&= \sup_r \left\{ |f_r \beta_r^2| \sup_l \frac{|Z_l|}{|\beta_l - \beta_r + Z_l v_l||\beta_l + Z_l v_l|} \left[\sum_{\substack{j=1 \\ j \neq l}}^{\infty} \frac{|Z_j|}{|\beta_j - \beta_r + Z_j v_j||\beta_j + Z_j v_j|} \right] \right\},
\end{aligned} \tag{4A.11}$$

$$\begin{aligned}
\Sigma_2 &= \sup_r \left\{ \sup_l \left| \frac{\partial^2 f_r}{\partial v_l^2} \right| \right\} \\
&= \sup_r \left\{ 2|f_r \beta_r| \sup_l \frac{Z_l^2}{|(\beta_l - \beta_r + Z_l v_l)(\beta_l + Z_l v_l^2)|} \right\}.
\end{aligned} \tag{4A.12}$$

We first handle Σ_1. In the sequel we shall use with no further mention that $Z_l = [\bar{w}_s]_k \leqq \text{const}/k \to 0$ as $k \to \infty$ (from (1.10′)). In order to show that Σ_1 is finite, it will suffice to establish that the following quantities Σ_1' and Σ_1'' be finite, where Σ_1' refers to the case

$l \neq r$ and Σ_1'' refers to the case $l = r$:

$$\begin{aligned}\Sigma_1' &= \sup_r \left\{ |f_r \beta_r^2| \sup_{l \neq r} \frac{|Z_l|}{|(\beta_l - \beta_r + Z_l v_l)\beta_l|} \left[\sum_{\substack{j=1 \\ j \neq l}}^{\infty} \frac{|Z_j|}{|(\beta_j - \beta_r + Z_j v_j)\beta_j|} \right] \right\} \\ &\leqq \sup_r \left\{ |f_r \beta_r^2| \sup_{l \neq r} \frac{|Z_l|}{|(\beta_l - \beta_r)\beta_l|} \left[\sum_{\substack{j=1 \\ j \neq l,r}}^{\infty} \frac{|Z_j|}{|(\beta_j - \beta_r)\beta_j|} + \frac{|Z_r|}{|Z_r v_r \beta_r|} \right] \right\}\end{aligned} \tag{4A.13}$$

$$\begin{aligned}\Sigma_1'' &= \sup_r \left\{ |f_r \beta_r^2| \frac{|Z_r|}{|Z_r v_r \beta_r|} \sum_{\substack{j=1 \\ j \neq r}}^{\infty} \frac{|Z_j|}{|(\beta_j - \beta_r + Z_j v_j)\beta_j|} \right\} \\ &\leqq \text{const} \sup_r \left\{ |\beta_r| \sum_{\substack{j=1 \\ j \neq r}}^{\infty} \frac{|Z_j|}{|(\beta_j - \beta_r)\beta_j|} \right\}, \quad \text{by (4A.7).}\end{aligned} \tag{4A.14}$$

To conclude that Σ_1' and Σ_1'', and hence Σ_1, are finite, we now need to invoke the following two estimates:

$$\sum_{\substack{j=1 \\ j \neq r}}^{\infty} \frac{|Z_j|}{|(\beta_j - \beta_r)\beta_j|} = O\left(\frac{1}{\beta_r}\right) \qquad \text{(see the independent Lemma 4.11),} \tag{4A.15}$$

$$\sup_{\substack{r,l \\ r \neq l}} \frac{|\beta_r Z_l|}{|\beta_l(\beta_l - \beta_r)|} \leqq \text{const} \qquad \text{(see the final Remark 4A.1).} \tag{4A.16}$$

We now use (4A.15) at the level of the infinite sum term in (4A.13). We then obtain from (4A.13), since $f_r \leqq \text{const}$ and $(f_r/v_r) \leqq \text{const}$ (cf. (4A.7′)):

$$\Sigma_1' \leqq \text{const} \sup_r \left\{ \sup_{r \neq l} \frac{|\beta_r Z_l|}{|(\beta_l - \beta_r)\beta_l|} \right\} < \infty,$$

from which the finiteness of Σ_1' follows via (4A.16). The finiteness of Σ_1'' follows directly from (4A.14) via (4A.15). The proof that $\Sigma_1 < \infty$ is complete. The proof that $\Sigma_2 < \infty$ is simpler. From (4B.12), we compute

$$\begin{aligned}\Sigma_2 &\leqq \text{const} \sup_r \left\{ |f_r \beta_r| \sup_{l \neq r} \frac{|Z_l^2|}{|(\beta_l - \beta_r)\beta_l^2|} \right\} \\ &\quad + \text{const} \sup_r \left\{ \frac{|f_r \beta_r Z_r^2|}{|Z_r v_r \beta_r^2|} \right\}.\end{aligned} \tag{4A.17}$$

Since $f_r \leqq \text{const}$, $(f_r/v_r) \leqq \text{const}$ and $(Z_l/\beta_l) \to 0$, we easily conclude from (4A.17) that $\Sigma_2 < \infty$ as desired. The proof that the second Fréchet derivative $F''(v)$ is a bounded operator $l_\infty \to l_\infty$ is thus complete.

To finish the proof of Proposition 4A.1, it remains to show statement (b) on the invertibility of $F'(v = 0)$. This is quickly done as follows. The Fréchet derivative $F'(v)$ is an infinite matrix with entries

$$F'(v) = \left| \left(\frac{\partial f_r}{\partial v_j} \right) \right|, \qquad j = r = K, K+1, \cdots.$$

Starting from (4A.5) (ii), we compute directly, again with $Z_k = [\bar{w}_s]_k$,

$$\frac{\partial f_r}{\partial v_j} = \gamma_r \prod_{\substack{k=K \\ k\neq j}}^{\infty} \left(1-\frac{\beta_r}{\alpha_k}\right) E(\beta_r, \beta_k, m) \frac{\partial}{\partial v_j}\left(1-\frac{\beta_r}{\beta_r + Z_j v_j}\right) E(\beta_r, \beta_j, m)$$

(4A.18)

$$= \frac{f_r}{\left(1-\dfrac{\beta_r}{\beta_j + Z_j v_j}\right)} \frac{-\beta_r Z_j}{(\beta_r + Z_j v_j)^2} = \frac{-f_r \beta_r Z_j}{(\beta_j - \beta_r + Z_j v_j)(\beta_r + Z_j v_j)}.$$

Setting $v = 0$, i.e. $v_j \equiv 0$, $j = K, K+1, \cdots$, implies, as we know (see below (4A.7)), $f_r \equiv 0$, $r = K, K+1, \cdots$. Thus from (4A.18) we get

$$\left.\frac{\partial f_r}{\partial v_j}\right|_{v=0} = 0 \quad \text{for } r \neq j, \qquad \left.\frac{\partial f_r}{\partial v_r}\right|_{v=0} = \frac{-f_r}{v_r}.$$

Thus, $F'(v=0)$ is an infinite matrix whose off-diagonal terms all vanish, and whose main diagonal terms are $-f_r/v_r$, $r = K, K+1, \cdots$. Since $f_r/v_r \sim 1$ by (4A.7), we can conclude that $F'(v=0)$ is invertible as an operator: $l_\infty \to l_\infty$.

Remark 4A.1. To prove estimate (4A.16), rewrite

$$\sup_{\substack{r,l \\ r\neq l}} \frac{|\beta_r Z_l|}{|\beta_l(\beta_l - \beta_r)|} = \sup_{\substack{r,l \\ r\neq l}} \frac{|Z_l|}{\left|\beta_l\left(\dfrac{\beta_l}{\beta_r}-1\right)\right|}, \tag{4A.19}$$

and since $Z_l/\beta_l \to 0$ as $l \to \infty$, we only have to worry if $[(\beta_l/\beta_r)-1]$ becomes unbounded. Thus, taking $r = l-1$ we estimate (4A.19) by using $Z_l = [\bar{w}_s]_l$ where $[\bar{w}_s]_l \leqq \text{const}/l$ (cf. (1.10′)) and $\beta_l \sim l^{2/\nu}$ (cf. (4.21)). It is left to the reader to check that the sup is bounded. Proposition 4A.1 is thus fully proved. □

Appendix 4B. Determination of $\{a_l\}_{l=1}^{K-1}$ from which the realizability conditions (4.33) hold; continuity of $G: \{v_k\}_{k=K}^{\infty} \to \{\alpha_l\}_{l=1}^{K-1}$ from $l_\infty \to \mathbb{R}^{K-1}$. Let $a = \{a_1, \cdots, a_{K-1}\}$ be a set of distinct, negative numbers in the sphere $\mathcal{S}_c$ defined by (4A.3), each different from all β_j. Motivated by (4.29), we define a sequence $\{\tilde{h}_r(a)\}_{r=1}^{\infty}$, depending on a, by

$$\tilde{h}_r(a) = \frac{\beta_r A_\infty \prod_{k=1}^{K-1}(1-\beta_r/a_k) E(\beta_r, \beta_k, m) \prod_{k=K}^{\infty}(1-\beta_r/\alpha_k) E(\beta_r, \beta_k, m)}{\prod_{k=1, k\neq r}^{\infty}(1-\beta_r/\beta_k) E(\beta_r, \beta_k, m)}. \tag{4B.0}$$

We then try to determine the parameters $\{a_1, \cdots, a_{K-1}\}$ in such a way that the sequence $\{\tilde{h}_r(a)\}_{r=1}^{\infty}$ satisfies the realizability condition (4.33).

The sequence $\tilde{h}_r(a)$ in (4B.0) can be more conveniently rewritten as

$$\tilde{h}_r(a) = e_r \prod_{k=1}^{K-1}\left(1-\frac{\beta_r}{a_k}\right), \qquad r = 1, 2, \cdots \tag{4B.1}$$

where the coefficients e_r are defined by

$$e_r = \frac{\beta_r A_\infty \prod_{k=K}^{\infty}(1-\beta_r/c_k) E(\beta_r, \beta_k, m) \prod_{k=1}^{K-1} E(\beta_r, \beta_k, m)}{\prod_{k=1, k\neq r}^{\infty}(1-\beta_r/\beta_k) E(\beta_r, \beta_k, m)}, \qquad r = 1, 2, \cdots, \tag{4B.2}$$

where $\alpha_k = \beta_k + [\bar{w}_s]_k v_k$ by (4.38) or (4A.2). We then determine the negative parameters $a_1, \cdots, a_{K-1}$ in such a way that the realizability conditions (4.33), rewritten now as

$$e_i \prod_{k=1}^{K-1}\left(1-\frac{\beta_i}{a_k}\right) = p_i(w_u)_i \sum_{r=K}^{\infty} \frac{e_r \prod_{k=1}^{K-1}(1-\beta_r/a_k)}{(\lambda_r - c_i)\{1-\sum_{j=1}^{K-1} p_j(w_u)_j/\lambda_r - c_i\}}, \tag{4B.3}$$

$$i = 1, \cdots, K-1,$$

are satisfied. To this end, we use the identity

$$\prod_{k=1}^{K-1}(\beta_r - a_k) = \beta_r^{K-1} - \Big(\sum_{k=1}^{K-1} a_k\Big)\beta_r^{K-2} + \Big(\sum_{\substack{i,j=1\\ i\neq j}}^{K-1} a_i a_j\Big)\beta_r^{K-3}$$
$$-\Big(\underbrace{\sum_{i,j,k=1}^{K-1} a_i a_j a_k}_{\text{distinct}}\Big)\beta_r^{K-4} + \cdots + (-1)^{K-1}\prod_{j=1}^{K-1} a_j,$$

on the right and, for $r = i$, on the left of (4B.3) and apply the sum $\sum_{r=K}^{\infty}$ on each power of β_r in the previous identity separately. Then, by setting

$$\bar{A}_{i,K-l} = p_i(w_u)_i \sum_{r=K}^{\infty} \frac{e_r \beta_r^{K-l}}{(\lambda_r - c_i)\{1 - \sum_{j=1}^{K-1} p_j(w_u)_j/\lambda_r - c_j\}}, \tag{4B.4}$$
$$l = 1, \cdots, K, \quad i = 1, \cdots, K-1$$

and

$$A_{i,K-l} = (-1)^{l-1}\bar{A}_{i,K-l} + (-1)^l e_i \beta_i^{K-l}, \qquad l = 1, \cdots, K, \quad i = 1, \cdots, K-1,$$

the realizability conditions (4B.3) can be rewritten as a multilinear algebraic system (see also (4.15)):

$$A_{i,K-1} + A_{i,K-2}\Big(\sum_{k=1}^{K-1} a_k\Big) + A_{i,K-3}\Big(\sum_{\substack{i,j=1\\ i\neq j}}^{K-1} a_i a_j\Big) + A_{i,K-4}\Big(\underbrace{\sum_{i,j,k=1}^{K-1} a_i a_j a_k}_{\text{distinct}}\Big) \tag{4B.5}$$
$$+ \cdots + \cdots A_{i,0}\Big(\prod_{j=1}^{K-1} a_j\Big) = 0,$$

of $(K-1)$ in $(K-1)$ unknowns, for which we seek a *negative* solution: $a_1, \cdots, a_{K-1}$ (that is, all a_i negative).

Notice that the infinite series defining each coefficient $\bar{A}_{j,K-l}$ through (4B.4) is $\sim \sum_{r=K}^{\infty} e_r \beta_r^{K-l-1}$. Therefore, the following claim is relevant.

CLAIM 4B.1. *Through* (4B.2), *define a nonlinear map*: $\{v_k\}_{k=K}^{\infty} \to \{e_r \beta_r^{K-l-1}\}_{r=K}^{\infty}$ *that we view from* $l_\infty \to l_1$. *Then this map is continuous.*

In fact, by simply comparing (4B.2) and (4A.6) (i) in Appendix 4A, with γ_r defined by (4A.6) (ii), we see that

$$e_r \sim \frac{(Fv)_r}{\prod_{k=1}^{K-1}(1 - \beta_r/\beta_k)}.$$

Proof. It suffices to consider the case $l = 1$. Here we get

$$\beta_r^{K-2} e_r \sim (Fv)_r/\beta_r \sim \frac{v_r[\bar{w}_s]_r}{\beta_r} = O\Big(\frac{v_r}{r^{1+2/\nu}}\Big) \quad \text{as } r \to \infty, \tag{4B.6}$$

by (1.10′) and also (4.21). Thus, the desired claim follows from (4B.6). We conclude that: *the map* $v = \{v_k\}_{k=K}^{\infty} \to A_{i,K-l}$ *is continuous from* $l_\infty \to R^1$.

We next want to show that when the l_∞-norm of v is sufficiently small, the system (4B.5) does admit a negative solution. To establish this, we make use of an observation plus a continuity argument.

The observation is that, when

$$\alpha_k \equiv \beta_k, \qquad k = K, K+1, \cdots, \tag{4B.7}$$

When a solution of distinct roots for the system (4B.5) is given by

$$a_i = \beta_i \overset{\text{df}}{=} c_i < 0, \qquad i = 1, \cdots, K-1, \tag{4B.8}$$

with the $c_i < 0$ coming from Lemma 2.1 (see (4.15)). In fact, under assumption (4B.7), it follows from (4B.2) that

$$e_r \equiv 0, \qquad r = K, K+1, \cdots,$$

and hence from the right-hand side of (4B.3) we deduce that system (4B.5) reduces to

$$e_i \prod_{k=1}^{K-1} \left(1 - \frac{\beta_i}{a_k}\right) = 0, \qquad i = 1, \cdots, K-1. \tag{4B.9}$$

In other words, $\bar{A}_{i,K-l} \equiv 0$ in this case (see (4B.5)).

Referring to (4.15), however, since

$$c_i \overset{\text{df}}{=} \beta_i \neq \alpha_k (\equiv \beta_k), \qquad i = 1, \cdots, K-1, \quad k = K, K+1, \cdots,$$

it follows from (4B.2) that $e_i \neq 0$, $i = 1, \cdots, K-1$. Hence, the desired conclusion (4B.8) is a consequence of (4B.9). This proves the observation.

For convenience of language, we shall call the situation under assumption (4B.7) the *original situation*. We now use a continuity argument. First, we argue that the roots of a multilinear system like (4B.6) depend continuously on the real coefficients of the system. Second, we argue that these coefficients, as shown above, depend continuously on the sequence $\{v_k\}_{k=K}^{\infty} \in l_\infty$. Therefore, if the vector v is sufficiently small in the l_∞-norm, the new coefficients $A_{i,K-l}$ are a slight perturbation of the original ones. Since the roots of the original situation (4B.9) are distinct and negative, as described in (4B.8), so will the new roots $\{\alpha_i\}_{i=1}^{K-1}$ be.

Thus the map G (needed in Appendix 4A), $\{v_k\}_{k=K}^{\infty} \to \{\alpha_i\}_{i=1}^{K-1}$ defined from $l_\infty \to \mathbb{R}^{K-1}$, is continuous. □

REFERENCES

[A1] P. J. ANTSAKLIS AND W. A. WOLOVICH, *Arbitrary pole placement using linear output feedback compensation*, Internat. J. Control, 25: 6 (1977), pp. 915–925.

[B1] A. V. BALAKRISHNAN, *Filtering and control problems for partial differential equations*, in Proc. 2nd Kingston Conf. on Differential Games and Control Theory, Univ. of Rhode Island, Marcel Dekker, New York, 1976.

[B2] ———, *Boundary control of parabolic equations: L-Q-R theory*, in Proc. Conf. on Theory of Nonlinear Equations, Sept. 1977, Akademie-Verlag, Berlin, 1978.

[B3] P. BUTZER AND H. BERENS, *Semigroups of Operators and Approximations*, Springer-Verlag, New York, 1967.

[C1] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics*, I, Interscience, New York, 1953.

[D1] E. J. DAVISON AND S. H. WANG, *On pole assignment in linear multivariable systems using output feedback*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 516–518.

[D2] G. DOETSCH, *Introduction to the Theory and Application of the Laplace Transform*, Springer-Verlag, New York, 1979.

[D3] N. DUNFORD AND J. SCHWARTZ, *Linear Operators*, Interscience Pubs., John Wiley, New York, vol. I, 1958, and vol. II, 1963.

[D4] H. B. DWIGHT, *Tables of Integrals and Other Mathematical Data*, Macmillan, New York, 1961.

[F1] A. FRIEDMAN, *Partial Differential Equations*, reprinted by Robert E. Krieger, Publ., Huntington, NY, 1976.

[F2] D. FUJIWARA, *Concrete characterization of the domains of fractional powers of some elliptic differential operators of the second order*, Proc. Japan Acad., 43 (1967), pp. 82–86.

[G1] W. GROBNER AND N. HOFREITER, *Integraltafel, Erster Teil, Unbestimmte Integrale*, 5th rev. ed., Springer-Verlag, Berlin, 1975.

[H1] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semigroups*, AMS Colloquium Publications, XXXI, American Mathematical Society, Providence, RI, 1957.

[K1] T. KATO, *Perturbation Theory of Linear Operators*, Springer-Verlag, New York, 1966.

[K2] H. KIMURA, *Pole assignment by gain output feedback*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 509–516.

[K3] K. KNOPP, *Theory of Functions*, Part 2, Dover, New York, 1947.

[K4] V. A. KONDRATIEV, *Boundary problems for elliptic equations in domain with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967).

[K5] T. KATO, *Fractional powers of dissipative operators*, J. Math. Soc. Japan, 14 (1961), pp. 241–248.

[L1] I. LASIECKA, *Unified theory for abstract parabolic boundary problems: A semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287–333.

[L2] I. LASIECKA AND R. TRIGGIANI, *A cosine operator approach in modeling $L_2[OT; L_2(\Gamma)]$ boundary input hyperbolic equations*, Appl. Math. Optim., 7 (1981), pp. 35–93.

[L3] N. LEVINSON AND R. M. REDHEFFER, *Complex Variables*, Holden-Day, San Francisco, 1970.

[L4] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, I, Springer (1972), II (1973).

[L5] ———, *Problème aux limites non homogènes*, IV, Ann. Scuola Superiore Pisa, 15 (1961), pp. 311–325.

[L6] I. LASIECKA AND R. TRIGGIANI, *Hyperbolic equations with Dirichlet boundary feedback via position vector: Regularity and almost periodic stabilization*; I, II, III, Appl. Math. Optim., 8 (1981), pp. 1–37; 8 (1982) pp. 103–130; 8 (1982), pp. 199–221.

[L7] ———, *Feedback semigroups and cosine operators for boundary feedback parabolic and hyperbolic equations*, J. Differential Equations, (to appear).

[L8] ———, *Analyticity and structural assignment of Dirichlet boundary feedback parabolic equations*, Report available upon request.

[M1] R. K. MILLER, *Nonlinear Volterra Integral Equations*, Benjamin, Reading, MA, 1971.

[N1] J. NECAS, *Les méthodes directes en théorie des équations elliptiques*, Masson et Cie., Paris, 1967.

[P1] A. PAZY, *Semigroups of Operators and Applications to Partial Differential Equations*, Lect. Notes, 10, Mathematics Dept., Univ. of Maryland, College Park, 1974.

[S1] E. J. P. G. SCHMIDT AND N. WECK, *On the boundary behavior of solutions to elliptic and parabolic equations*, this Journal, 16 (1978), pp. 533–598.

[T1] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, VEB Deutscher Verlag, Berlin, 1978.

[T2] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403; *Addendum*, 56 (1976), pp. 492–493.

[T3] ———, *Well posedness and regularity of boundary feedback parabolic systems*, J. Differential Equations, 36 (1980), pp. 347–362.

[T4] ———, *Boundary feedback stabilizability of parabolic equations*, Appl. Math. Optim., 6 (1980), pp. 201–220.

[T5] A. E. TAYLOR, *Introduction to Functional Analysis*, John Wiley, New York, 1958.

[W1] D. C. WASHBURN, *A bound on the boundary input map for parabolic equations with application to time optimal control*, this Journal, 17 (1979), pp. 652–671. (Based on Ph.D. thesis at UCLA, 1974.)

[W2] M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automat. Control, AC-12 (1967), pp. 660–665.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 5, September 1983

0363-0129/83/2105-0009 $01.25/0

NESTED BASES OF INVARIANTS FOR MINIMAL REALIZATIONS OF FINITE MATRIX SEQUENCES*

YUVAL BISTRITZ†

Abstract. The problem of finding minimal realizations of linear constant systems from finite order input-output Markov matrix sequences is considered. The paper identifies from the sequences sets of independent structural and numerical quantities which are invariants of equivalent state space representations and completely characterize any minimal realization of the sequence. These sets, termed bases of invariants, acquire a "nesting" property by which a subsequent basis of a higher order finite sequence is obtained from the previous basis by addition of some new invariants. Two canonical state space representations of special forms that reflect the input and output structural properties of the underlying systems are presented and readily derived from these bases by a simple algorithm which is provided. Necessary and sufficient conditions for the existence of a unique minimal partial realization to a given finite Markov sequence are given in terms of the invariants of its nested basis. The set of all minimal partial realizations $S^r_{n_r}$ that, in the case of existence of more than one solution, corresponds to many distinct systems, is thoroughly investigated. A minimal set of undetermined quantities that parametrize $S^r_{n_r}$ is obtained. These parameters are used to characterize $S^r_{n_r}$ either in the form of bases of invariants or in the form of the canonical representations, and it is also shown that an arbitrary assignment of values to these parameters leads to a minimal realization of the given finite sequence. Additional properties of these parameters that may be desirable in certain identification problems are also discussed.

Key words. minimal partial realization, system invariants, canonical forms

1. Introduction.

The problem of minimal realization of a finite sequence of Markov matrices of a multivariable linear constant system has been considered by various authors [1]–[8]. The early results of Kalman and Tether [1]–[3] showed that a minimal realization, or equivalently, a minimal extension sequence for a finite Markov sequence, always exists but may not be unique. Necessary and sufficient conditions on the incomplete Hankel matrix for the existence of a unique extension sequence as well as the derivation of a corresponding realization have also been described in these papers. The approach of Dickinson, Kailath and Morf in [4] is different in that they derive a matrix fraction representation by direct operation on the matrices of the sequence. References [5]–[8] also consider the incomplete Hankel matrix. Roman and Bullock [7] represent an invariant approach to the problem which is further developed by Candy, Warren and Bullock in [8] by deriving the partial realization from a set of Popov invariants [9].

The present paper provides a comprehensive treatment of the minimal partial realization (m.p.r.) problem of a finite sequence of r Markov matrices, using an invariant description. It puts a special emphasis on the common situation where a unique solution to the problem does not exist. It obtains a characterization of the set $S^r_{n_r}$ of all partial realizations of minimal dimension n_r for the sequence of r Markov matrices.

We show the existence of bases of invariants [10] for the description of equivalent classes of m.p.r.'s which have the property that a basis for a Markov sequence of a subsequent order is obtained from the basis of the former order by the addition of a few new invariants without altering the previous set of invariants. A basis acquiring this property is termed a nested basis. The nested bases are constructed from a set of entries of specified locations in the Markov sequence that were recently suggested

* Received on January 14, 1981 revised on April 20, 1982.
† School of Engineering, Tel-Aviv University, Tel-Aviv, Israel.

by Bosgra and Van der Weiden [11] and from a modified set of integer invariants that describe the structure of the underlying system. An important feature of the present approach is that the invariants that compose these bases are not dependent on the choice of some specific canonical representation. This differs from the system descriptions by canonical invariants suggested by Popov and refined and studied by Rissanen [10] and Denham [12]. Instead of the nonuniform descriptions of the set of all m.p.r.'s obtained by methods which adopt canonical invariants [7], [8], we obtain an intrinsic set of parameters $\mathscr{P}_r$ which completely characterizes $S^r_{n_r}$. By applying additional properties of the nested bases of invariants, it is also shown that $\mathscr{P}_r$ is a minimal set of independent parameters for a complete characterization of $S^r_{n_r}$ and that the mapping from the set of equivalent realizations in $S^r_{n_r}$ to $\mathscr{P}_r$ is one-to-one and onto.

Descriptions for m.p.r.'s other than the nested bases are also presented. In fact, any equivalent canonical representation can be derived from a nested basis. Two canonical state space representations of a special form that reflects the input and output invariant structure of the underlying system are presented and a simple algorithm for their derivation from a nested basis is provided. The two canonical forms tie together, in the special case of an infinite order Markov sequence, the realizations of Rissanen [10] and Silverman [14]. They also supply a simplified algorithm for the derivation of the invariants of Rissanen and provide a system invariant description for the realizations of Silverman.

It is desirable in general, to have a system description by a minimal set of parameters [9], [11], [5] and [7]. This is advantageous, for example, in solving the problem of system identification from statistical data which is possibly the most important practical implication of the present study. The stochastic interpretation of a deterministic partial realization is discussed by Akaike [15]. The problem of selection of free parameters for the description of all possible minimal realizations of a finite Markov sequence, which Ledwich and Fortman [6] recognized as a difficult one, is solved by the above-mentioned set $\mathscr{P}_r$. The set $\mathscr{P}_r$ is not only a minimal set of independent parameters but is composed of entries of specified locations of the input-output data, which become available in further measurements.

The paper is written in continuous-time formulation but all the results apply also to discrete-time systems with some obvious redefinition of concepts. Section 2 contains the necessary definitions for the representation of the results, including the definition of a nested basis of system invariants. Section 3 represents bases of system invariants and suggests the above two canonical representations. Section 4 deals with the partial realization problem. The background of §§ 2 and 3 is used to derive nested bases of invariants for the descriptions of m.p.r.'s. The existence of a unique m.p.r. can be tested by its invariants. In the case where there exists more than one solution, the set of all m.p.r.'s is described by nested bases of invariants. These bases are expressed in terms of the minimal set of independent parameters $\mathscr{P}_r$. The m.p.r.'s can also be presented in the canonical forms described in § 3. These results are illustrated by a demonstrative example taken from [2]. This example appears also in [5], [7], [8] and allows a convenient comparison with former results.

2. System invariants of equivalent realizations. Let $\Sigma_n(A, B, C)$ denote the set of all matrix triples A, B, C, $A \in R^{n\times n}$, $B \in R^{n\times m}$, $C \in R^{l\times n}$ with (A, B) controllable and (A, C) observable. The elements $(A, B, C) \in \Sigma_n$ are state space representations of a linear system and each defines a transfer function matrix

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

The transfer function matrix can be expressed in a Laurent series about infinity

$$G(s) = G_1 s^{-1} + G_2 s^{-2} + \cdots, \tag{2.2}$$

where $G_i \in R^{l \times m}$ are called the Markov matrices of the system represented by (A, B, C) and are related to the representation by

$$G_i = CA^{i-1}B, \qquad i = 1, 2, \cdots. \tag{2.3}$$

Let E_n denote the equivalence of state space coordinate transformations, defined on elements of Σ_n, (A, B, C), $(\tilde{A}, \tilde{B}, \tilde{C}) \in \Sigma_n$, by

$$(A, B, C) E_n (\tilde{A}, \tilde{B}, \tilde{C}) \leftrightarrow CA^{i-1}B = \tilde{C}\tilde{A}^{i-1}\tilde{B}, \qquad i = 1, 2, \cdots. \tag{2.4}$$

The relation E_n partitions Σ_n into equivalence classes

$$E_n(\tilde{A}, \tilde{B}, \tilde{C}) = \{(A, B, C) \,|\, (A, B, C) \in \Sigma_n, (A, B, C) E_n (\tilde{A}, \tilde{B}, \tilde{C})\}. \tag{2.5}$$

The set of all such equivalence classes is called the quotient set and is denoted by Σ_n / E_n. Given an infinite sequence of Markov matrices G_i, $i = 1, 2, \cdots$, a representation $(A, B, C) \in \Sigma_n$ is called a minimal realization if (2.3) is satisfied. Given a finite sequence of only r Markov matrices $\{G_1, G_2, \cdots, G_r\}$ the representation (A, B, C) is called an rth order partial realization if

$$CA^{i-1}B = G_i, \qquad i = 1, 2, \cdots, r \tag{2.6}$$

and it is called a minimal partial realization (m.p.r.) of r if n is the minimal dimension of a system which satisfies (2.6).

An rth order m.p.r. is said to be unique if there exists only one infinite extension sequence G_{r+i}, $i = 1, 2, \cdots$ such that a m.p.r. is also a (complete) minimal realization of the infinite sequence $\{G_1, G_2, \cdots, G_r, G_{r+1}, G_{r+2}, \cdots\}$. If it is not unique, other triples of matrices exist that also minimally realize the rth order sequence but determine different extension sequences.

Let S_n^r be the set of all representations of m.p.r.'s of $\{G_1, G_2, \cdots, G_r\}$

$$S_n^r = \{(A, B, C)\, CA^{i-1}B = G_i, i = 1, \cdots, r, E_n(A, B, C) \subset \Sigma_n\}. \tag{2.7}$$

The m.p.r. of $\{G_1, G_2, \cdots, G_r\}$ is called unique if S_n^r consists of a single equivalence class. If it is not unique, the equivalence relation E_n partitions S_n^r into distinct classes that represent different systems whose first r Markov matrices are $\{G_1, G_2, \cdots, G_r\}$. The set of all these classes is denoted by S_n^r / E_n and is a subset of Σ_n / E_n. The set of all m.p.r.'s S_n^r is discussed in § 4, the main section of this paper. The characterization and derivation of S_n^r uses system invariant descriptions and canonical representations. The required concepts are defined below and elaborated in § 3. Many of the following definitions can be found elsewhere [16], [10]–[12].

DEFINITION 2.1. An *invariant* of the equivalence relation E_n is a function $f : \Sigma_n \to R$ for which $(A, B, C) E_n (\tilde{A}, \tilde{B}, \tilde{C})$ implies $f(A, B, C) = f(\tilde{A}, \tilde{B}, \tilde{C})$.

DEFINITION 2.2. An invariant $f : \Sigma_n \to R$ is a *complete invariant of* E_n if $f(A, B, C) = f(\tilde{A}, \tilde{B}, \tilde{C})$ implies $(A, B, C) E_n (\tilde{A}, \tilde{B}, \tilde{C})$.

A set of invariants $f, \cdots, f_N$ is called complete if Definition 2.2 is satisfied for $F = (f_1, \cdots, f_N) : \Sigma_n \to R^N$.

DEFINITION 2.3. The set of invariants $f_i : \Sigma_n \to R$ $i = 1, \cdots, N$ is said to be *independent* if the complement of the range of $F = (f_1, \cdots, f_N)$ in its codomain is a finite union of sets V_i

$$V_i = \{x \,|\, x \in R^N, P_{ij}(x) = 0, j = 1, \cdots, L; \text{finite } L\}, \tag{2.8}$$

where P_{ij} are polynomials.

Definition 2.3 implies that F is surjective on its codomain except possibly on a subset of "measure zero" and that no f_i can be expressed as a function of any f_j, $j \neq i$. This definition is a refinement due to [10] of the definition in [9] for independence of invariants. A complete set of invariants for the equivalence relation E_n of (2.4) may be divided into two sets $F = (F_\sigma; F_\alpha)$ where F_σ is a set of integers called arithmetic invariants that correspond to the structure of $E_n(A, B, C)$ and F_α is a complementary set of numerical values called algebraic invariants for the description of $E_n(A, B, C)$ [11], [10]. We adopt the term basis of invariants [10] for the following intuitive notion of a complete set of independent invariants.

DEFINITION 2.4. A set of invariants F is called a *basis of invariants* for the equivalence class $E_n(A, B, C)$ if $F = (F_\sigma; F_\alpha)$ is complete, the set of arithmetic invariants $F = (\sigma_1, \cdots, \sigma_{N_1})$ is surjective and the complementary set of algebraic invariants $F_\alpha = (\alpha_1, \cdots, \alpha_{N_2})$ is independent.

A subset $\Sigma_c \subset \Sigma_n$ is called a canonical form if for each $(A, B, C) \in \Sigma_n$ there exists one and only one $(A_c, B_c, C_c) \in \Sigma_c$ for which $(A, B, C) E_n (A_c, B_c, C_c)$. A canonical representation induces a (trivial) complete set of invariants for $E_n(A, B, C)$ simply by $F_c(A, B, C) = (A_c, B_c, C_c)$ with the arithmetic and algebraic invariants being the location and content, respectively, of the entries of the matrices A_c, B_c, C_c. It is well understood that such a set is not in general a basis because the entries in the canonical representation satisfy certain constraints (e.g., minimal dimensionality) by which they are dependent. However subsets of independent invariants can be extracted from (A_c, B_c, C_c) and a complete set of independent invariants determined [7], [8], [10]. We call a basis of invariants whose algebraic invariants are a subset of entries of a canonical representation a *canonical basis*. Two canonical bases are presented in § 3 of this paper where a different type of basis is introduced. The new basis of invariants does not depend on any specific canonical form and it is shown later to acquire the additional "nesting" property which is defined below. Nested bases of invariants play a major role in our forthcoming investigation of the set of all minimal partial realizations. Let $G_i \in R^{l \times m}$ $i = 1, 2, \cdots$ be a sequence of matrices and let $S^r_{n_r}$ be the set of all rth order partial realizations of minimal dimension n_r (2.7). Let $F^r = (F^r_\sigma; F^r_\alpha)$ be a basis of invariants for E_n (A, B, C) where $(A, B, C) \in S^r_{n_r}$.

DEFINITION 2.5. The basis $F^r = (F^r_\sigma; F^r_\alpha)$ of E_n (A, B, C) is said to be a *nested basis of invariants* if for $j < r$ there exist subsets $F^j_\sigma \subset F^r_\sigma$ and $F^j_\alpha \subset F^r_\alpha$ such that $F^j = (F^j_\sigma; F^j_\alpha)$ is a basis of invariants for some equivalence class in $S^j_{n_j}$ $(n_j \leqq n_r)$, the set of all m.p.r.'s of the jth order sequence of the same Markov matrices $(j = r-1, r-2, \cdots \geqq m, l)$.

Nested bases of invariants can be considered as a natural elaboration on concepts of the previous system invariants for the descriptions of partial realizations. Subsets of nested bases of invariants form bases of invariants for lower order m.p.r.'s in a manner reminiscent of that by which subspaces of projections of linear spaces are spanned by subsets of their bases. Thus, the nested bases add, to the previous notion of independence and completeness of the bases of invariants, an additional notion of familiarity with bases in linear algebra. These bases provide a useful tool for the investigation of the partial realization problem and also have important consequences for efficient sequential realization algorithms of partial realizations of successive orders.

3. Bases of invariants and canonical forms for minimal realizations. Consider the infinite sequence of Markov matrices G_i, $i = 1, 2, \cdots$, and define the infinite Hankel block matrix H whose (i, j) block is G_{i+j-1}. Denote by $H_{i,j}$ the finite submatrix of the first i block rows and j block columns of H. It is well known that if the infinite Markov

sequence has a minimal realization $(A, B, C) \in \Sigma_n$ than this realization is completely determined by a submatrix $H_{i,j}$ not larger than $H_{n,n}$, where n is the rank of H [1], [14]. The matrix $H_{n,n}$ satisfies

$$(3.1)\quad H_{n,n} = \begin{bmatrix} G_1 & G_2 & \cdots & G_n \\ G_2 & & & \\ \vdots & & & \\ G_n & & \cdots & G_{2n-1} \end{bmatrix} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = H_C H_B,$$

where H_C and H_B are the observability and controllability matrices for $(A, B, C) \in \Sigma_n$. The row and column dependencies of $H_{n,n}$ are equivalent to row dependencies of H_C and the column dependencies of H_B, respectively.

Let $I_n = \{i_1, \cdots, i_n\}$ and $J_n = \{j_1, \cdots, j_n\}$ denote the indices of the first independent rows and columns of H or $H_{n,n}$. A selection of rows I_n and columns J_n is called a nice selection [12] if they satisfy

$$(3.2)\qquad l < i_k \in I_n \rightarrow i_k - l \in I_n, \qquad m < j_k \in J_n \rightarrow j_k - m \in J_n.$$

The choice of the first n independent rows and columns is recognized as a nice selection by the decomposition of $H_{n,n}$ in (3.1) into H_C and H_B. The sets of integers I_n and J_n thus defined on H are invariants of the equivalence relation E_n. They are closely related to the observability and controllability indices

$$(3.3)\qquad \nu = \{\nu_1, \cdots, \nu_l\} \quad \text{and} \quad \mu = \{\mu_1, \cdots, \mu_m\}$$

of the underlying system. The observability index $\nu_i \in \nu$ is the highest integer k for which the row $c_i^t A^{k-1}$ (c_i^t is the ith row of C) still appears in the selection of rows I_n in H_C. Similarly, the controllability index $\mu_j \in \mu$ is the highest integer k for which column $A^{k-1}b_j$ (b_j is the jth column of B) is in the selection of columns J_n in H_B. It is therefore obvious from the decomposition of $H_{n,n}$ in (3.1) that ν and μ are related to I_n and J_n by

$$(3.4)\qquad \begin{aligned} &I_n \leftrightarrow \nu, \quad \nu_i = \#I_n/i, \ i \in \mathbf{l}, \\ &J_n \leftrightarrow \mu, \quad \mu_j = \#J_n/j, \ j \in \mathbf{m}, \end{aligned}$$

where $\#S$ denotes the number of elements in the set S, $\mathbf{n} = \{1, 2, \cdots, n\}$ and I_n/i and J_n/j denote the subsets of the arithmetic series $\{i, i+l, i+2l, \cdots\}$ and $\{j, j+m, j+2m, \cdots\}$ that are included in the sets I_n and J_n, respectively. The relation between the sets I_n and J_n and the sets ν and μ is bijective and an alternative way to derive them is to use the crate diagram [11], [17], [18]. Assume for example that $J_6 = \{1, 2, 3, 4, 5, 7\}$ and $m = 2$ then $J_6/1 = \{1, 3, 5, 7\} \rightarrow \mu_1 = 4$ and $J_6/2 = \{2, 4\} \rightarrow \mu_2 = 2$ and therefore $\mu = \{4, 2\}$. The integers β and α defined on ν and μ by

$$(3.5)\qquad \beta = \max_{i \in \mathbf{l}} \nu_i, \qquad \alpha = \max_{j \in \mathbf{m}} \mu_j$$

are the first integers for which the realizability condition, $n = \rho H_{\beta,\alpha} = \rho H_{\beta+1,\alpha} = \rho H_{\beta,\alpha+1}$, is satisfied [14]. The submatrices of H in the following definition are uniquely determined by I_n and J_n and can be recognized as the submatrices defined also by Silverman for his realization algorithm [14].

DEFINITION 3.1. The following submatrices of the Hankel matrix H are defined for the sets I_n and J_n:

Q: The nonsingular $n \times n$ matrix formed from $H_{\beta,\alpha}$ by the intersection of the columns J_n and the rows I_n.

$\hat{A}$: The $n \times n$ matrix whose entries in $H_{\beta,\alpha+1}$ are positioned m columns to the right of the positions of corresponding entries of Q.

$\hat{B}$: The $n \times m$ matrix formed from $H_{\beta,\alpha}$ by the intersection of the columns $\mathbf{m}$ with the rows I_n.

$\hat{C}$: The $l \times n$ matrix formed from $H_{\beta,\alpha}$ by the intersection of the rows $\mathbf{l}$ with the columns J_n.

Remark 3.1. $\hat{A}$ is equivalently formed by the $n \times n$ matrix whose entries in $H_{\beta+1,\alpha}$ are positioned l rows below the position of corresponding entries of Q.

Remark 3.2. The columns J_n of $[B, A]$ and the rows I_n of $\begin{bmatrix} \hat{C} \\ \hat{A} \end{bmatrix}$, each separately, form Q.

Remark 3.3. The matrix triple $(\hat{A}Q^{-1}, \hat{B}, \hat{C}Q^{-1})$ is a realization of the infinite sequence G_i, $i = 1, 2, \cdots$ [14].

The first two remarks result from the special structure of the Hankel matrix. The triple of matrices $(\hat{A}, \hat{B}, \hat{C})$ involves the following collection of $n(m+l)$ entries of the infinite Markov sequence [11]

$$\mathcal{G} = \{g_{ijk} \mid k = 1, 2, \cdots, \nu_i + \mu_j,\ i \in \mathbf{l},\ j \in \mathbf{m}\}, \tag{3.6}$$

where $g_{ijk} = (G_k)_{ij}$. It follows from [11] and the bijective relation between I_n, J_n and ν and μ that I_n, J_n and $\mathcal{G}$ define a complete set of independent invariants in the sense of Definitions 2.3 and 2.4.

THEOREM 3.1. *$\mathcal{B} = (I_n, J_n; \mathcal{G})$ is a basis of invariants for $E_n(A, B, C)$, the equivalence class of minimal realizations of the infinite Markov sequence G_i, $i = 1, 2, \cdots$, with I_n, J_n the sets of arithmetic invariants and $\mathcal{G}$ the associated set of algebraic invariants.*

The basis $\mathcal{B}$ deserves the name of *Markov basis* to indicate that its set of algebraic invariants are entries of the Markov matrices. This is in contrast to the canonical invariants and bases of [9], [10], in which the algebraic invariants form entries of the canonical representations. It must be noted that other bases whose algebraic invariants are Markov entries may be defined in association with nice selections other than the choice I_n, J_n of first independent rows and columns [11]. The advantage of the above basis $\mathcal{B}$ over these other bases for the partial realization problem will be clarified in the next section.

Example 3.1. We shall illustrate the Markov basis for the following Markov sequence

$$G_1, G_2, G_3, G_4, \cdots = \begin{bmatrix} ① & ① \\ ① & ① \end{bmatrix}, \begin{bmatrix} ① & ② \\ ① & ④ \end{bmatrix}, \begin{bmatrix} ③ & ⑤ \\ 7 & ⑨ \end{bmatrix}, \begin{bmatrix} 7 & ⑥ \\ 11 & 4 \end{bmatrix}, \cdots. \tag{3.7}$$

The Hankel matrix is then

$$H = \begin{bmatrix} 1 & 1 & 1 & 2 & 3 & 5 & \cdots \\ 1 & 1 & 1 & 4 & 7 & 9 & \cdots \\ 1 & 2 & 3 & 5 & 7 & 6 & \cdots \\ 1 & 4 & 7 & 9 & 11 & 4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}.$$

The rank of H is $n = 3$ and the first independent rows and columns are $I_3 = \{1, 2, 3\}$ and $J_3 = \{1, 2, 4\}$. A systematic elimination procedure to determine these values will be described later. Thus the observability and controllability indices are $\nu = \{2, 1\}$ and

$\mu = \{1, 2\}$ and the Hankel submatrices of Definition 3.1 are

$$Q = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 4 \\ 1 & 2 & 5 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} 1 & 2 & 5 \\ 1 & 4 & 9 \\ 3 & 5 & 6 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \quad \hat{C} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 4 \end{bmatrix}.$$

The set of algebraic invariants $\mathcal{G}$ of (3.6) consists of the encircled entries in (3.7). In the rest of this section we shall present two canonical forms and bases of invariants of a special structure and derive them from the triple of matrices $(\hat{A}, \hat{B}, \hat{C})$ associated with $\mathcal{B}$.

THEOREM 3.2. *Given the Markov basis $\mathcal{B} = (I_n, J_n; \mathcal{G})$ for the equivalence class $E_n(A, B, C)$ of minimal realization of G_i, $i = 1, 2, \cdots$, two possible canonical realizations and two corresponding canonical bases of invariants for $E_n(A, B, C)$ are the following:*

1a) *The realization $(A_1, B_1, C_1) \in E_n(A, B, C)$ where*

$$A_1 = Q^{-1}\hat{A}, \quad B_1 = Q^{-1}\hat{B}, \quad C_1 = \hat{C}. \tag{3.8}$$

1b) *The columns J_n are the first independent columns of the controllable pair $[B_1, A_1]$ and they form the $n \times n$ identity matrix. The entries in the remaining m columns of $[B_1, A_1]$, denoted by S_1, form part of the corresponding canonical basis $\mathcal{B}_1$ defined below.*

1c) *The canonical basis of invariants for (A_1, B_1, C_1) is $\mathcal{B}_1 = (J_n; \mathcal{G}_1)$ where*

$$\mathcal{B}_1 = \text{Se}\,\{C_1\} \cup S_1 \tag{3.9}$$

Se $\{C_1\}$ *denotes the set of entries in C_1 and S_1 is defined in statement* (1b).

2a) *The realization $(A_2, B_2, C_2) \in E_n(A, B, C)$ where*

$$A_2 = \hat{A}Q^{-1}, \quad B_2 = \hat{B}, \quad C_2 = \hat{C}Q^{-1}. \tag{3.10}$$

2b) *The rows I_n are the first independent rows of the observable pair $\begin{bmatrix} C_2 \\ A_2 \end{bmatrix}$ and they form the $n \times n$ identity matrix. The set of entries in the remaining rows of $\begin{bmatrix} C_2 \\ A_2 \end{bmatrix}$, denoted by S_2, form part of the corresponding canonical basis $\mathcal{B}_2$ defined below.*

2c) *The canonical basis of invariants for (A_2, B_2, C_2) is $\mathcal{B}_2 = (I_n; \mathcal{G}_2)$ where*

$$\mathcal{G}_2 = \text{Se}\,\{B_2\} \cup S_2. \tag{3.11}$$

Se $\{B_2\}$ *denotes the set of entries of B_2, and S_2 is defined in statement* (2b).

Proof. See Appendix 1.

The two canonical forms and their corresponding canonical bases of invariants may be derived without explicit calculations involving the matrix Q. To achieve this purpose we define the following restricted elimination procedure.

DEFINITION 3.2. A *row (column) reserving elimination operation* represented by the matrix T_1^r of size $p \times p$ (T_2^r of size $q \times q$), is defined as a restricted Gaussian elimination procedure that acts only on the rows (columns) of some matrix M of rank n and size $p \times q$ $p, q \geqq n$. The action of T_1^r (T_2^r) is to change the first n independent rows (columns) of $T_1^r M$ (MT_2^r) to unity column (row) vectors without interchanging row (column) positions.

Note that T_1^r brings the first n columns of $T_1^r M$ to n unity vectors that may form a nonordered arbitrary selection of n columns of the $p \times p$ identity matrix. The procedure that changes the first n independent columns of M to n ordered unity vectors will be called a *complete row elimination* and is denoted by T_1. T_1 combines the action of T_1^r followed by a proper row interchange procedure. Similarly for the

dual case we shall denote by T_2 the complete column elimination procedure that changes the first n independent rows of M to the ordered sequence of unity row vectors of the $q \times q$ identity matrix.

Note that Q^{-1} in (3.8) and (3.10) stands for the operations of T_1 and T_2, respectively, so that the canonical forms can be derived from $(\hat{A}, \hat{B}, \hat{C})$ by a complete elimination procedure without finding Q^{-1} explicitly. Definition 3.2 suggests the following even more simple algorithm.

ALGORITHM 3.1

1. To obtain (A_1, B_1, C_1)

(i) $C_1 = \hat{C}$;

(ii) $T_1^r[\hat{B}, \hat{A}] = [\tilde{B}_1, \tilde{A}_1]$ where $\tilde{B}_1 \in R^{n \times m}$, $\tilde{A}_1 \in R^{n \times n}$ are intermediate matrices resulting from the implicit action of the row reserving operation T_1^r of Definition 3.2;

(iii) $[B_1, A_1] = P^t[\tilde{B}_1, \tilde{A}_1]$ where P is a permutation of the $n \times n$ identity matrix formed by columns J_n of $[\tilde{B}_1, \tilde{A}_1]$. Columns J_n are identified at stage (ii) as the pivotal columns of the action of T_1^r.

2. To obtain (A_2, B_2, C_2)

(i) $B_2 = \hat{B}$

(ii)
$$\begin{bmatrix} \hat{C} \\ \hat{A} \end{bmatrix} T_2^r = \begin{bmatrix} \tilde{C}_2 \\ \tilde{A}_2 \end{bmatrix},$$

where T_2^r is the column reserving elimination of Definition 3.2 and $\tilde{C}_2 \in R^{l \times n}$, $\tilde{A}_2 \in R^{n \times n}$ are the intermediate results of its action;

(iii)
$$\begin{bmatrix} C_2 \\ A_2 \end{bmatrix} = \begin{bmatrix} \tilde{C}_2 \\ \tilde{A}_2 \end{bmatrix} P^t,$$

where the permutation P is the matrix formed by rows I_n of $\left[\begin{smallmatrix} \tilde{C}_2 \\ \tilde{A}_2 \end{smallmatrix}\right]$ which are the pivotal rows revealed at stage (ii).

Any canonical representation can be derived from its Markov basis of invariants. The two canonical forms of Theorem 3.2 have been chosen as suitable forms for system invariant descriptions in having a structure that reflects the output or the input structural properties of the system and as forms that are easily derived from the basis. The derivation of the canonical bases of invariants $\mathscr{B}_1$ and $\mathscr{B}_2$ shows the connection between the Markov sets of invariants and previous descriptions of canonical invariants. The significance of canonical forms that reflect some of the invariant properties has been recognized in [10] and more recently in [11]. In fact the second canonical form and canonical basis of Theorem 3.2, derived here from the Markov basis, are identical to the results of Rissanen which are derived in [10] directly from the Hankel matrix. The first canonical form also coincides with a realization obtained by the Silverman procedure [14]. Silverman has not considered the invariant structure of the pair $[B_1, A_1]$ or any related invariant aspect of his realization. The main reason for stating Theorem 3.2 and subsequent algorithms is to provide for the next section alternative equivalent descriptions for the solution of the minimal partial realization problem other than the nested bases. However, these results are also significant for the previous invariant descriptions and derivation of complete minimal realizations. They show that the realization obtained by Rissanen [10] is a dual form of the earlier realization derived by Silverman [14]. These results also supply a simplified elimination procedure for the derivation of the invariants of Rissanen and provide a system invariant description framework for the realization of Silverman.

Example 3.2. We illustrate Theorem 3.2 and the subsequent algorithm by continuation of Example 3.1. To derive from $\hat{A}, \hat{B}, \hat{C}$, the first canonical form we follow Algorithm 3.1 to obtain:

(i) $$C_1 = \hat{C} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 4 \end{bmatrix}.$$

(ii) Perform reserving row elimination on $[\hat{B}, \hat{A}]$

$$[\hat{B}, \hat{A}] = \left[\begin{array}{cc|ccc} \boxed{1} & 1 & 1 & 2 & 5 \\ 1 & 1 & 1 & 4 & 9 \\ 1 & 1 & 3 & 5 & 6 \end{array}\right] \to \left[\begin{array}{cc|ccc} 1 & 1 & 1 & 2 & 5 \\ 0 & 0 & 0 & \boxed{2} & 4 \\ 0 & 1 & 2 & 3 & 1 \end{array}\right] \to \left[\begin{array}{cc|ccc} 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & \boxed{1} & 2 & 0 & -5 \end{array}\right]$$

$$\to \left[\begin{array}{cc|ccc} 1 & 0 & -1 & 0 & 6 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 2 & 0 & -5 \end{array}\right] = [\tilde{B}_1, \tilde{A}_1],$$

where the squared entries indicate the pivotal element at each step.

(iii) Rearranging the rows of $[\tilde{B}_1, \tilde{A}_1]$ to obtain the identity matrix at columns $J_3 = \{1, 2, 4\}$ or equivalently extracting P from these columns and performing the row changes by premultiplication by P^t results in

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \to P^t[\tilde{B}_1, \tilde{A}_1] = \left[\begin{array}{cc|ccc} 1 & 0 & -1 & 0 & 6 \\ 0 & 1 & 2 & 0 & -5 \\ 0 & 0 & 0 & 1 & 2 \end{array}\right] = [B_1, A_1],$$

note that columns $J_3 = \{1, 2, 4\}$ of $[B_1, A_1]$ form the identity matrix. The first canonical form is therefore

$$A_1 = \begin{bmatrix} -1 & 0 & 6 \\ 2 & 0 & -5 \\ 0 & 1 & 2 \end{bmatrix}, \qquad B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad C_1 = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 1 & 4 \end{bmatrix}.$$

The basis is $\mathscr{B}_1 = (J_3; \mathscr{G}_1)$ where $\mathscr{G}_1$ is formed by the set of entries of C_1 and of columns 1 and 3 of A_1.

Using the second part of Algorithm 3.1, the dual canonical form (A_2, B_2, C_2) is

$$A_2 = \begin{bmatrix} 0 & 0 & 1 \\ -1 & -1 & 3 \\ -2 & -3 & 2 \end{bmatrix}, \qquad B_2 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The corresponding canonical basis is $\mathscr{B}_2 = (I_3; \mathscr{G}_2)$ where $\mathscr{G}_2$ is formed by the entries of B_2 and of rows 2 and 3 of A_2.

4. Nested bases of invariants and minimal partial realizations. Given a finite sequence of r Markov matrices $\{G_1, \cdots, G_r\}$. We construct the Hankel matrix

(4.1) $$H^r = \begin{bmatrix} G_1 & G_2 & \cdots & G_r & G_{r+1}^* & \cdots \\ G_2 & & & & & \\ \vdots & & & & & \\ G_r & & & & & \\ G_{r+1}^* & & & & & \end{bmatrix}$$

where $\{G^*_{r+1}, G^*_{r+2}, \cdots\}$ represents some unknown complementary sequence. This matrix is closely related to the incomplete Hankel matrix used by Tether [2] with the slight difference that we explicitly write entries $(G^*_k)_{ij} = g^*_{ijk}$ for $k > r$ instead of the common asterisks put, in [2] and [7], [8], in all the locations of the unknown data. This modification proves to be powerful if the following "asterisk convention" is adopted: (i) Asterisked entries g^*_{ijk} and their combinations are carried along in any submatrix of H^r and any operation on such submatrices. (ii) Asterisked entries of matrices are assumed not to influence the internal dependencies between the rows and columns that are determined by the numerically specified entries. Consequently the indices of the first independent rows and columns and the rank of the matrix are not changed by any specific choice of values for the asterisked entries. The rank of a matrix that contains asterisked entries is by this convention the minimal rank that is admissible by its numerically specified parts.

Following the above convention, let n_r be the rank of H^r and denote the indices of the first n_r independent rows and columns of H^r by $I^r_{n_r}$ and $J^r_{n_r}$, respectively. Thus, $I^r_{n_r}$ ($J^r_{n_r}$) are the first n_r rows (columns) in H^r which, considering for each row (column) only columns (rows) that correspond to its numerically specified positions, do not depend linearly on preceding rows (columns).

Let β_r denote the smallest integer for which every row of the block row $\beta_r + 1$ of H^r (i.e., $[G_{\beta r+1}, G_{\beta r+2}, \cdots]$) depends on the previous rows and similarly let α_r denote the smallest integer for which every column of the block column $\alpha_r + 1$ depends on the preceding columns. We have the following important result on the existence of m.p.r.'s [2], [3].

THEOREM 4.1. *Given the finite sequence* $\{G_1, \cdots, G_r\}$: (1) *There exists an extension sequence* $\{G_{r+1}, G_{r+2}, \cdots\}$ *for which* n_r *is the dimension of the minimal realization of the infinite sequence* G_i, $i = 1, 2 \cdots$. *This realization is not, in general unique.* (2) *Every extension fixed up to* $r_0 = \alpha_r + \beta_r$ *is uniquely determined thereafter.*

The invariants description approach developed and discussed in this section will provide an alternative verification of this well known theorem. The theorem indicates that values for the g^*_{ijk} exist for which the structure of the Hankel matrix as well as the row and column dependencies are retained. Later we shall be able to specify the required g^*_{ijk} values and construct the minimal extension sequences. Let G_i, $i = 1, 2, \cdots$ be the infinite Markov sequence associated with some equivalence class in Σ_n and let $\mathcal{B} = (I_n, J_n; \mathcal{G})$ be its Markov basis described in Theorem 3.1. The next theorem establishes $\mathcal{B}$ as a nested basis of invariants (Definition 2.5).

THEOREM 4.2. *Let* $\{G_1, G_2, \cdots, G_r\}$ *be an rth order subsequence of the infinite sequence* G_i, $i = 1, 2, \cdots$ *whose Markov basis is* $\mathcal{B} = (I_n, J_n; \mathcal{G})$. *Let also* $n_r = \rho H^r$ *where* H^r *is the incomplete Hankel matrix associated with the finite subsequence. There exist subsets of* $I^r_{n_r} \subset I_n$, $J^r_{n_r} \subset J_n$ *and a subset* $\mathcal{G}_r$ *of* $n_r(m+l)$ *elements of* $\mathcal{G}_r \subset \mathcal{G}$ *such that* $\mathcal{B}_r = (I^r_n, J^r_n; \mathcal{G}_r)$ *forms a Markov basis for a m.p.r. of* $\{G_1, \cdots, G_r\}$ *of dimension* n_r.

Proof. Let $I^r_{n_r}$ and $J^r_{n_r}$ be the indices of the first independent rows and columns of H^r of (4.1). Let $\hat{A}_r \in R^{n_r \times n_r}$, $B_r \in R^{n_r \times m}$ and $C_r \in R^{l \times n_r}$ be the submatrices of H of (3.1) derived in association with I_n and J_n in accordance with Definition 3.1. Note that the matrices $\hat{A}_r$, $\hat{B}_r$, $\hat{C}_r$ are derived from H of (3.1), not from H^r of (4.1), and thus all their entries are specified and completely determined by $I^r_{n_r}$, $J^r_{n_r}$ and the infinite sequence G_i, $i = 1, 2, \cdots$. Clearly $n_r \leqq n$ and as the process of successive replacement of asterisked entries in H^r by numerically specified entries G_{r+i}, $i = 1, 2, \cdots$ may add new independent rows and columns but cannot cancel former independencies, we have $I^r_{n_r} \subset I_n$ and $J^r_{n_r} \subset J_n$. It therefore follows that the following algebraic set of

invariants defined for $I^r_{n_r}$ and $J^r_{n_r}$

$$\mathcal{G}_r = \{g_{ijk} \mid k = 1, \cdots, \bar{\nu}_i + \bar{\mu}_j,\ i \in \mathbf{l},\ j \in \mathbf{m}\}, \tag{4.2}$$

where

$$\bar{\nu}_i = \#I^r_{n_r}/i, \qquad \bar{\mu}_j = \#J^r_{n_r}/j \tag{4.3}$$

is a subset of $\mathcal{G}$. Construct from $(\hat{A}_r, \hat{B}_r, \hat{C}_r)$ a representation $(\bar{A}, \bar{B}, \bar{C}) \in \Sigma_{n_r}$ (using Algorithm 3.1 say) clearly $\mathcal{B}_r = (I^r_{n_r}, J^r_{n_r}; \mathcal{G}_r)$ is a Markov basis for $E_{n_r}(\bar{A}, \bar{B}, \bar{C})$. The Markov entries $\bar{G}_i = \bar{C}_i \bar{A}^{(i-1)} \bar{B}$ satisfy $\bar{G}_i = G_i$ for (at least) $i = 1, \cdots, r$ therefore $\mathcal{B}_r$ is a Markov basis for a m.p.r. of $\{G_1, \cdots, G_r\}$ where we have also shown that $I^r_{n_r} \subset I^r_n$, $J^r_{n_r} \subset J_n$ and $\mathcal{G}_r \subset \mathcal{G}$. □

It follows from Theorem 4.2 and Definition 2.5 that all the bases $\mathcal{B}_r$ are nested bases. The set of invariants $\mathcal{G}_r$ is either completely composed of entries that are selected from $\{G_1, \cdots, G_r\}$, in which case $\mathcal{B}_r$ represents a basis for the unique m.p.r. of the rth order sequence, or it contains also entries from $\{G_{r+1}, G_{r+2}, \cdots\}$. In the latter case $\mathcal{B}_r$ represents a basis of an equivalence class in $S^r_{n_r}$, the one which is induced by the higher order basis $\mathcal{B}$. In this case it is understood that other infinite Markov sequences of minimal dimensions n^*, $n^* \geqq n_r$ that have $\{G_1, \cdots, G_r\}$ for their first r matrices may induce other sub-bases for equivalence classes in $S^r_{n_r}$.

The last observation leads to the following condition for the uniqueness of a m.p.r.

PROPOSITION 4.3. *The sequence $\{G_1, \cdots, G_r\}$ yields a unique m.p.r. if and only if it acquires a Markov basis $\mathcal{B}_r = (I^r_{n_r}, J^r_{n_r}; \mathcal{G}_r)$ for which the set $\mathcal{G}_r$, defined in* (4.2), *is completely formed by entries of the sequence $\{G_1, \cdots, G_r\}$.*

Define for $\bar{\nu}_i$ and $\bar{\mu}_j$ of (4.3)

$$\beta_r = \max_{i \in \mathbf{l}} \bar{\nu}_i, \quad \alpha_r = \max_{j \in \mathbf{m}} \bar{\mu}_j, \quad \nu_0 = \alpha_r + \beta_r. \tag{4.4}$$

It follows from (4.1) that the condition expressed in Proposition 4.3 is satisfied if and only if $r_0 = \alpha_r + \beta_r \leqq r$. It is easy to verify that α_r and β_r of (4.4) are identical to the integers in Theorem 4.1. This proposition therefore assures the uniqueness conditions stated in Theorem 4.1.

We now proceed to investigate the case where $\{G_1, \cdots, G_r\}$ has more than one m.p.r. Assume that $r < r_0$ and thus that the set $S^r_{n_r}$ of all m.p.r.'s of $\{G_1, \cdots, G_r\}$ consists of distinct equivalence classes to each of which there corresponds a different extension sequence. Denote a general form of an infinite Markov sequence whose first r Markov matrices are $\{G_1, \cdots, G_r\}$ by

$$\{G_1, G_2, \cdots, G_r, G^*_{r+1}, G^*_{r+2}, \cdots\}, \tag{4.5}$$

where G^*_{r+1}, G^*_{r+2} are some unknown matrices. The sequence (4.5) may have realizations of any minimal dimension $n^* \geqq n_r$.

Applying the derivation of the rth order Markov basis as in the proof of Theorem 4.2, to the sequence (4.5) and following the discussion that preceded this theorem the Markov bases $\mathcal{B}^*_r = (I^r_{n_r}, J^r_{n_r}; \mathcal{G}^*_r)$ are obtained where $\mathcal{G}^*_r$ may be divided into two disjoint sets $\mathcal{G}^*_r = \tilde{\mathcal{G}}_r \cup \mathcal{P}_r$. The first set

$$\tilde{\mathcal{G}}_r = \{g_{ijk} \mid k = 1, \cdots, \min(\bar{\nu}_i + \bar{\mu}_j, r),\ i \in \mathbf{l},\ j \in \mathbf{m}\} \tag{4.6}$$

with $\bar{\nu}_i$ and $\bar{\mu}_j$ as defined in (4.3) represents the specified invariants that form a selection of entries of $\{G_1, \cdots, G_r\}$ while the second set

$$\mathcal{P}_r = \{g_{ijk} \mid k = r+1, \cdots, \bar{\nu}_i + \bar{\mu}_j > r,\ i \in \mathbf{l},\ j \in \mathbf{m}\} \tag{4.7}$$

represents a complementary set of unspecified invariants that form a selection of entries of the extension segment $\{G^*_{r+1}, \cdots, G^*_{r_0}\}$ of positions specified by $I^r_{n_r}$ and $J^r_{n_r}$. It follows from the nested property of Markov bases (Theorem 4.2), that the above set of bases $\mathcal{B}^*_r$ represents the set of rth order sub-bases of any general sequence (4.5). Therefore, any admissible extension sequence of dimension n_r is represented by $\mathcal{B}^*_r$ for some suitable choice of values for $\mathcal{P}_r$. The set $\mathcal{P}_r$ is a complete set of parameters $\{g^*_{ijk}\}$ for $S^r_{n_r}$, labelled by the locations of the required unspecified entries in the extension sequence. Two m.p.r.'s of $\{G_1, \cdots, G_r\}$ that assign values to $\mathcal{P}_r$ and are different even in one labelled parameter value represent different equivalence classes in $S^r_{n_r}$. The question now arises whether by arbitrarily assigning numerical values to the set of parameters $\mathcal{P}_r$, the resultant set of invariants $\{I^r_{n_r}, J^r_{n_r}; \tilde{\mathcal{G}}_r \cup \mathcal{P}_r\}$ is a basis of some m.p.r. of $\{G_1, \cdots, G_r\}$, or in other words, whether the relation between $S^r_{n_r}/E_{n_r}$ and $\mathcal{P}_r$ is also surjective (onto). Since the set of parameters $\mathcal{P}_r$ is taken from locations in H^r whose specification cannot affect the rank condition $n_r = \rho H^r$, we get the following result:

PROPOSITION 4.4. *There exists a one-to-one and onto (a bijective) relationship between $S^r_{n_r}/E_{n_r}$, the set of equivalent classes in $S^r_{n_r}$, and the set of parameters $\mathcal{P}_r$.*

It has been noted, in the paragraph following Theorem 3.1, that other bases which correspond to nice selections of arithmetic invariants other than the choice of the first set of independent rows and columns may be found. Choice of such bases for the partial realization would lead to an algebraic set of invariants which would contain both specified and unspecified invariants. It can be shown that though the unspecified invariants form alternative candidates for the parametrization of the set $S^r_{n_r}$ and satisfy the one-to-one relationship of the last proposition, they do not satisfy the onto relationship. The set $\mathcal{P}_r$ is the largest set of unspecified Markov entries to which we may assign values independently and the smallest set of parameters for $S^r_{n_r}$ that covers all m.p.r.'s of order r. We restate and prove this claim as Proposition 4.5.

PROPOSITION 4.5. *The set $\mathcal{P}_r$ is* (i) *an independent set, and equivalently,* (ii) *a minimal set of parameters for the parametrization of the set of all minimal partial realizations of $\{G_1, \cdots, G_r\}$.*

Proof. See Appendix 2.

It is considered important in some fields of system theory, such as certain problems of adaptive modelling identification to have a description of the set of all m.p.r.'s with the least possible set of parameters. It follows from the last proposition that only $\mathcal{P}_r$ results in such a description. This useful complete invariant description of the set $S^r_{n_r}$ of all minimal partial realization is summarized by the following:

THEOREM 4.6. *The set of all minimal partial realizations $S^r_{n_r}$ of a finite rth order sequence is completely determined by the set of nested Markov bases $\mathcal{B}^*_r = (I^r_{n_r}, J^r_{n_r}; \tilde{\mathcal{G}}_r \cup \mathcal{P}_r)$ where $\tilde{\mathcal{G}}_r$ and $\mathcal{P}_r$ are defined in* (4.6) *and* (4.7) *respectively. $\mathcal{P}_r$ is a minimal set of independent parameters for $S^r_{n_r}$ and the relation between the set of equivalence classes in $S^r_{n_r}$ and the set of parameters $\mathcal{P}_r$, $S^r_{n_r}/E_{n_r} \to \mathcal{P}_r$ is one-to-one and onto.*

Remark 4.1. The number of parameters in $\mathcal{P}_r$ is determined by the arithmetic invariants (implicitly via (4.7)).

Remark 4.2. The equivalence classes in $S^r_{n_r}$ have the following list of system invariants in common: The arithmetic invariants $I^r_{n_r}, J^r_{n_r}$ (and as a consequence the controllability and observability indices), the subset of the algebraic invariants $\tilde{\mathcal{G}}_r$ of (4.6) and $\#\mathcal{P}_r$, the minimal number of the above-mentioned parameters. These equivalence classes in $S^r_{n_r}$ differ only in the numerical values acquired by the set $\mathcal{P}_r$.

Remark 4.3. In the special case where the m.p.r. of $\{G_1, G_2, \cdots, G_r\}$ is unique the theorem implies the following: $S^r_{n_r}$ reduces to a single equivalence class for which $\mathscr{B}_r = (I^r_{n_r}, J^r_{n_r}; \tilde{\mathscr{G}}_r)$ is the corresponding Markov basis and the minimal set of parameters, $\mathscr{P}_r$, is empty.

The description of the minimal partial realizations need not be confined to nested Markov bases of invariants. It has been mentioned in the preceding section that any canonical representation can be derived from an ordinary Markov basis. In a similar manner any canonical representation can be derived from $\mathscr{B}^*_r$ for equivalent descriptions of $S^r_{n_r}$. Let $(\hat{A}_r, \hat{B}_r, \hat{C}_r)$ be the triple of matrices of Definition 3.1, derived from H^r in association with $\mathscr{B}^*_r = (I^r_{n_r}, J^r_{n_r}; \tilde{\mathscr{G}}_r \cup \mathscr{P}_r)$. The first and the second canonical forms of Theorem 3.2, (A_1, B_1, C_1), $(A_2, B_2, C_2) \in S^r_{n_r}$, can be derived from $(\hat{A}_r, \hat{B}_r, \hat{C}_r)$ by using a method analogous to the method of § 3,

$$[B_1, A_1] = T_1[\hat{B}_r; \hat{A}_r] \quad \text{and} \quad C_1 = C_r \tag{4.8}$$

and

$$B_2 = B_r \quad \text{and} \quad \begin{bmatrix} C_2 \\ A_2 \end{bmatrix} = \begin{bmatrix} \hat{C}_r \\ \hat{A}_r \end{bmatrix} T_2, \tag{4.9}$$

where T_1 and T_2 represent, respectively, the row and the column elimination operations, of Algorithm 3.1. Bases of canonical invariants $\mathscr{B}^*_1 = (I^r_{n_r}; \mathscr{G}^*_1)$ and $\mathscr{B}^*_2 = (J^r_{n_r}; \mathscr{G}^*_2)$ can also be derived for these canonical representations in accordance with Theorem 3.2. The difference between the m.p.r. canonical descriptions in the present case and the minimal (complete) realization description by system invariants of § 3 becomes significant in the case of $r < r_0$. In this case, which corresponds to the existence of more than one solution to the m.p.r. problem, the canonical representations as well as their corresponding canonical bases of invariants contain undetermined entries which are expressed by combinations of the minimal set of parameters $\mathscr{P}_r$.

Some other points of significance about the set $\mathscr{P}_r$ that make it further useful in certain problems of system identification are as follows. The set $\mathscr{P}_r$ is formed by assembling parameters in a form that can directly use further data that may be available under excessive measurements. Furthermore, as the parameters $\{g^*_{ijk}\}$ in the set $\mathscr{P}_r$ are labelled by their position in the extending data set, $\mathscr{P}_r$ contains information that indicates precisely which output-input pairs of relations (i, j) require further exploration and in what way can the model be completely specified.

Example 4.1. We shall illustrate the invariant description concepts presented above for m.p.r.'s by deriving nested bases of invariants and canonical realizations for sequences of order $r = 2, 3, 4$ for the numerical example of Tether [2].

$$G_1, G_2, G_3, G_4 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 4 & 3 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 10 & 7 \\ 1 & 1 \end{bmatrix}, \begin{bmatrix} 22 & 15 \\ 3 & 3 \end{bmatrix}. \tag{4.10}$$

Fourth order m.p.r.'s for this example were also derived in [5]–[8]. Nested bases of invariants are suggestive of recursive algorithms of realizations of sequences of successive higher orders. Since an efficient algorithm of this kind requires details which were not discussed in the present context, we shall derive invariant descriptions of m.p.r.'s separately for each order. For the sake of brevity we shall derive realizations only in the second canonical forms.

(a) *Fourth order sequence.* Construct for $r=4$, H^4, the fourth order incomplete Hankel matrix of (4.1)

$$
(4.11)\qquad H^r = \left[\begin{array}{cc|cc|cc|cc}
\boxed{1} & 1 & 4 & 3 & 10 & 7 & 22 & 15 \\
0 & 0 & 0 & 0 & \boxed{1} & 1 & 3 & 3 \\
\hline
4 & \boxed{3} & 10 & 7 & 22 & 15 & g_{115} & g_{125} \\
0 & 0 & \boxed{1} & 1 & 3 & 3 & g_{215} & g_{225} \\
\hline
10 & 7 & 22 & 15 & g_{115} & g_{125} & g_{116} & g_{126} \\
1 & 1 & 3 & \boxed{3} & g_{215} & g_{225} & g_{216} & g_{226} \\
\hline
22 & 15 & g_{115} & g_{125} & g_{116} & g_{126} & g_{117} & g_{127} \\
3 & 3 & g_{215} & g_{225} & g_{216} & g_{226} & g_{217} & g_{227}
\end{array}\right].
$$

The squared entries in (4.11) represent the pivotal elements determined by the numerically specified entries. (In a numerical example we may drop the asterisks used to mark unspecified entries. These can be found, for example, by the row reserving elimination operation (Definition 3.2). We observe that a m.p.r. of order $r=4$ is of dimension $\rho H^4 = 5$. The first independent rows are $I_5^4 = \{1, 2, 3, 4, 6\}$ thus $\bar{\nu}_i = \#\{1, 3\} = 2$, $\bar{\nu}_2 = \#\{2, 4, 6\} = 3$ by which $\nu^4 = \{2, 3\}$. Similarly, the first independent columns are $J_5^4 = \{1, 2, 3, 4, 5\}$ hence $\mu^4 = \{3, 2\}$ and a fourth order m.p.r. is determined by entries of the first $r_0 = 3+3 = 6$ Markov matrices. The set of Markov bases therefore consists of I_5^4, J_5^4 as the arithmetic invariants and $\tilde{\mathcal{G}}_4 \cup \mathcal{P}_4$ as the algebraic invariants, where $\tilde{\mathcal{G}}_4$ and $\mathcal{P}_4$ are determined by (4.6) and (4.7), respectively, to be $\tilde{\mathcal{G}}_4 = \{(g_{11k}, g_{12k}, g_{21k}, g_{22k}),\ k \in 4\}$ and $\mathcal{P}_4 = \{g_{115}, g_{215}, g_{225}, g_{216}\}$. These invariants are summarized in the upper part of Table 4.1 where the algebraic invariants appear as encircled entries in the Markov matrices. Associated with the set of bases $\mathcal{B}_4^* = \{I_5^4, J_5^4; \tilde{\mathcal{G}}_4 \cup \mathcal{P}_4\}$ are the triple of matrices $(\hat{A}, \hat{B}, \hat{C})$ of Definition 3.1,

$$
\hat{A} = \begin{bmatrix}
4 & 3 & 10 & 7 & 22 \\
0 & 0 & 1 & 1 & 3 \\
10 & 7 & 22 & 15 & g_{115} \\
1 & 1 & 3 & 3 & g_{215} \\
3 & 3 & g_{215} & g_{225} & g_{216}
\end{bmatrix}, \qquad
B = \begin{bmatrix}
1 & 1 \\
0 & 0 \\
4 & 3 \\
0 & 0 \\
1 & 1
\end{bmatrix},
$$

$$
\hat{C} = \begin{bmatrix}
1 & 1 & 4 & 3 & 10 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
$$

from which the second canonical form (A_2, B_2, C_2) can be obtained by Algorithm 3.1 resulting in $B_2 = \hat{B}$ and

$$
\begin{bmatrix} C_2 \\ A_2 \end{bmatrix} = \left[\begin{array}{ccccc}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
\hline
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
-2 & a & 3 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
b & d & 0 & c & -b+3
\end{array}\right],
$$

where $a = g_{115} - 46$, $b = g_{215} - g_{225}$, $c = g_{225} - 9$, $d = g_{216} - 3g_{215} + 6 - (g_{215} - 7) \times (g_{125}g_{215} + 3)$. The corresponding canonical invariant basis is $\mathscr{B}_2^* = (I_5^4; \mathscr{G}_2^*)$ where $\mathscr{G}_2^*$ is formed by the entries of B_2 and of rows 3, 5 of A_2.

(b) *Third order sequence*: $\{G_1, G_2, G_3\}$. The three upper-block diagonals of (4.8) reveal that H^3, the incomplete Hankel matrix required for $r = 3$, is characterized by $n_3 = 4$ and that the first four independent rows and columns of H^3 are, respectively, $I_4^3 = \{1, 2, 3, 4\} \to \nu^3 = \{2, 2\}$ and $J_4^3 = \{1, 2, 3, 5\} \to \mu^3 = \{3, 1\}$ from which $r_0 = 5$. The set of Markov bases are $\mathscr{B}_3^* = \{I_4^3, J_4^3; \tilde{\mathscr{G}}_3 \cup \mathscr{P}_3\}$ where $\tilde{\mathscr{G}}_3$ and $\mathscr{P}_3$ are formed by the encircled entries in the middle part of Table 4.1. The associated triple of matrices $(\hat{A}, \hat{B}, \hat{C})$ for these invariants are,

$$\hat{A} = \begin{bmatrix} 4 & 3 & 10 & g_{114} \\ 0 & 0 & 1 & g_{214} \\ 10 & 7 & g_{114} & g_{115} \\ 1 & 1 & g_{214} & g_{215} \end{bmatrix}, \qquad \hat{B} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 4 & 3 \\ 0 & 0 \end{bmatrix}, \qquad \hat{C} = \begin{bmatrix} 1 & 1 & 4 & 10 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The second canonical form (A_2, B_2, C_2) is readily obtained from these matrices by Algorithm 3.1; $B_2 = B$ and

$$\begin{bmatrix} C_2 \\ A_2 \end{bmatrix} = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -2 & c & 3 & a \\ 1 & d & 0 & b \end{array}\right],$$

where $a = g_{114} - 22$, $b = g_{214} - 4$, $c = g_{115} - 3g_{114} - g_{214}(g_{114} - 20) + 20$, $d = g_{215} - g_{214}(g_{214} - 4) - 10$. The corresponding canonical invariant bases are $\mathscr{B}_2^* = (I_4^3; \mathscr{G}_2^*)$ where $\mathscr{G}_2^*$ is composed of the entries of B_2 and of rows 3, 4 of A_2.

(c) *Second order sequence* $\{G_1, G_2\}$. Repetition of the above procedure for $r = 2$ yields $n_2 = 2$ $I_2^2 = \{1, 3\} \to \nu^2 = \{2, 0\}$, $J_2^2 = \{1, 2\} \to \mu^2 = \{1, 1\}$, by which $r_0 = 3$. The set of Markov bases are $\mathscr{B}_2^* = \{I_2^2, J_2^2; \tilde{\mathscr{G}}_2 \cup \mathscr{P}_2\}$ where $\tilde{\mathscr{G}}_2$ and $\mathscr{P}_2$ respectively are formed by the specified and the unspecified encircled entries in the $r_0 = 3$ Markov matrices in the lower part of Table 4.1. From the associated triple of matrices $(\hat{A}, \hat{B}, \hat{C})$,

$$\hat{A} = \begin{bmatrix} 4 & 3 \\ g_{113} & g_{123} \end{bmatrix}, \qquad \hat{B} = \begin{bmatrix} 1 & 1 \\ 4 & 3 \end{bmatrix}, \qquad \hat{C} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

the following realization in the second form is found

$$A_2 = \begin{bmatrix} 0 & 1 \\ a & b \end{bmatrix}, \qquad B_2 = \begin{bmatrix} 1 & 1 \\ 4 & 3 \end{bmatrix}, \qquad C_2 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix},$$

where $a = 4g_{113} - 3g_{123}$, $b = g_{113} - g_{123}$ and the corresponding canonical set of bases are $(I_2^2; \mathscr{G}_2^*)$ with $\mathscr{G}_2^*$ containing the entries of B_2 and the second column of A_2.

Table 4.1 summarizes the invariants of the realizations of orders $r = 4, 3, 2$ and exhibits their nested property. Our results can be compared for the $r = 4$ case, with the previous realizations in [2], [5]–[8]. Tether [2] suggests a minimal extension segment $\{G_{r+1}, \cdots, G_{r_0}\} = \{G_5, G_6\}$ that contains only two free parameters which in comparison with our results corresponds to two unnecessary constraints on $\mathscr{P}_4$, namely $g_{115} = 46$, $g_{215} = g_{225}$. The realization in [5] identifies only three free parameters for $S_{n_4}^4$ and has other weaknesses discussed in [6]. The authors in [7] and [8] correctly

TABLE 4.1
Nested bases of invariants for $r = 2, 3, 4$ *realization of* (4.10).

$r = 4, \quad n_4 = 5, \quad I_5^4 = \{1, 2, 3, 4, 6\}, \quad J_5^4 = \{1, 2, 3, 4, 5\}, \quad r_0 = 6$

$$G_1, G_2, G_3, G_4, G_5^*, G_6^* = \begin{bmatrix} ① & ① \\ ⓪ & ⓪ \end{bmatrix}, \begin{bmatrix} ④ & ③ \\ ⓪ & ⓪ \end{bmatrix}, \begin{bmatrix} ⑩ & ⑦ \\ ① & ① \end{bmatrix}, \begin{bmatrix} ㉒ & ⑮ \\ ③ & ③ \end{bmatrix}, \begin{bmatrix} ⓖ_{115} & g_{125} \\ ⓖ_{215} & ⓖ_{225} \end{bmatrix}, \begin{bmatrix} g_{116} & g_{126} \\ ⓖ_{216} & g_{226} \end{bmatrix}$$

$\mathscr{P}_4 = \{g_{115}, g_{215}, g_{225}, g_{216}\}$

$r = 3, \quad n_3 = 4, \quad I_4^3 = \{1, 2, 3, 4\} \subset I_5^4, \quad J_4^3 = \{1, 2, 3, 5\} \subset J_5^4, \quad r_0 = 5$

$$G_1, G_2, G_3, G_4^*, G_5^* = \begin{bmatrix} ① & ① \\ ⓪ & ⓪ \end{bmatrix}, \begin{bmatrix} ④ & ③ \\ ⓪ & ⓪ \end{bmatrix}, \begin{bmatrix} ⑩ & ⑦ \\ ① & ① \end{bmatrix}, \begin{bmatrix} ⓖ_{114} & g_{124} \\ ⓖ_{214} & g_{224} \end{bmatrix}, \begin{bmatrix} ⓖ_{115} & g_{125} \\ ⓖ_{215} & g_{225} \end{bmatrix}$$

$\mathscr{P}_3 = \{g_{114}, g_{214}, g_{115}, g_{215}\}$

$r = 2, \quad n_2 = 2, \quad I_2^2 = \{1, 3\} \subset I_4^3, \quad J_2^2 = \{1, 2\} \subset J_4^3, \quad r_0 = 3$

$$G_1, G_2, G_3^* = \begin{bmatrix} ① & ① \\ ⓪ & ⓪ \end{bmatrix}, \begin{bmatrix} ④ & ③ \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} ⓖ_{113} & ⓖ_{123} \\ g_{213} & g_{223} \end{bmatrix}$$

$\mathscr{P}_2 = \{g_{113}, g_{123}\}$

identify four independent parameters. Their descriptions use Popov type system invariants [9] and the representation is admitted in [7] to be nonunique. The realization there is into arbitrary Luenberger forms [13] by which the unity vectors in $[{}_A^C]$ or $[B, A]$ appear in arbitrary order and in positions that are not related to the system output or input structure. The computation in [7] requires solutions of sets of linear equations and the elimination procedure in [8] requires an auxiliary matrix. By comparison with these former invariant description approaches to m.p.r. our method is also advantageous computationally.

5. Conclusions. This paper studies the minimal partial realization (m.p.r.) problem using system invariant descriptions. The concept of nested bases of invariants is introduced and these bases are derived from entries in specified positions of the Markov sequences. These bases form invariant descriptions for m.p.r.'s which, in contrast to previous approaches, do not depend on any particular choice of a canonical representation. The existence of a unique solution to the m.p.r. problem can be tested on these invariants and when more than one solution exists the set of all m.p.r.'s for the given finite sequence can be expressed as a set of bases that contains a subset of undetermined invariants. The nesting property of these bases is used to prove that this set of undetermined invariants forms a minimal set of independent parameters that covers all possible m.p.r.'s of the sequence and that for any arbitrarily assigned values of these parameters there corresponds some admissible solution.

Two canonical state space representations have been suggested and an efficient algorithm for their derivation from the nested bases is provided. These canonical forms reflect the structural properties of the underlying system and also compare favorably in their numerical aspects with previous approaches to m.p.r.'s.

Any other canonical representation can alternatively be derived from these bases, and the solution to the m.p.r. problem can be expressed by combinations of the minimal set of parameters obtained. The complete freedom in assigning values to

these parameters may be used to search for further properties of the constructed models. (e.g., to ask for stable models). These parameters form entries of specific positions in the unknown extension sequence of the Markov matrices which may be of importance in certain identification problems. The formulation may also be advantageous in building adaptive real time identification models from input-output data. In this latter case an estimated state space model can be continuously updated by measuring only specific locations in the input-output map prescribed by the basis of invariants (where the model may be taken to be valid so long as the arithmetic invariants remain unchanged).

An obvious property of the suggested nested bases of invariants which has not been put to use in the present context is that these bases are ideal for recursive m.p.r. algorithms for sequences of Markov matrices of successive orders. This stems from the projective property of nested bases, i.e., they present sub-bases not only for all possible minimal extension sequences but also for arbitrary extension sequences of higher dimensions. Such a sequential algorithm, whose detailed numerical aspects have yet to be developed, will have the following features. The dimension of a realization of a sequence of a given length need not be known in advance. Subsequent order realizations require the calculation of only a few new invariants which add to the former set of invariants to form the new basis. The final important feature is that at each stage either the unique m.p.r. or in the nonunique case, the set of all possible m.p.r.'s are obtained and in the latter case these m.p.r.'s are described in terms of a minimal set of parameters.

Appendix 1. Proof of Theorem 3.2. We shall prove only the first part of the theorem, as the second part follows by an obvious dual reasoning. Statements (1a) and (1b) follow from Remarks 3.2 and 3.3. We have to show that $\mathcal{B}_1 = (I_n; \mathcal{G}_1)$ is a basis of invariants. I_n represents the arithmetic invariants associated with the observability indices (3.4) of the underlying system. The elements of the set $\mathcal{G}_1$ are entries in a canonical representation, thus they are canonical invariants. The set $(I_n; \mathcal{G}_1)$ is complete because it completely determines (A_1, B_1, C_1) via statements (1a) and (1b). The pair (B_1, A_1) is controllable by statement (1b) by which an arbitrary choice of $(I_n; \mathcal{G}_1)$ fails to give rise to a representation $(A_1, B_1, C_1) \in \Sigma_n$ if and only if it yields an unobservable pair (A_1, C_1). This condition is equivalent to $\rho\left[\begin{smallmatrix} C \\ A \end{smallmatrix}\right] < n$ and it can be expressed by suitable sets V_i of (2.8). Consequently the map $\mathcal{G}_1 : \Sigma_n \to R^{n(m+l)}$ is surjective except possibly on some hypersurfaces of "measure zero" in its codomain, thus $\mathcal{G}_1$ is also an independent set of invariants in the sense of Definition 2.3.

Appendix 2. Proof of Proposition 4.5. Assume that $\mathcal{P}_r$ is not independent and let $\mathcal{P}_r = \mathcal{P}_1 \cup \mathcal{P}_2$ where $\mathcal{P}_1$ and $\mathcal{P}_2$ are subsets of independent and dependent parameters, respectively. Once $\mathcal{P}_1$ has been arbitrarily assigned values the set $\mathcal{P}_2$ is uniquely determined in contrast to the surjective relationship between $S^r_{n_r}/E_{n_r}$ and $\mathcal{P}_r$ stated in Proposition 4.4. Therefore all the parameters in $\mathcal{P}_r$ can be assigned values independently. Now we show the equivalence of (i) and (ii). For (ii) $\to$ (i), a minimal set of parameters has to be independent or else a smaller set can be extracted for the parametrization of $S^r_{n_r}$. For the converse, (i) $\to$ (ii), assume there exists another basis $\hat{\mathcal{B}}^*_r$ for $S^r_{n_r}$ whose set of algebraic invariants $\hat{\mathcal{G}}_r \cup \hat{\mathcal{P}}_r$ is composed of a smaller set of unspecified values $\#\hat{\mathcal{P}}_r < \#\mathcal{P}_r$. Then $\mathcal{G}_r \cup \mathcal{P}_r$ could be expressed as a function of $\hat{\mathcal{G}}_r \cup \hat{\mathcal{P}}_r$ which implies the contradiction that not all the parameters in $\mathcal{P}_r$ can be assigned values independently.

Acknowledgments. The author wishes to express his thanks to Dr. U. Shaked for his preview and discussion of the paper. His valuable suggestions significantly affected the revised form of this paper.

REFERENCES

[1] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[2] A. J. TETHER, *Construction of minimal linear state variable models from finite input-output data*, IEEE Trans. Automat. Contr., AC-15 (1970), pp. 427–436.

[3] R. E. KALMAN, *On minimal partial realization of linear input-output map*, in Aspects of Network and System Theory, Kalman and Declaris, eds., Holt, Rinehart and Winston, New York, 1971, pp. 385–407.

[4] B. W. DICKINSON, T. KAILATH AND M. MORF, *A minimal realization algorithm for matrix sequences*, IEEE Trans. Automat. Contr., AC-19 (1974), pp. 31–38.

[5] J. E. ACKERMAN, *On partial realization*, IEEE Trans. Automat. Contr., AC-17 (1972), p. 381.

[6] G. LEDWICH AND T. E. FORTMAN, *Comment on 'On partial realization' and "Author's reply"*, IEEE Trans. Automat. Contr., AC-19 (1974), pp. 625–627.

[7] J. E. ROMAN AND T. E. BULLOCK, *Minimal partial realizations in canonical forms*, IEEE Trans. Automat. Contr., AC-20 (1975), pp. 529–533.

[8] J. V. CANDY, M. E. WARREN AND T. E. BULLOCK, *Partial realization of invariant system descriptions*, Int. J. Contr., 28 (1978), pp. 113–127.

[9] V. M. POPOV, *Invariant descriptions of linear time invariant controllable systems*, this Journal, 10 (1972), pp. 254–264.

[10] J. RISSANEN, *Basis of invariants and canonical forms for linear dynamic systems*, Automatica, 10 (1974), pp. 175–182.

[11] O. H. BOSGRA AND A. J. VAN DER WEIDEN, *Input-output invariants for linear multivariable systems*, IEEE Trans. Automat. Contr., AC-25 (1980), pp. 20–36.

[12] M. J. DENHAM, *Canonical forms for the identification of multivariable linear systems*, IEEE Trans. Automat. Contr., AC-19 (1974), pp. 646–656.

[13] D. G. LUENBERGER, *Canonical forms for linear multivariable systems*, IEEE Trans. Automat. Contr., AC-12 (1967), pp. 290–293.

[14] L. M. SILVERMAN, *Realization of linear dynamical systems*, IEEE Trans. Automat. Contr., AC-16 (1971), pp. 554–567.

[15] H. AKAIKE, *Stochastic theory of minimal realization*, IEEE Trans. Automat. Contr., AC-19 (1974), pp. 667–674.

[16] G. MCLANE AND G. BIRKHOFF, *Algebra*, Macmillan, New York, 1967.

[17] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[18] R. E. KALMAN, *Kronecker invariants and feedback*, Proc. of N.R.L. Math. Res. Center Conference on Ordinary Differential Equations, June 1971.

IN MEMORIAM

This issue of *SIAM Journal on Control and Optimization* is dedicated to Joseph P. LaSalle, who died on July 7, 1983, at his home in Little Compton, Rhode Island. He had been ill for several years. Joe's pioneering work in control theory and differential equations has had tremendous impact on both fields and his leadership in the scientific community is of lasting importance.

SIAM and this journal in particular were fortunate recipients of his long-time support. Joe initiated the idea that we publish this journal and served on our editorial board from its beginning in 1963 until 1981. He was president of SIAM in 1962–63 and a member of its board of trustees from 1964 to 1967.

Joe was born on May 28, 1916, in State College, Pennsylvania, the son of Professor Leo J. LaSalle and Aline Mistric LaSalle. At that time his father taught at Pennsylvania State College. Four years later the family moved to Louisiana, where his father joined the faculty of Louisiana State University.

Joe majored in political science for two years at Louisiana State University. During the summer of 1935 he took courses in mathematics and logic because the dean of the Law School recommended mathematics as a good prerequisite for the study of law. Mathematics quickly became his major interest. In 1937 he graduated from Louisiana State University and then went to the California Institute of Technology to begin his graduate studies in mathematics. He was a Henry Laws Fellow at Cal Tech from 1939 until he received his Ph.D. in 1941 for a dissertation, directed by A. D. Michal, on pseudo-normed spaces.

During World War II LaSalle devoted his energies to more applied areas. He obtained his first academic appointment in 1942 as an instructor in the Department of Applied Mathematics and Astronomy at the University of Texas. In 1943, he was an instructor at the Radar School of the Massachusetts Institute of Technology. At Princeton in 1944, he served with a group of mathematicians as a scientific advisor to the U.S. Army Office of the Chief of Ordnance. From 1944 to 1946, he worked at Cornell with physicists on the design of a magnetron for a naval communications system.

In 1946, LaSalle accepted an Assistant Professorship in the Mathematics Department of the University of Notre Dame. He remained there until 1958, rising to the rank of professor in 1956. While a visitor at Princeton (1947–48) he became interested in differential equations through his association with Solomon Lefschetz. Here, also, he met Richard Bellman, and they became close personal friends. LaSalle taught Bellman to play tennis; in return Bellman taught LaSalle something about differential equations. This stimulated some of his early research on nonlinear oscillations, an area in which he continued to make significant contributions.

In 1958 LaSalle was invited by Lefschetz to join his research group on differential equations at the Research Institute for Advanced Studies (RIAS) in Baltimore. LaSalle's research on stability theory began with his collaboration with Lefschetz on what has become a classic work on Liapunov's direct method. It was here that he first published an extension of Liapunov theory based on what we now call the "invariance principle."

In 1964 LaSalle, Lefschetz and some of the members of RIAS moved to Brown and formed, within the Division of Applied Mathematics, what was to become the Lefschetz Center for Dynamical Systems. LaSalle served as Director of the center from 1964 to 1980, and as Chairman of the Division of Applied Mathematics at Brown University from 1968 to 1973.

In 1964, Joe founded the *Journal of Differential Equations,* and served as Editor-in-Chief until 1980. His high standards made this journal one of the most important in the field. LaSalle also played important roles as an editor of the *Journal of Mathematical Analysis and Applications,* and the Springer Applied Mathematical Sciences Series.

Joe LaSalle's early work in control theory dealt with the bang-bang principle, initially a conjecture which he stated as follows: "If a control system is being operated from a limited source of power and if one wishes to have the system change from one state to another in minimum time, then this can be done by at all times utilizing properly all the power available." It was typical of Joe to phrase mathematical ideas verbally whenever possible, rather than to limit his audience via complicated symbolism.

The bang-bang principle was an accepted hypothesis in the important dissertation of D. Bushaw (Princeton, 1952). LaSalle's first published contribution to the problem appeared in 1954 in a Bulletin of the American Mathematical Society abstract entitled "Study of the basic principle underlying the bang-bang servo." Here he showed that if, for a special form of a controlled, second order, nonlinear equation, there exists a unique bang-bang control which is best of all bang-bang controls, then it is also best among all controls.

The proof of a generalized bang-bang principle eluded researchers for the next several years. In conversation LaSalle confided that during the late 1950's he finally had the basic ideas of a proof for linear systems, albeit a complicated one. He discussed these with a young Russian mathematician who informed him of the similarity with a 1940 paper by A. Liapunov dealing with vector measures. Indeed, the Liapunov theorem on the range of a vector measure provided an elegant method of obtaining the bang-bang principle for linear systems, which was given in LaSalle's celebrated paper "The Time Optimal Problem" (Contributions to Diff. Eqs., 1959).

Here he showed that for an n-dimensional, time varying, linear control system having an r-dimensional control vector taking admissible values in the closed unit cube of R^r, the set of points attainable at any time $t_1 \geq 0$ from a given initial point by utilizing all admissible (measurable) controls equals the set attainable at time t_1 utilizing controls taking values in the set of vertices of the cube (the bang-bang controls). In doing so, he included a characterization of the form of a time optimal control (i.e., the maximum principle for time optimal problems governed by linear systems).

He next introduced the notion of a proper linear control system, i.e., one in which time-optimal controls are determined (except on set of zero measure) by the maximum principle. (Today, proper systems would be generalized as systems which exclude singular arcs.) The paper continues with "Think now of removing all of the constraints on the admissible control functions, and consider any two states x_1 and x_2 and any two times t_1 and t_2. If for each pair of states and pair of times there is a steering function such that starting at x_1, at time t_1, the system is brought to the state x_2 at time t_2, then the system is said to be completely controllable." His next theorem shows proper control systems are completely controllable; this is followed by the characterization that the autonomous n-dimensional, linear control system $\dot{x}=Ax+Bu$ is proper if and only if rank $[B, AB, \ldots, A^{n-1}B]=n$. One should reflect, for a moment, on the influence that these ideas have had on the last twenty years of control theory!

LaSalle's work focussed more on stability theory than on control. In 1962, his paper "Stability and Control" appeared in the first issue of this journal. Here he considers an n-dimensional control system $\dot{x}=F(x,u)$ with $F(0,0)=0$, the problem being to find a feedback control $u=u(x)$ which "drives the system to zero." Restated, if $F(x)=f(x,u(x))$, then zero is to be an asymptotically stable solution of $\dot{x}=F(x)$. The concept of introducing a Liapunov function V and generating a suboptimal (or optimal if V is the actual cost function) feedback control by choosing it to minimize $\dot{V}$, appears here.

The number of LaSalle's publications in control was small but the number of basic ideas these papers contained was substantial, and the influence they have had on the development of the theory during the last three decades has been monumental.

Although LaSalle's scientific achievements alone are of major significance, many feel that perhaps his greatest contributions were made in his role as a leader of mathematicians.

He provided a chance for many young mathematicians, from all over the country, to get a start in research. He created an atmosphere in which there was pressure to produce, but it was a healthy pressure which made the effort fun.

RIAS, under his leadership and the sponsorship of the Martin-Marietta Corporation, assembled an outstanding group of extremely productive mathematicians, physicists and engineers. Later, as director of the Lefschetz Center for Dynamical Systems at Brown, Joe influenced the thinking of many young visitors from the U.S. and abroad. Indeed, the authors of this memorial note are among the many indebted to Joe LaSalle for providing a stimulating atmosphere which was an important factor in their mathematical development.

H. T. Banks
H. G. Hermes
M. Q. Jacobs

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0001 $01.25/0

A DIRECT APPROACH TO COMPENSATOR DESIGN FOR DISTRIBUTED PARAMETER SYSTEMS*

J. M. SCHUMACHER†

Abstract. We present a direct approach to finite-order compensator design for distributed parameter systems, i.e., one that is not based on reduced order modelling. Instead, we use a parametrization around an initial compensator which displays both controller order and closed-loop stability in a convenient way. The main result is an existence theorem which holds for a wide class of linear time-invariant systems (parabolic, delay, damped hyperbolic). The most important assumptions are: bounded inputs and outputs, finitely many unstable modes, completeness of eigenvectors. An example is included to illustrate the feasibility of our method for purposes of design.

1. Introduction. In the context of systems described by linear partial differential equations or functional differential equations, the problem of stabilization by feedback gains some challenging features that are not present in the finite-dimensional situation. For instance, it is no longer easy to establish necessary and sufficient conditions for the existence of a finite-dimensional compensator that will produce a closed-loop system with a prescribed stability margin. It is an important practical problem to find at least sufficient conditions which will hold for a wide class of interesting systems, since implementation of state feedback [1], [2] or of controllers of infinite order [3], [4], [5] is often not possible. The most popular approach consists of replacing the infinite-dimensional system by a finite-dimensional "reduced order model" and applying standard techniques to obtain a finite-dimensional compensator for this model. The pertinent question is, of course, how we can be sure that the compensator will also stabilize the original, infinite-dimensional system. It has been shown by examples that, under unfavorable circumstances, the interaction of the controller with the unmodelled part of the system (sometimes termed "spillover") may be such as to destabilize the closed-loop system as a whole [6]. Existence results for finite-dimensional compensators have been established recently on the basis of a "zero spillover" assumption [5], [7], [8], but this assumption is severely restrictive. Also, existence results can be based on a suitable concept of "closeness" of the reduced-order model and the actual system. This approach is taken in [9], where the results are still limited in nature. At this point, it should be emphasized that a concept of "closeness" is also crucial in any study of parameter uncertainty. This aspect is, as well as order reduction, inherent in many discussions of modelling. For the sake of theoretical clarity, we shall keep these two issues apart. In the present paper, we shall assume that the infinite-dimensional system to be controlled is known precisely, and we shall construct a finite-dimensional compensator under this assumption. It is expected that this result can then be used in a further study of what can be done under conditions of parameter uncertainty.

Our approach is not based on reduced-order modelling, and therefore we call it a "direct approach". The core of our method is a certain parametrization of compensators for a given system, which displays both the stability properties of the closed-loop system and the order of the compensator in a convenient way. We shall

* Received by the editors December 15, 1980, and in final revised form August 13, 1982.

† The author was with the Department of Mathematics, Vrije Universiteit, Amsterdam, and with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts, where he was supported by the Netherlands Organization for the Advancement of Pure Scientific Research (ZWO). He is now with the Department of Mathematics, Erasmus Universiteit, Rotterdam, the Netherlands.

try to explain the basic idea in § 2. In § 3, the set-up is described in a more rigorous fashion. The main result, which establishes the existence of finite dimensional compensators for a wide class of time-invariant linear systems (including parabolic systems, delay systems and damped hyperbolic systems), will be given in § 4. The method of proof is constructive and can be turned into an actual design method, as will be shown by an example in § 5. Some final remarks follow in § 6.

2. Heuristics. The purpose of this section is to describe the main idea behind the development in the rest of the paper, without entering into technical details. A rigorous set-up will be described in the next section; here, we just want to give a heuristic discussion.

So let us consider a linear system in its standard state-space form

$$\begin{aligned} x'(t) &= Ax(t)+Bu(t), \\ y(t) &= Cx(t), \end{aligned} \tag{2.1}$$

where we assume that the pair (A, B) is stabilizable and the pair (C, A) is detectable. We can then choose F such that $A+BF$ is stable and G such that $A+GC$ is stable, and the standard full-order compensator (see, for instance, [10]) is then formed by

$$\begin{aligned} \hat{x}'(t) &= (A+GC)\hat{x}(t)-Gy(t)+Bu(t), \\ u(t) &= F\hat{x}(t). \end{aligned} \tag{2.2}$$

In the finite dimensional situation, it is well known that the closed-loop system obtained by combining (2.1) and (2.2) is described by a system matrix whose eigenvalues are those of $A+BF$ and $A+GC$ taken together [10, 5.2]. Let us examine the compensator (2.2) a little more closely. We can rewrite the compensator equations as

$$\begin{aligned} \hat{x}'(t) &= (A+BF+GC)\hat{x}(t)-Gy(t), \\ u(t) &= F\hat{x}(t) \end{aligned} \tag{2.3}$$

and hence the compensator transfer matrix is

$$\phi_c(s) = -F(sI-A-BF-GC)^{-1}G. \tag{2.4}$$

Now, there is no reason why (2.3) should represent a minimal realization of this transfer function. If it is not, then the compensator order can be reduced. Even if the McMillan degree of ϕ_c coincides with the order of the system (2.3), there may be transfer matrices with considerably lower McMillan degree that are close enough to ϕ_c to guarantee that they as well will stabilize (2.1). In order to find such transfer matrices, one possible strategy would be to take ϕ_c and to change it a little bit by turning near-cancellations into actual cancellations, thereby decreasing the order of its minimal realization.

The question is, of course, under what conditions we can be sure that such a procedure will lead to a finite-dimensional compensator, if the original system (2.1) is infinite-dimensional. To get at least a partial answer to this, let us return to the state-space setting. The realization (2.3) is nonminimal if the pair $(A+BF+GC, G)$ is not reachable or the pair $(F, A+BF+GC)$ is not observable. We shall concentrate on the reachable set of the pair $(A+BF+GC, G)$, which is of course the same as the reachable set of the pair $(A+BF, G)$. This set is characterized as the smallest subspace $\mathcal{V}$ such that $(A+BF)\mathcal{V} \subset \mathcal{V}$ and $\operatorname{im} G \subset \mathcal{V}$. The basic idea which underlies the present paper is the observation that, by manipulation of G alone, we can implement a strategy

of slightly perturbing the compensator transfer matrix to decrease its McMillan degree. Even if the original im G is not contained in any $(A+BF)$-invariant subspace of interesting dimension, it may very well be true that close to G there is a $\tilde{G}$ such that im $\tilde{G}$ does fit into a low-dimensional $(A+BF)$-invariant subspace. Then the reachable set of the pair $(A+BF+\tilde{G}C, \tilde{G})$ will also be low-dimensional, say equal to k, and it will be possible to construct a compensator of order k based on F and $\tilde{G}$. The stability of the closed-loop system will then depend on $A+BF$ and $A+\tilde{G}C$. We didn't change $A+BF$, so there is no problem for that part, and it follows from the theorem on continuity of eigenvalues that the stability of $A+\tilde{G}C$ follows from that of $A+GC$ if $\tilde{G}$ is close enough to G. (Actually, we shall use another theorem below, which gives us a ball around G where stability of $A+\tilde{G}C$ is guaranteed: see Lemma 4.3.)

It can also be seen directly from the differential equations (2.2) that a reduction of compensator order is possible if there is a nontrivial subspace $\mathcal{V}$ with $(A+BF)\mathcal{V} \subset \mathcal{V}$ and im $G \subset \mathcal{V}$. For this purpose, rewrite (2.2) as

$$\begin{aligned} &\hat{x}'(t) = (A+BF)\hat{x}(t) + G(C\hat{x}(t) - y(t)), \\ &u(t) = F\hat{x}(t). \end{aligned} \tag{2.5}$$

The equation for $\hat{x}(t)$ is seen to be given by the evolution operator $A+BF$ together with a driving input which enters through G. Since the stabilization action of the compensator should take place for any initial value of $\hat{x}(\cdot)$, we may as well suppose that $\hat{x}(0) = 0$. Then it is clear that $x(t)$ will be in $\mathcal{V}$ for all time. Consequently, no larger state space than $\mathcal{V}$ is necessary for $\hat{x}$.

As a third possible interpretation, consider the following matrix argument. Again, if $\mathcal{V}$ is a subspace such that $(A+BF)\mathcal{V} \subset \mathcal{V}$ and im $G \subset \mathcal{V}$, then we obviously have the following matrix representations for $A+BF$ and G, with respect to a suitable basis.

$$A+BF = \begin{pmatrix} A_{11}+B_1F_1 & A_{12}+B_1F_2 \\ 0 & A_{22}+B_2F_2 \end{pmatrix}, \qquad G = \begin{pmatrix} G_1 \\ 0 \end{pmatrix}. \tag{2.6}$$

As is easily established from (2.1) and (2.2), the equation describing the closed-loop system is

$$\frac{d}{dt}\begin{pmatrix} x(t) \\ \hat{x}(t) \end{pmatrix} = A_e \begin{pmatrix} x(t) \\ \hat{x}(t) \end{pmatrix}, \qquad A_e = \begin{pmatrix} A & BF \\ -GC & A+BF+GC \end{pmatrix}. \tag{2.7}$$

Using the special forms in (2.6) to describe the compensator dynamics, we see that the evolution operator A_e in (2.7) can be given as a three-by-three block matrix:

$$A_e = \begin{pmatrix} A & BF_1 & BF_2 \\ -G_1C & A_{11}+B_1F_1+G_1C_1 & A_{12}+B_1F_2+G_1C_2 \\ 0 & 0 & A_{22}+B_2F_2 \end{pmatrix}. \tag{2.8}$$

It is evident from this representation that if A_e is stable, then the two-by-two left upper block in A_e must also be stable. This means that we are able to build a stabilizing compensator (of order dim $\mathcal{V}$) based on G_1, F_1 and $A_{11}+B_1F_1+G_1C_1$. Technically speaking, this is perhaps the cleanest way to describe the situation, and we shall use basically this approach in the rigorous development of later sections.

In summary, the proposed method is the following. We start by selecting a full-order compensator that stabilizes the original system. Then, we parametrize a set of nearby compensators on the basis of the "injection mapping" G. This parametrization is not necessarily complete, but the stability of the resulting closed-loop systems is easily monitored, and, in particular, there is a ball around the original injection

mapping where stability is guaranteed. Moreover, the points in the parameter space where the compensator order is reduced to a given number k are easily spotted, because they correspond to the k-dimensional invariant subspaces of $A+BF$, which are, at least theoretically speaking, known. So this parametrization allows us to do an effective search for low-order stabilizing compensators. In the infinite-dimensional case, we expect that it will be possible to prove the existence of a finite-dimensional stabilizing compensator if there are finite-dimensional $(A+BF)$-invariant subspaces arbitrarily close to any given subspace, i.e., if we have completeness of eigenvectors. No further essential restrictions will be required. We shall now proceed to make this precise. It should be emphasized that the procedure we have sketched is meant for theoretical purposes; several alterations may be made to advantage, when a similar method is to be used for practical design purposes. This will be illustrated in the example of § 5.

3. Assumptions and preliminaries. We shall consider systems of the form

$$\begin{aligned} &x'(t)=Ax(t)+Bu(t), \qquad x(t)\in\mathscr{X}, \quad u(t)\in\mathscr{U}, \\ &y(t)=Cx(t), \qquad y(t)\in\mathscr{Y}, \end{aligned} \tag{3.1}$$

under the following basic assumptions:

(A1) A is the generator of a strongly continuous semigroup $T(\cdot)$ of bounded linear operators on the Banach space $\mathscr{X}$.

(A2) B is a bounded linear mapping from the finite dimensional input space $\mathscr{U}$ into $\mathscr{X}$.

(A3) C is a bounded linear mapping from $\mathscr{X}$ into the finite dimensional output space $\mathscr{Y}$.

For the general theory of semigroups, we refer to [11]. The condition (A2) requires that the control enters the system in a "distributed" way, i.e., as a forcing term, rather than via the boundary conditions. The condition (A3) excludes, for instance, taking point observations on an L_2-space. The case of unbounded input and output operators has been considered in [29], where an approach is used that is similar to ours.

Following [12, p. 181], we shall say that the spectrum of an operator is *discrete* if it consists only of isolated eigenvalues with finite multiplicities. We shall make the following assumption because it is convenient and also because it covers the commonly encountered cases.

(A4) The spectrum of A is discrete.

As a measure of stability, we shall use the *growth constant.* This constant is obtained for every semigroup $T(t)$ (from now on, we shall use the term "semigroup" as a synonym for "strongly continuous semigroup of bounded linear operators on a Banach space") by the following formula [11, p. 306]:

$$\omega_0 := \inf_{t\in[0,\infty)} \frac{1}{t}\log\|T(t)\| = \lim_{t\to\infty}\frac{1}{t}\log\|T(t)\| < \infty. \tag{3.2}$$

The semigroup is said to be *asymptotically stable* if its growth constant is negative, and the absolute value of the growth constant is then also called the *stability margin.* Obtaining a reasonable stability margin is a primary purpose of feedback control, and we shall suppose that a desired minimum degree of stability has been specified by a growth constant $\omega<0$ which will be fixed from now on. A semigroup will be called simply *stable* if its growth constant is smaller than or equal to ω. We shall assume that there are only finitely many unstable or nearly unstable modes.

(A5) There exists $\delta > 0$ such that the half-plane $\{\lambda \in \mathbb{C} | \operatorname{Re} \lambda > \omega - \delta\}$ contains only finitely many eigenvalues of A.

Under this assumption, we can draw a simple closed curve enclosing precisely those eigenvalues of A that have real parts larger than ω. From this, we obtain a decomposition of the state space $\mathscr{X}$ as in [12, p. 178]. We shall write $\mathscr{X} = \mathscr{X}_u \oplus \mathscr{X}_s$ where $\mathscr{X}_u$ is called the *unstable modal subspace* and $\mathscr{X}_s$ is the *stable modal subspace.* Correspondingly, the following notation will be used with respect to this decomposition:

$$A = \begin{pmatrix} A_u & 0 \\ 0 & A_s \end{pmatrix}, \quad B = \begin{pmatrix} B_u \\ B_s \end{pmatrix}, \quad C = (C_u \quad C_s). \tag{3.3}$$

As in the finite-dimensional case, we shall need assumptions on the stabilizability of the pair (A, B) and the detectability of the pair (C, A). In the present context, these are most easily expressed in the following way.

(A6) The pair (A_u, B_u) is controllable.

(A7) The pair (C_u, A_u) is observable.

Note that both pairs involve only operators between finite-dimensional spaces, so that we can rely on the familiar finite-dimensional concepts.

Next, we need an assumption of a somewhat more technical nature. Let $\delta > 0$ satisfy the condition of (A5). Then it is clear that one can also do a decomposition of $\mathscr{X}$ with respect to the eigenvalues of A that have real parts larger than $\omega - \delta$ (rather than ω). Let $A_s^{\omega-\delta}$ denote the operator that is obtained in this way, similarly to A_s. It has been shown in [2, App. 2] that $A_s^{\omega-\delta}$ generates a semigroup. We shall assume the following.

(A8) The growth constant of the semigroup generated by $A_s^{\omega-\delta}$ is smaller than ω.

We know, of course, that the eigenvalues of $A_s^{\omega-\delta}$ all have real parts smaller than or equal to $\omega - \delta$, but counterexamples [11, p. 665], [13] show that this in itself does not guarantee that the growth constant of the semigroup will be bounded by $\omega - \delta$ or by ω. One solution, then, is to introduce a "spectrum determined growth assumption" like (A8). This solution has been proposed in [2], where it has also been argued that the assumption holds for various important classes of semigroups.

For an alternative, we should consider our ultimate purposes. To the system (3.1), we want to add a finite dimensional *compensator* of the form

$$\begin{aligned} &w'(t) = A_c w(t) + G_c y(t), \qquad w(t) \in \mathscr{W}, \quad \dim \mathscr{W} < \infty, \\ &u(t) = F_c w(t) + K y(t). \end{aligned} \tag{3.4}$$

Doing so, we obtain a *closed-loop system* which looks like

$$\frac{d}{dt}\begin{pmatrix} x \\ w \end{pmatrix}(t) = A_e \begin{pmatrix} x \\ w \end{pmatrix}(t), \tag{3.5}$$

where the *closed-loop system mapping* A_e is given by

$$A_e = \begin{pmatrix} A + BKC & BF_c \\ G_c C & A_c \end{pmatrix}. \tag{3.6}$$

This operator generates a semigroup on $\mathscr{X} \oplus \mathscr{W}$, since it is a bounded perturbation of

$$\tilde{A}_e = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \tag{3.7}$$

[11, p. 389]. For our purposes, it will be easily sufficient if we know the following.

(A8)′ For any choice of the matrices K, F_c, G_c and A_c in (3.6), the growth constant of the semigroup generated by A_e is equal to $\sup\{\operatorname{Re}\lambda \mid \lambda \in \sigma(A_e)\}$.

We shall primarily use (A8), because this assumption is probably in most cases more directly verifiable (see [2]). However, in some instances it may be easy to check that (A8)′ is true, and then (A8) can be dispensed with. In engineering contexts, (A8)′ is often assumed without mentioning.

For our final assumption, we point out that we shall call any non-zero vector in the range of the eigenprojection associated with a given eigenvalue [12, p. 181] an eigenvector, so this includes "generalized eigenvectors". A set of elements of $\mathscr{X}$ is called *complete* (in $\mathscr{X}$) if the finite linear combinations of these elements form a dense set in $\mathscr{X}$. We assume the following.

(A9) The eigenvectors of A form a complete set in $\mathscr{X}$.

Completeness of eigenvectors is a common property for diffusion operators, delay operators and wave operators as well; see, for instance, [14, p. 325], [15, pp. 465–470], [16, pp. 278–289], [17], [18] and [19, p. 250]. Under the stated assumptions, it will be shown below (Lemma 4.5) that there exists a feedback mapping $F: \mathscr{X} \to \mathscr{U}$ such that the spectrum of $A+BF$ is discrete and contained in $\{\lambda \in \mathbb{C} \mid \operatorname{Re}\lambda \leqq \omega\}$ and such that the eigenvectors of $A+BF$ form a complete set in $\mathscr{X}$. We could use this statement to replace both (A6) and (A9), but since these assumptions are stated directly in terms of A, we prefer to use them, rather than an indirect (be it weaker) expression.

For easy reference, we shall state here the following lemma, which will be used repeatedly. The proof presents no basic difficulties and will be omitted.

LEMMA 3.1. *Suppose that A_{11} and A_{22} are generators of semigroups on the Banach spaces $\mathscr{X}_1$ and $\mathscr{X}_2$, respectively, with growth constants ω_1, and ω_2. Suppose also that $A_{21}: \mathscr{X}_1 \to \mathscr{X}_2$ is a bounded linear mapping. Then the operator on $\mathscr{X}_1 \oplus \mathscr{X}_2$ defined by*

$$A = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \tag{3.8}$$

generates a semigroup whose growth constant equals $\max(\omega_1, \omega_2)$.

4. Existence result. Our aim in this section is to prove the following result.

THEOREM 4.1. *Consider the system* (3.1) *and suppose that the assumptions* (A1)–(A8) *hold for some given growth constant ω. Then there exists a compensator of finite order such that the evolution of the controlled system is described by a strongly continuous semigroup with growth constant smaller than or equal to ω.*

For convenience, we shall break up the proof of this theorem into four separate lemmas.

LEMMA 4.2. *Consider the system* (3.1) *under the assumptions* (A1)–(A3). *Let ω be a given growth constant and suppose that there exist a finite dimensional subspace $\mathscr{V} \subset D(A)$ and linear mappings $F: \mathscr{V} \to \mathscr{U}$ and $G: \mathscr{Y} \to \mathscr{X}$ with the following properties*:

(4.1) $\operatorname{im} G \subset \mathscr{V}$,

(4.2) *the semigroup generated by $A+GC$ has growth constant $\omega_1 \leqq \omega$,*

(4.3) $(A+BF)x \in \mathscr{V}$ *for all* $x \in \mathscr{V}$,

(4.4) *the (finite-dimensional) semigroup generated by $A+BF|_{\mathscr{V}}$ has growth constant $\omega_2 \leqq \omega$.*

Then there exists a compensator of the form (3.4), *which has (finite) order equal to* $\dim \mathscr{V}$ *and which is such that the evolution of the controlled system is described by a semigroup with growth constant* $\max(\omega_1, \omega_2) \leqq \omega$.

Proof. Introduce a new linear space $\mathcal{W}$ isomorphic to $\mathcal{V}$ and let $R:\mathcal{V}\to\mathcal{W}$ be the mapping that provides the isomorphism. Define a compensator of the form (3.4) by setting $K=0$, $F_c=FR^{-1}$, $G_c=-RG$ and $A_c=R(A+BF+GC)R^{-1}$. (Note that it follows from (4.1) and (4.3) that G_c and A_c are well defined, even though R is not defined on all of $\mathcal{X}$.) We can write the following differential equation for the controlled system:

$$\frac{d}{dt}\begin{pmatrix}x\\w\end{pmatrix}(t)=A_e\begin{pmatrix}x\\w\end{pmatrix}(t) \tag{4.5}$$

with the extended system mapping A_e given by

$$A_e=\begin{pmatrix}A & BF_c\\ G_cC & A_c\end{pmatrix}=\begin{pmatrix}A & BFR^{-1}\\ -RGC & R(A+BF+GC)R^{-1}\end{pmatrix}. \tag{4.6}$$

Consider the following subspace of the extended state space $\mathcal{X}_e:=\mathcal{X}\oplus\mathcal{W}$:

$$\mathcal{M}:=\left\{\begin{pmatrix}x\\Rx\end{pmatrix}\middle|\, x\in\mathcal{V}\right\}. \tag{4.7}$$

There is an obvious isomorphism between $\mathcal{V}$ and $\mathcal{M}$, given by

$$Tx=\begin{pmatrix}x\\Rx\end{pmatrix},\qquad x\in\mathcal{V}. \tag{4.8}$$

The space $\mathcal{X}_e$ can also be decomposed as $\mathcal{X}\oplus\mathcal{M}$, rather than as $\mathcal{X}\oplus\mathcal{W}$. Written with respect to this decomposition, A_e will have the form

$$\tilde{A}_e:=HA_eH^{-1}, \tag{4.9}$$

where the isomorphism $H:\mathcal{X}\oplus\mathcal{W}\to\mathcal{X}\oplus\mathcal{M}$ is defined by

$$H=\begin{pmatrix}I & -R^{-1}\\ 0 & TR^{-1}\end{pmatrix}. \tag{4.10}$$

By straightforward computation, we find that

$$\tilde{A}_e=\begin{pmatrix}A+GC & 0\\ -TGC & T(A+BF)T^{-1}\end{pmatrix}. \tag{4.11}$$

Noting that $T(A+BF)T^{-1}$ is similar to $A+BF|_{\mathcal{V}}$, we now immediately get the result by an application of Lemma 3.1.

LEMMA 4.3. *Consider a pair of mappings (C,A) under the assumptions* (A1), (A3), (A4), (A5), (A7) *and* (A8). *Then we can find a linear mapping $G:\mathcal{Y}\to\mathcal{X}$ and a constant $\eta>0$ such that, for every $\tilde{G}:\mathcal{Y}\to\mathcal{X}$ satisfying $\|G-\tilde{G}\|<\eta$, the semigroup generated by $A+\tilde{G}C$ is stable.*

Proof. We shall use the same modal decomposition that has been used to formulate (A8), and we shall further decompose the "unstable" parts $A_u^{\omega-\delta}$ and $C_u^{\omega-\delta}$ (cf. (3.3)) in order to display the unobservable subspace of this pair. The final result of these operations is a decomposition of the form

$$A=\begin{pmatrix}A_{11} & 0 & 0\\ 0 & A_{22} & 0\\ 0 & A_{32} & A_{33}\end{pmatrix},\qquad C=(C_1\quad C_2\quad 0), \tag{4.12}$$

where $\operatorname{Re}\lambda\leqq\omega-\delta$ for $\lambda\in\sigma(A_{11})$, the pair (C_2,A_{22}) is observable, and $\omega-\delta<\operatorname{Re}\lambda\leqq\omega$ for $\lambda\in\sigma(A_{33})$. (The last inequality follows from (A7).) By the observability of the

pair (C_2, A_{22}), there exists a G_2 such that all eigenvalues of $A_{22}+G_2C_2$ have real parts smaller than ω. Define G by

$$G = \begin{pmatrix} 0 \\ G_2 \\ 0 \end{pmatrix}. \tag{4.13}$$

In general, for $G = (\tilde{G}_1^t \; \tilde{G}_2^t \; \tilde{G}_3^t)^t$, we get

$$A + \tilde{G}C = \begin{pmatrix} A_{11}+\tilde{G}_1C_1 & \tilde{G}_1C_2 & 0 \\ \tilde{G}_2C_1 & A_{22}+\tilde{G}_2C_2 & 0 \\ \tilde{G}_3C_1 & A_{32}+\tilde{G}_3C_2 & A_{33} \end{pmatrix}. \tag{4.14}$$

If $\tilde{G} = G$, it follows from our construction, from Lemma 3.1 and from assumption (A8), that the two-by-two left upper block in (4.14) generates a semigroup whose growth constant is smaller than ω. By the general result on bounded perturbation of semigroups (see, for instance, [20, p. 38]), this entails that the same block will also generate a stable semigroup if $\|\tilde{G} - G\|$ is small enough. Since the eigenvalues of A_{33} all have real parts smaller than or equal to ω, this means again by Lemma 3.1, that the semigroup generated by $A + \tilde{G}C$ is stable as well.

LEMMA 4.4. *Consider the system* (3.1) *under the assumptions* (A1)–(A3). *Let* $G : \mathcal{Y} \to \mathcal{X}$ *by a given injection mapping and suppose that there exists* $F : \mathcal{X} \to \mathcal{U}$ *such that the eigenvectors of* $A + BF$ *are complete in* $\mathcal{X}$. *Then, for any* $\eta > 0$, *there exist a finite dimensional subspace* $\mathcal{V} \subset D(A)$ *and a mapping* $\tilde{G}$: $\mathcal{Y} \to \mathcal{X}$ *such that*

$$\|\tilde{G} - G\| < \eta, \tag{4.15}$$

$$(A + BF)x \in \mathcal{V} \quad \textit{for all } x \in \mathcal{V} \tag{4.16}$$

$$\operatorname{im} \tilde{G} \subset \mathcal{V}. \tag{4.17}$$

Proof. Pick some orthonormal basis $\{y_1, \cdots, y_p\}$ of $\mathcal{Y}$ and write $g_i := Gy_i$. Let $\eta > 0$ be given. For every $i = 1, \cdots, p$, there exists a finite set $\{x_{i1}, \cdots, x_{iN(i)}\}$ of generalized eigenvectors of $A + BF$ such that

$$\left\| g_i - \sum_{j=1}^{N(i)} \alpha_{ij} x_{ij} \right\| < \eta \tag{4.18}$$

for suitable constants α_{ij} $(i = 1, \cdots, p; j = 1, \cdots, N(i))$. To every (i, j) there exist a $\lambda_{ij} \in \mathbb{C}$ and an $n_{ij} \in \mathbb{N}$ such that

$$(\lambda_{ij} - (A + BF))^{n_{ij}} x_{ij} = 0. \tag{4.19}$$

Now define $\tilde{G} : \mathcal{Y} \to \mathcal{X}$ by $\tilde{G}y_i = \tilde{g}_i$ $(i = 1, \cdots, p)$, where

$$\tilde{g}_i := \sum_{j=1}^{N(i)} \alpha_{ij} x_{ij}, \tag{4.20}$$

and let $\mathcal{V}$ be the subspace defined by

$$\mathcal{V} := \operatorname{span} \{(\lambda_{ij} - (A + BF))^k x_{ij} \mid i = 1, \cdots, p; j = 1, \cdots, N(i); k = 0, \cdots, n_{ij} - 1\}. \tag{4.21}$$

Then $\tilde{G}$ and $\mathcal{V}$ satisfy the requirements.

LEMMA 4.5. *Consider a pair of operators* (A, B) *and suppose that the assumptions* (A1), (A2), (A4), (A5), (A6) *and* (A9) *hold. Then there exists a bounded linear mapping* $F : \mathcal{X} \to \mathcal{U}$ *such that the spectrum of* $A + BF$ *is discrete, all eigenvalues of* $A + BF$ *have real parts smaller than or equal to* ω *and the eigenvectors of* $A + BF$ *are complete in* $\mathcal{X}$.

Proof. Doing a modal decomposition with respect to the eigenvalues of A in $\{\lambda \in \mathbb{C} | \operatorname{Re} \lambda > \omega\}$, we obtain a direct sum representation $\mathscr{X} = \mathscr{X}_u \oplus \mathscr{X}_s$ and corresponding block representations for A and B:

$$A = \begin{pmatrix} A_u & 0 \\ 0 & A_s \end{pmatrix}, \qquad B = \begin{pmatrix} B_u \\ B_s \end{pmatrix}. \tag{4.22}$$

By (A6), we can choose F_u such that the eigenvalues of $A_u + B_u F_u$ are in $\{\lambda \in \mathbb{C} | \operatorname{Re} \lambda \leqq \omega\}$ and such that they are distinct from the eigenvalues of A_s. Define F by

$$F = (F_u \quad 0). \tag{4.23}$$

Then the spectrum of $A + BF$ will consist of the eigenvalues of A_s together with those of $A_u + B_u F_u$. Because the two sets are separated, there is a corresponding modal decomposition, which we shall indicate by $\mathscr{X} = \mathscr{X}_s \oplus \mathscr{X}_n$ ("n" for "new"). Hence, every vector $x \in \mathscr{X}$ can be written as $x = x_s + x_n$ with $x_s \in \mathscr{X}_s$ and $x_n \in \mathscr{X}_n$. By (A9), x_s can be approximated by linear combinations of eigenvectors of A in $\mathscr{X}_s$, which are, as a consequence of the special form of F, also eigenvectors of $A + BF$. Because $\mathscr{X}_n$ is a finite dimensional $(A + BF)$-invariant subspace, x_n is equal to some linear combination of eigenvectors of $A + BF$. We conclude that x can be approximated by linear combinations of eigenvectors of $A + BF$. Thus, the eigenvectors of $A + BF$ are complete in $\mathscr{X}$.

Proof of Theorem 4.1. Choose G as in Lemma 4.3 and F as in Lemma 4.5. Let $\eta > 0$ be the constant from Lemma 4.3 and use Lemma 4.4 to obtain $\tilde{G} : \mathscr{Y} \to \mathscr{X}$ and $\mathscr{V} \subset D(A)$ satisfying (4.15)–(4.17). Finally, apply Lemma 4.2 to the subspace $\mathscr{V}$ and the mappings F and $\tilde{G}$.

The proof of the theorem is constructive, and therefore it suggests a design method. Depending on the particular type of equation one has at hand, one may vary the actual form of this method in order to avoid unnecessary work. In the next section, we shall illustrate this by an example.

5. Design example. Consider the following system, which is of the "delay" type:

$$x_1'(t) = -\frac{\pi}{2} x_1(t-1) + x_2(t), \qquad x_2'(t) = u(t), \tag{5.1}$$

$$y(t) = x_1(t). \tag{5.2}$$

To write these equations in the standard form (3.1), we use the following set-up (cf. [21]). Let $M_2(-1, 0)$ denote the product space $\mathbb{R} \times L_2(-1, 0)$ and let $H^1(-1, 0)$ be the set of functions on $[-1, 0]$ whose distributional derivative is in $L_2(-1, 0)$ [22, p. 44]. By Sobolev's lemma [22, p. 97], the mappings $\phi \mapsto \phi(-1)$ and $\phi \mapsto \phi(0)$ are well-defined and continuous functions on $H^1(-1, 0)$. For (5.1), the state space will be

$$\mathscr{X} := M_2(-1, 0) \oplus \mathbb{R}. \tag{5.3}$$

The elements of this linear space will be written as column vectors with two components, where the first component is in $M_2(-1, 0)$ and will be written as a row vector (ϕ_0, ϕ) with $\phi_0 \in \mathbb{R}$ and $\phi \in L_2(-1, 0)$ and the second component is in $\mathbb{R}$. The operator A is defined by

$$D(A) := \left\{ \begin{pmatrix} \phi_0, \phi \\ \alpha \end{pmatrix} \middle| \phi_0 \in \mathbb{R}, \phi \in H^1(-1, 0), \alpha \in \mathbb{R}, \phi(0) = \phi_0 \right\}, \tag{5.4}$$

$$A \begin{pmatrix} (\phi_0, \phi) \\ \alpha \end{pmatrix} := \begin{pmatrix} (-\frac{1}{2}\pi\phi(-1) + \alpha, \phi') \\ 0 \end{pmatrix}. \tag{5.5}$$

The input space $\mathcal{U}$ and the output space $\mathcal{Y}$ are both equal to $\mathbb{R}$, and the mappings B and C are given by

$$B\alpha = \begin{pmatrix} (0,0) \\ \alpha \end{pmatrix}, \tag{5.6}$$

$$C\begin{pmatrix} (\phi_0, \phi) \\ \alpha \end{pmatrix} = \phi_0. \tag{5.7}$$

We shall also use the complexifications of these spaces and operators, without change of notation.

It follows from the results of [23] (see also [21]) that the operator A generates a semigroup on $\mathcal{X}$. It is seen immediately that the operators B and C are bounded. The spectrum of A is discrete, and the eigenvalues are precisely the roots of the characteristic equation

$$\det \begin{pmatrix} \lambda + \frac{\pi}{2} e^{-\lambda} & -1 \\ 0 & \lambda \end{pmatrix} = 0 \tag{5.8}$$

[18, Prop. 4.2]. The characteristic function

$$\Delta_A(\lambda) := \lambda\left(\lambda + \frac{\pi}{2} e^{-\lambda}\right) \tag{5.9}$$

has roots at 0, $\pm\pi i/2$, and at infinitely many other points in the complex plane which are given approximately by

$$\lambda_k \cong -\log(4k+1) \pm \frac{\pi}{2}(4k+1)i \qquad (k \in \mathbb{N}). \tag{5.10}$$

Rules for deriving such formulas are given in [30]. All roots are simple. We see that there are only finitely many eigenvalues of A to the right of any vertical line in the complex plane, as is true in general for delay equations [24, p. 114]. The stabilizability of the pair (A, B) and the detectability of the pair (C, A) can be verified conveniently using the generalization of the Hautus test [25], [26] that was given in [3]. Because

$$\operatorname{rank} \begin{pmatrix} \lambda + \frac{\pi}{2} e^{-\lambda} & 1 & 0 \\ 0 & \lambda & 1 \end{pmatrix} = 2 \quad \text{for all } \lambda \in \mathbb{C}, \tag{5.11}$$

the pair (A, B) is stabilizable no matter how the desired growth constant ω is chosen. Likewise, detectability of the pair (C, A) also holds for any ω because

$$\operatorname{rank} \begin{pmatrix} \lambda + \frac{\pi}{2} e^{-\lambda} & 1 \\ 0 & \lambda \\ 1 & 0 \end{pmatrix} = 2 \quad \text{for all } \lambda \in \mathbb{C}. \tag{5.12}$$

Adding a compensator of the form (3.4) to the system (5.1)–(5.2) will lead to a closed-loop system which still has the basic form of a delay equation:

$$x'(t) = A_1 x(t-1) + A_0 x(t). \tag{5.13}$$

Consequently, the closed-loop semigroup will be compact for $t > 1$ ([31]; see also [24]), and this is sufficient to guarantee that its growth constant is determined by the

spectrum of its generator [11, p. 467]. So we can use assumption (A8)′ instead of (A8). Finally, the completeness of the eigenvectors of A follows from [17, Cor. 5.5].

We have verified that all assumptions of § 3 are satisfied for any choice of the desired growth constant ω. Hence, it follows from Thm. 4.1 that any degree of stability can be obtained by adding a finite dimensional compensator to the system (3.1). Let us design such a compensator to obtain a stability margin of 1; so we set $\omega = -1$.

First step. By the stabilizability of the pair (A, B), there exists an F such that the eigenvalues of A at 0 and $\pm\pi i/2$ are shifted to new eigenvalues at -1 and $-1 \pm \pi i/2$ for $A+BF$. If μ is an eigenvalue of $A+BF$, it is easily verified that the corresponding eigenvector is given by

$$\psi = \begin{pmatrix} (\phi_0, \phi) \\ \alpha \end{pmatrix}, \quad \phi(\theta) = e^{\mu\theta}\phi_0 \quad (\theta \in [-1, 0]), \quad \alpha = \left(\mu + \frac{\pi}{2}e^{-\mu}\right)\phi_0. \tag{5.14}$$

The eigenvector will be normalized such that $C\psi = 1$ if we put $\phi_0 = 1$. In that case, we also have

$$F\psi = \mu\left(\mu + \frac{\pi}{2}e^{-\mu}\right) = \Delta_A(\mu). \tag{5.15}$$

Second step. The matrices of A_u and C_u with respect to the basis

$$\left\langle \begin{pmatrix} 1, \cos\frac{\pi}{2}\theta \\ 0 \end{pmatrix}, \begin{pmatrix} 0, \sin\frac{\pi}{2}\theta \\ 0 \end{pmatrix}, \begin{pmatrix} \left(\frac{2}{\pi}, \frac{2}{\pi}\right) \\ 1 \end{pmatrix} \right\rangle \tag{5.16}$$

of $\mathscr{X}_u$ are given by

$$A_u = \begin{pmatrix} 0 & \pi/2 & 0 \\ -\pi/2 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad C_u = \begin{pmatrix} 1 & 0 & \frac{2}{\pi} \end{pmatrix}. \tag{5.17}$$

A straightforward pole placement procedure leads to the conclusion that $A+GC$ will have new eigenvalues at $-(1/2)\pi$ (double) and $-\pi$ if we take

$$G = -\pi \begin{pmatrix} \left(2, \cos\frac{\pi}{2}\theta + 2\sin\frac{\pi}{2}\theta + 1\right) \\ \frac{\pi}{2} \end{pmatrix}. \tag{5.18}$$

Third step. Although it is possible, in principle, to compute η such that $A+\tilde{G}C$ will be stable for each $\tilde{G}$ with $\|\tilde{G}-G\| < \eta$, it does not seem attractive to perform the actual computations and, moreover, the bound we obtain may be unnecessarily conservative. Rather, we shall proceed in an algorithmic way. Let us select

$$\tilde{G} = 2.08 \begin{pmatrix} \left(1, e^{-\theta}\cos\frac{\pi}{2}\theta\right) \\ -1 \end{pmatrix} - 9.08 \begin{pmatrix} \left(0, e^{-\theta}\sin\frac{\pi}{2}\theta\right) \\ \frac{\pi(1-e)}{2} \end{pmatrix} - 8.36 \begin{pmatrix} (1, e^{-\theta}) \\ -1 + \frac{\pi e}{2} \end{pmatrix} \tag{5.19}$$

which is obtained by orthogonally projecting G into the subspace spanned by the eigenvectors of $A+BF$ corresponding to the eigenvalues at -1 and $-1 \pm \pi i/2$. A convenient way to compute the eigenvalues of $A+\tilde{G}C$ is provided by the Weinstein-Aronszajn theory [12, p. 244], from which it follows that these eigenvalues can be

found as the zeros of

$$\Delta_{A+\tilde{G}C}(\lambda) := \Delta_A(\lambda)(1 - C(\lambda - A)^{-1}G). \tag{5.20}$$

If $\tilde{G}$ maps into a subspace spanned by finitely many eigenvectors of $A+BF$, so that

$$G = \sum_{k=1}^{m} \gamma_k \begin{pmatrix} (1, e^{\mu_k \theta}) \\ \mu_k + \frac{\pi}{2} e^{-\mu_k} \end{pmatrix}, \tag{5.21}$$

then we have the more explicit formula

$$\Delta_{A+\tilde{G}C}(\lambda) = \Delta_A(\lambda) - \sum_{k=1}^{m} \gamma_k (\lambda - \mu)^{-1} (\Delta_A(\lambda) - \Delta_A(\mu)). \tag{5.22}$$

Using this, we can employ a simple Newton method to compute the eigenvalues of $A+\tilde{G}C$, where $\tilde{G}$ is given by (5.19). Initial guesses are provided by (5.10) and by the assigned values $-\pi/2$ and $-\pi$. The results are given in Table 1. We see that this trial is easily successful, and so we shall base our design on F, $\tilde{G}$ and the subspace $\mathcal{V}$ spanned by the eigenvectors of $A+BF$ associated with the eigenvalues at -1 and $-1 \pm \pi i/2$.

TABLE 1
Effects of perturbation of G

roots of $\Delta_{A+GC}(\lambda)$	roots of $\Delta_{A+\tilde{G}C}(\lambda)$
-1.571 (double)	$-1.491 \pm 0.288\, i$
-3.142	-3.401
$-1.604 \pm 7.647\, i$	$-1.609 \pm 7.854\, i$
$-2.198 \pm 13.98\, i$	$-2.197 \pm 14.14\, i$
$-2.567 \pm 20.29\, i$	$-2.565 \pm 20.42\, i$
$-2.835 \pm 26.60\, i$	$-2.833 \pm 26.70\, i$
$-3.046 \pm 32.89\, i$	$-3.045 \pm 32.99\, i$
$-3.220 \pm 39.19\, i$	$-3.219 \pm 39.27\, i$
$-3.368 \pm 45.48\, i$	$-3.367 \pm 45.55\, i$
$-3.497 \pm 51.77\, i$	$-3.497 \pm 51.84\, i$
$-3.612 \pm 58.06\, i$	$-3.611 \pm 58.12\, i$

Fourth step. Written in a somewhat sloppy way (with omission of isomorphisms), our compensator is given by

$$w'(t) = (A + BF + \tilde{G}C)w(t) - \tilde{G}y(t), \tag{5.23}$$

$$u(t) = Fw(t), \tag{5.24}$$

where the state space of $w(t)$ is the three-dimensional subspace of $M_2(-1, 0) \oplus \mathbb{R}$ that is spanned by the vectors

$$w_1 = \begin{pmatrix} \left(1, e^{-\theta} \cos \frac{\pi}{2}\theta\right) \\ -1 \end{pmatrix}, \quad w_2 = \begin{pmatrix} \left(0, e^{-\theta} \sin \frac{\pi}{2}\theta\right) \\ \frac{\pi(1-e)}{2} \end{pmatrix}, \quad w_3 = \begin{pmatrix} (1, e^{-\theta}) \\ -1 + \frac{\pi e}{2} \end{pmatrix}. \tag{5.25}$$

The coordinates of $\tilde{G}$ with respect to this basis are given by (5.19). The matrices of

$A+BF$ and C are easily found to be

$$A+BF=\begin{pmatrix}-1 & \frac{1}{2}\pi & 0\\ -\frac{1}{2}\pi & -1 & 0\\ 0 & 0 & -1\end{pmatrix}, \qquad C=(1 \quad 0 \quad 1). \tag{5.26}$$

Finally, we can use (5.15) to calculate $Fw_1=5.24$, $Fw_2=1.13$ and $Fw_3=-3.27$. We finally arrive at the following compensator equations:

$$w'(t)=\begin{pmatrix}1.08 & 1.57 & 2.08\\ -10.65 & -1 & -9.08\\ -8.63 & 0 & -9.36\end{pmatrix}w(t)+\begin{pmatrix}-2.08\\ 9.08\\ 8.36\end{pmatrix}y(t), \tag{5.27}$$

$$u(t)=(5.24 \quad 1.13 \quad -3.27)w(t). \tag{5.28}$$

The eigenvalues of the closed-loop system are given by -1, $-1\pm\pi i/2$, and the eigenvalues of $A+\tilde{G}C$ as listed in Table 1. Consequently, the closed-loop growth constant is exactly equal to -1. The ability to prescribe the location of k closed-loop poles exactly, when a compensator of order k is used, is a particular feature of the method we have used, but one should not get the impression that in general it takes a kth order compensator to stabilize a system with k unstable poles—this is far from being true.

In conclusion, we can say that the computational work needed to obtain the finite-dimensional compensator has been quite moderate: nothing was needed that goes beyond the power of hand-held calculators. Also, note that it has not been necessary to compute the modal projection. The method could be implemented as an iterative procedure, with the third step as the iteration step. The iteration consists of projecting G into a series of trial $(A+BF)$-invariant subspaces of increasing dimension. In this interpretation, Theorem 4.1 can be viewed as a convergence result, guaranteeing that the procedure will terminate after a finite number of steps. Finally, we note that the compensator we obtain is in the standard finite-dimensional form, unlike the compensators obtained from algebraic methods (see, for instance, [27]), which in general contain delay elements.

6. Final remarks. Although we have worked an example to show that the method presented here is in principle feasible as a design procedure, the main emphasis of this paper has been on establishing the existence result on finite-dimensional compensators for a wide class of infinite-dimensional systems. There are many other design considerations, besides the stability margin, that have to be taken into account in any practical situation, such as robustness properties and sensitivity reduction. Fortunately, the method we have employed leaves a great deal of freedom and in particular the selection of the initial F and G is expected to be helpful in obtaining good closed-loop properties. We did not really scrutinize our method to arrive at as low as possible controller orders; here, too, further research promises to be fruitful. The parameterization on the basis of the injection mapping G is particularly suited for situations in which we have few outputs and many inputs; in the reverse situation, one should work with a parametrization on the basis of the feedback mapping F and with subspaces of finite codimension. It has been shown in [28] that ideas very similar to the ones presented here will lead to finite-dimensional compensators that solve tracking and regulation problems for distributed parameter systems. It is, of course, of interest to

extend our results to situations in which we have unbounded control and sensing; results in this direction have been reported recently in [29].

Acknowledgment. The author wants to thank Prof. R. F. Curtain for her help and stimulation in connection with the present paper.

REFERENCES

[1] J. L. Lions, *Optimal Control for Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.
[2] R. Triggiani, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.
[3] M. K. P. Bhat, *Regulator theory for evolution systems*, Ph.D. thesis, Univ. of Toronto, 1976.
[4] N. Fuji, *Feedback stabilization of distributed parameter systems by a functional observer*, this Journal, 18 (1980), pp. 108–121.
[5] R. F. Curtain, *Finite-dimensional compensator design for parabolic distributed systems with point sensors and boundary input*, IEEE Trans. Automat. Contr., AC-27 (1982), pp. 98–104.
[6] M. J. Balas, *Active control of flexible systems*, AIAA Symposium on Dynamics and Control of Large Flexible Spacecraft, Blacksburg, VA, June 13–15, 1977.
[7] ———, *Modal control of certain flexible dynamic systems*, this Journal, 16 (1978), pp. 450–462.
[8] ———, *Feedback control of linear diffusion processes*, Internat. J. Control, 29 (1979), pp. 523–533.
[9] ———, *Reduced-order feedback control of distributed parameter systems via singular perturbation methods*, J. Math. Anal. Appl., 87 (1982), pp. 281–294.
[10] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, John Wiley, New York, 1972.
[11] E. Hille and R. S. Phillips, *Functional Analysis and Semigroups*, AMS Colloquium Publications 31, American Mathematical Society, Providence, RI, 1957.
[12] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1966.
[13] J. Zabczyk, *A note on C_0 semigroups*, Bull. l'Acad. Pol. des Sci., Sér. Math., Astr. et Phys., 23 (1975), pp. 895–898.
[14] F. Trèves, *Basic Linear Partial Differential Equations*, Academic Press, New York, 1975.
[15] S. Mizohata, *The Theory of Partial Differential Equations*, Cambridge Univ. Press, Cambridge, 1973.
[16] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, Van Nostrand, Princeton, NJ, 1965.
[17] A. Manitius, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1–29.
[18] M. C. Delfour and A. Manitius, *The structural operator F and its role in the theory of retarded systems* II, J. Math. Anal. Appl., 74 (1980), pp. 359–381.
[19] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Frederick Ungar, New York, 1972.
[20] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences, 8, Springer-Verlag, Berlin, 1978.
[21] M. C. Delfour, *The largest class of hereditary systems defining a C_0 semigroup on the product space*, Canad. J. Math., 32 (1980), pp. 969–978.
[22] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[23] J. G. Borisovic and A. S. Turbabin, *On the Cauchy problem for linear non-homogenous differential equations with retarded argument*, Soviet Math. Doklady, 10 (1969), pp. 401–405.
[24] J. Hale, *Functional Differential Equations*, Springer-Verlag, New York, 1971.
[25] V. M. Popov, *Hyperstability and optimality of automatic systems with several control functions*, Rev. Roum. Sci. Tech., Sér. Electrotech. Energ., 9 (1964), pp. 629–690.
[26] M. L. J. Hautus, *Controllability and observability conditions of linear autonomous systems*, Proc. Kon. Ned. Akad. Wetensch. Ser. A, 72 (1969), pp. 443–448.
[27] F. M. Callier and C. A. Desoer, *An algebra of transfer functions for distributed linear time invariant systems*, IEEE Trans. Circuits and Systems, CAS-25 (1978), pp. 651–662.
[28] J. M. Schumacher, *Dynamic feedback in finite-and infinite-dimensional linear systems*, Ph.D. thesis, Vrije Univ., Amsterdam, 1981. Also MC Tract 143, Mathematisch Centrum, Amsterdam, 1981.
[29] R. F. Curtain, *Finite dimensional compensators for parabolic distributed systems with unbounded control and observation*, Report TW-234, Univ. of Groningen, 1982.
[30] R. Bellman and K. Cooke, *Differential-Difference Equations*, Academic Press, New York, 1963.
[31] M. C. Delfour and S. K. Mitter, *Hereditary differential systems with constant delays. I. General case*, J. Differential Equations, 12 (1972), pp. 213–235.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0002 $01.25/0

OPTIMIZATION AND CONTROLLABILITY WITHOUT DIFFERENTIABILITY ASSUMPTIONS*

J. WARGA†

Abstract. Let $\mathscr{X}$ and $\mathscr{Y}$ be real normed vector spaces, $K \subset \mathscr{X}$ convex and compact, $C \subset \mathscr{Y}$ a convex body, $\mathscr{U} \subset K$, and $(\phi, \Phi): K \to \mathbb{R}^m \times \mathscr{Y}$ a function that can be appropriately approximated by functions (ϕ_i, Φ_i) whose compositions with linear maps of small finite-dimensional simplices are C^1. We derive sufficient conditions for ϕ to be (locally) controllable on $\mathscr{U}$ subject to the restriction $\Phi(u) \in C$, and obtain, as a corollary, corresponding necessary conditions for a related restricted minimum. These conditions are formulated in terms of directional derivate containers which are a type of set-valued "derivatives" of (ϕ, Φ), and they improve on and extend previously obtained results. They are used elsewhere to obtain new conditions for controllability and restricted minimum in nonsmooth optimal control problems defined by differential or functional-integral equations with isoperimetric and unilateral restrictions and involving either relaxed or original controls.

Key words. nondifferentiable functions, set-valued derivatives, directional derivate containers, Lagrange multipliers, extremals of an optimization problem

1. Introduction. The Lagrange multiplier rule for restricted minimization problems of the calculus can be based on the following proposition: let $V \subset \mathbb{R}^n$, V be open, and $f = (f^1, \cdots, f^m): V \to \mathbb{R}^m$ be C^1 near x_0; then $f(V)$ contains a ball centered at $f(x_0)$ if $f'(x_0)V$ contains a ball centered at $f'(x_0)x_0$. If x_0 minimizes $f^1(x)$ subject to $f^2(x) = \cdots = f^m(x) = 0$ then the above conclusion cannot hold and thus dimension $(f'(x_0)V) < m$; therefore there exists a Lagrange multiplier vector $\lambda = (\lambda^1, \cdots, \lambda^m) \neq 0$ such that $\lambda^T f'(x_0) = 0$.

For C^1 problems of nonlinear programming and for C^1 problems of relaxed optimal control with finitely many scalar restrictions, the arguments can be based on a similar proposition but with V assumed to be a simplex with a vertex at 0. In these cases, the equality $\lambda^T f'(x_0) = 0$ is replaced by the inequality $\lambda^T f'(x_0)(x - x_0) \geqq 0$ for all $x \in V$. In unilateral problems, an additional restriction is introduced of the form $g(x) \in C$, where C is a convex body in a normed vector space $\mathscr{Y}$. (Typically, C may be the set of continuous functions with values in the negative "octant.") Then the corresponding proposition asserts that, for $C^0 \triangleq$ interior of C, the set

$$\{f(x) | x \in V, g(x) \in C^0\} \tag{1}$$

covers a ball centered at $f(x_0)$ if $g(x_0) \in C^0$ and

$$\{f'(x_0)\omega | \omega \in V, g'(x_0)\omega \in C^0 - g(x_0)\} \tag{2}$$

covers a ball centered at 0.

This approach to C^1 problems (that underlies many of the arguments in [8]) is thus seen to be based on propositions P such that "P is valid locally for $(f - f(x_0), g - g(x_0))$ if P is valid for $(f'(x_0), g'(x_0))$." In dealing with optimization problems defined by nondifferentiable functions, we have attempted, starting with [9], to describe the local properties of (f, g) in terms of those that uniformly characterize C^1 functions (f_i, g_i) that converge uniformly to (f, g). Typically, we searched for propositions P such that "P is valid locally for $(f - f(x_0), g - g(x_0))$ if P is valid for $(f_i'(x), g_i'(x))$ uniformly for all x near x_0 and large i." This led to the concepts of a derivate container

* Received by the editors November 25, 1981, and in revised form October 19, 1982. This work was supported in part by the National Science Foundation under grant MCS 8102079.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

[9], [10], [11] (unbounded derivate container [15]) for Lipschitzian (continuous) functions from $\mathbb{R}^n$ to $\mathbb{R}^m$ and of a directional derivate container [13] for functions between normed vector spaces with compact convex domains. All of these concepts replace a nonexistent derivative at x_0 with nested sets of derivatives of approximating functions (or of their restrictions) at "nearby" points. However, both in finite- and infinite-dimensional cases, there appear in the literature many other constructions of set-valued "derivatives" such as those due to Aubin [1] (as quoted in [5]), Clarke [2], Dolecki and Rolewicz [3], Halkin [4], Ioffe [5], Mordukhovich [6], Rockafellar [7], and others, most of them classified, studied or compared in Ioffe's paper [5].

The present paper improves on and extends the results of [13]. Its basic new tool is Theorem 2.3 which states, crudely speaking, that if (f, g) is C^1 and f' Lipschitzian and if all sets defined similarly to (2) with x_0 replaced by arbitrary x near x_0 cover the same ball, then all sets similar to (1) cover a common ball when (f, g) is replaced by arbitrary continuous functions close enough to it.

We apply this tool to an optimization problem defined by (not necessarily differentiable) functions

$$\phi^0 : K \to \mathbb{R}, \quad \phi^1 : K \to \mathbb{R}^{m_1}, \quad \Phi : K \to \mathcal{Y},$$

a set $\mathcal{U} \subset K$, and a convex body $C \subset \mathcal{Y}$, where $m_1 \in \{1, 2, \cdots\}$, $\mathcal{X}$ and $\mathcal{Y}$ are real normed vector spaces, and K is a convex and compact subset of $\mathcal{X}$. We study the related questions of restricted minimization and (local) controllability, both subject to the "unilateral" inclusion $\Phi(q) \in C$; specifically, necessary conditions for a point q_0 to yield the minimum of ϕ^0 on the set

$$\mathcal{A}(\mathcal{U}) \triangleq \{q \in \mathcal{U} \mid \phi^1(q) = 0, \Phi(q) \in C\}$$

respectively on the closure of $\mathcal{A}(\mathcal{U})$; and, given a function $\phi : K \to \mathbb{R}^m$, sufficient conditions for the existence of some $\kappa > 0$ such that

$$S^F(\phi(q_0), \kappa) \subset \{\phi(q) \mid q \in \mathcal{U}, \Phi(q) + S^F(0, \kappa) \subset C\},$$

where $S^F(a, r)$ denotes the closed ball of center a and radius r in the appropriate space.

The optimization problem that we consider is an abstract version of a nonsmooth optimal control problem. Thus K may represent the collection of relaxed (measure-valued) controls, $\mathcal{U}$ some collection of special controls (in particular, some set of ordinary, point-valued, controls), ϕ^0 the cost functional, the relation $\phi^1(q) = 0$ the "isoperimetric" restrictions, and the relation $\Phi(q) \in C$ the unilateral or other functional restrictions. We have studied this problem in [13] but only for the special case where $\mathcal{U} = K$ and with stronger assumptions; and our present results supersede those of [13].

Our results apply to all sets $\mathcal{U} \subset K$ that contain continuous images of every simplex in K that lie arbitrarily close to the simplex. This property (one of two properties that define "abundant" sets of controls [8, § IV.3, pp. 279 ff.]) characterizes in particular the usual sets $\mathcal{U}$ of ordinary controls embedded in the space K of relaxed controls.

Our basic result is the local controllability Theorem 2.2. As corollaries, we derive necessary conditions (based on weaker assumptions than in [13]) for a point q_0 to yield the minimum of ϕ^0 on the set $\mathcal{A}(K)$; and, with the additional assumptions that (ϕ, Φ) is continuous, necessary conditions for q_0 to minimize ϕ^0 on $\mathcal{A}(\mathcal{U})$ respectively on its closure $\overline{\mathcal{A}(\mathcal{U})}$ when $\mathcal{U}$ is a proper subset of K. In particular, we show that, as in the case of smooth problems [8, Thm. V.3.4, p. 314], if q_0 minimizes ϕ^0 on $\overline{\mathcal{A}(\mathcal{U})}$ but not on $\mathcal{A}(K)$ then the problem in abnormal (i.e., there exists an admissible "extremal" point with a vanishing Lagrange coefficient corresponding to ϕ^0). We also

show, subject to some additional assumptions, that if q_0 minimizes ϕ^0 on $\overline{\mathcal{A}(\mathcal{U})}$ then either q_0 yields a local minimum on $\mathcal{A}(K)$ or q_0 itself is an abnormal extremal.

The results of the present paper, together with those of [14], are applied elsewhere [16] to provide necessary conditions for minimum and sufficient conditions for controllability in nonsmooth optimal control problems defined by hereditary functional-integral (and, in particular, differential or functional-differential) equations controlled by ordinary or relaxed controls. When this approach is applied directly in [16] to the special case of nonsmooth problems defined by ordinary differential equations, it enables us to obtain improved results with shorter arguments and much simpler constructions than in [10] and [12, §§ XI.3, XI.4].

2. Assumptions and results. We denote by $h|_A$ or $h|A$ the restriction of a function h to A, by $\mathcal{T}_N$ the simplex

$$\left\{\theta=(\theta^1,\cdots,\theta^N)\in\mathbb{R}^N\middle|\,\theta^j\geqq 0,\ \sum_{j=1}^{N}\theta^j\leqq 1\right\},$$

by $\mathcal{L}(U, V)$ the collection of linear operators from a real vector space U to a real vector space V, and by $C(K,\mathcal{Z})$ the normed vector space of continuous functions from K to a real normed vector space $\mathcal{Z}$ with the sup norm $|\cdot|_{\sup}$. We write $\triangleq$ for "equal by definition," $A^0(\bar{A}, \operatorname{co} A)$ for the interior (closure, convex hull) of a set A, $d[b, A]$ for the distance from a point b to a set A, $S^F(a, r)(S(a, r))$ for the closed (open) ball of center a and radius r, and $S^F(A, r)$ for $\{b|d[b, A]\leqq r\}$. A "convex body" means "a closed convex set with a nonempty interior." We use the concepts of a derivative and of differentiability in the sense of "a derivative relative to a convex set." Thus, let U and V be (real) vector spaces, A a convex subset of U, $a_0\in A$, and $h: A\to V$. We define the (one-sided) *directional derivative* by

$$Dh(a_0; a-a_0)\triangleq\lim_{\alpha\to 0+}\alpha^{-1}[h(a_0+\alpha[a-a_0])-h(a_0)].$$

We say that $Dh(a_0)$ is a *Gâteaux derivative* of h at a_0 (relative to A) if $Dh(a_0)\in\mathcal{L}(U, V)$ and

$$Dh(a_0)(a-a_0)=Dh(a_0; a-a_0)\quad\text{for all } a\in A.$$

If U and V are normed, $h'(a_0)\in\mathcal{L}(U, V)$, $h'(a_0)$ is continuous, and

$$\lim|a-a_0|^{-1}[h(a)-h(a_0)-h'(a_0)(a-a_0)]=0\quad\text{as } a\to a_0,\quad a\in A\sim\{a_0\}$$

then we refer to $h'(a_0)$ as a (Fréchet) derivative of h at a_0 (relative to A). Such a derivative is unique if $A^0\neq\varnothing$ [8, Thm. II.3.1, p. 170].

We denote by $\mathcal{Z}^*$ the topological dual of a normed vector space $\mathcal{Z}$, and identify $(\mathbb{R}^m)^*$ with $\mathbb{R}^m$, writing $l_1 z$ or $l_1(z)$ for the scalar product of $l_1\in\mathbb{R}^m$ with $z\in\mathbb{R}^m$.

DEFINITION 2.1.

Directional derivate container. Let $q_0\in K$, $m\in\{1, 2,\cdots\}$ and $(\phi,\Phi): K\to\mathbb{R}^m\times\mathcal{Y}$. For each choice of $N\in\{1, 2,\cdots\}$, $\delta>0$ and $Q\triangleq(q^1,\cdots,q^N)\in K^N$ we write

$$h^Q(\omega)\triangleq q_0+\sum_{j=1}^{N}\omega^j(q^j-q_0)\qquad(\omega\triangleq(\omega^1,\cdots,\omega^N)\in\mathcal{T}_N),$$

$$\hat{Q}\triangleq h^Q(\mathcal{T}_N)=\operatorname{co}\{q_0, q^1,\cdots,q^N\},\qquad Q_\delta\triangleq h^Q(\delta\mathcal{T}_N).$$

A collection $\{\Lambda^\varepsilon(\phi,\Phi)(q_0)|\varepsilon>0\}$ of nonempty subsets of $\mathcal{L}(\mathcal{X},\mathbb{R}^m\times\mathcal{Y})$, also referred to as $\Lambda^\varepsilon(\phi,\Phi)(q_0)$, is a *directional derivate container for* (ϕ,Φ) *at* q_0 if

(1) $\Lambda^\varepsilon(\phi,\Phi)(q_0)\subset\Lambda^{\varepsilon'}(\phi,\Phi)(q_0)$ $(\varepsilon'>\varepsilon)$,

(2) for some $\varepsilon_0 > 0$, the set $\{M|K \mid M \in \Lambda^{\varepsilon_0}(\phi, \Phi)(q_0)\}$ is a bounded and equicontinuous subset of $C(K, \mathbb{R}^m \times \mathcal{Y})$,

and there exist functions $(\phi_i, \Phi_i): K \to \mathbb{R}^m \times \mathcal{Y}$ $(i = 1, 2, \cdots)$ such that, for every choice of $N \in \{1, 2, \cdots\}$, $Q \triangleq (q^1, \cdots, q^N) \in K^N$ and $\varepsilon > 0$, there exist $\delta(Q) > 0$ and $i^*(Q, \varepsilon) \in \{1, 2, \cdots\}$ such that

(3) the functions $\omega \to (\phi_i, \Phi_i)(h^Q(\omega)): \delta(Q)\mathcal{T}_N \to \mathbb{R}^m \times \mathcal{Y}$ are continuously differentiable,

(4) for every $q' \in Q_{\delta(Q)}$ and $i \geqq i^*(Q, \varepsilon)$ there exists $M \in \Lambda^{\varepsilon}(\phi, \Phi)(q_0)$ satisfying

$$D(\phi_i, \Phi_i)(q'; q - q') = M(q - q') \qquad (q \in \hat{Q}),$$

(5) $\lim_i (\phi_i, \Phi_i)(h^Q(\omega)) = (\phi, \Phi)(h^Q(\omega))$ uniformly for all $\omega \in \delta(Q)\mathcal{T}_N$.

Scalar directional derivate container. Let $\Lambda^{\varepsilon}(\phi, \Phi)(q_0)$ define a directional derivate container for (ϕ, Φ) at q_0. We define a corresponding *scalar directional derivate container* $\mathcal{L}\Lambda(\phi, \Phi, C)(q_0)$ as the collection of all triplets (l_1, l_2, λ) such that $l = (l_1, l_2) \in \mathbb{R}^m \times \mathcal{Y}^*$, $l \neq 0$, $\lambda \in \mathcal{L}(\mathcal{X}, \mathbb{R})$, $\lambda|K$ is continuous, and there exist sequences $l^i = (l_1^i, l_2^i) \in \mathbb{R}^m \times \mathcal{Y}^*$ and (M^i) with $M^i \in \Lambda^{1/i}(\phi, \Phi)(q_0)$ such that l_2 is a cluster point of the l_2^i in the weak* topology of $\mathcal{Y}^*$,

$$\lim l_1^i = l_1, \qquad |l_1^i| + |l_2^i| = 1,$$

$$\lim_i l^i M^i q = \lambda q \qquad (q \in K),$$

$$l_2^i y \leqq 0 \quad \text{if } S^F(y, 1/i) \subset C - \Phi(q_0).$$

Remark 1. The present definition of a directional derivate container is more general than the one in [13, Def. 2.1], every directional derivate container in the sense of [13] being also a directional derivate container as defined above. Since our present results supersede those of [13], there appeared to be no need to coin a new expression.

Remark 2. Conditions 2.1(3) and 2.1(5) imply that if (ϕ, Φ) has a directional derivate container at q_0 then $(\phi, \Phi)|Q_{\delta(Q)}$ is continuous for each Q.

The definition of $\mathcal{L}\Lambda(\phi, \Phi, C)(q_0)$ implies that

$$l_2\Phi(q_0) = \max_{c \in C} l_2 c$$

because every point $c \in C^0$ satisfies the relation

$$S^F(c - \Phi(q_0), 1/i) \subset C - \Phi(q_0)$$

for all sufficiently large i. Furthermore, since K is compact, l^i and $M^i|_K$ equicontinuous and $M^i|_K$ uniformly bounded, we have

$$\lim_i l^i M^i = \lambda \quad \text{uniformly on } K.$$

We can also verify that if (ϕ, Φ) admits a directional derivate container at q_0 then the corresponding scalar directional derivate container $\mathcal{L}\Lambda(\phi, \Phi, C)(q_0)$ is nonempty. Indeed, for $i = 1, 2, \cdots$, let M^i be an arbitrary element of $\Lambda^{1/i}(\phi, \Phi)(q_0)$, $l = (l_1, l_2) = (l_1, 0) \in R^m \times \mathcal{Y}^*$ with $|l_1| = 1$, and $l^i = (l_1^i, l_2^i) = (l_1, 0)$. Then $M^i|_K$ are equicontinuous and uniformly bounded and therefore a subsequence $(l^i M^i)_{i \in J}$ converges uniformly to some continuous $\lambda: K \to \mathbb{R}$ which is linear under convex combinations and can therefore be extended to an element of $\mathcal{L}(\mathcal{X}, \mathbb{R})$. If $I = (j_1, j_2, \cdots)$ and we replace M^i by M^{j_i} then we conclude that $(l_1, l_2, \lambda) \in \mathcal{L}\Lambda(\phi, \Phi, C)(q_0)$.

Remark 3. The definition of the directional derivate container was motivated by the study of nondifferentiable functions (ϕ, Φ) arising in optimal control in terms of appropriate "approximations" with the properties of functions (ϕ_i, Φ_i) of Definition 2.1. For example, such functions (ϕ, Φ) may be of the form

$$\phi(q) = h^1(y(q)(t_1)), \qquad \Phi(q)(t) = h^2(t, y(q)(t)),$$

where $h^1 : \mathbb{R}^n \to \mathbb{R}^m$ and $h^2(t, \cdot) : \mathbb{R}^n \to \mathbb{R}^{m_2}$ are nondifferentiable and $y(q)$ is the solution of a differential or a functional-integral equation such as

$$y(t) = \int_{t_0}^{t} f(\tau, y(\tau), q(\tau))\, d\tau \qquad (t \in [t_0, t_1]),$$

where q is a control function and $f(t, \cdot, q(t))$ nondifferentiable. Then a directional derivate container for (ϕ, Φ) can be constructed (see [16]) in terms of C^1 functions $h_i^1, h_i^2(t, \cdot)$ and $f_i(t, \cdot, q(t))$ approximating $h^1, h^2(t, \cdot)$ and $f(t, \cdot, q)$ and of corresponding finite-dimensional derivate containers. However, it appears difficult to characterize the class of functions (ϕ, Φ) that admit directional derivate containers at q_0 by such simple properties as continuity or Lipschitz continuity. It is easy to show, using, e.g., the arguments of [11] or [16], that Lipschitz continuity is sufficient to ensure the existence of a directional derivate container when K and $\mathcal{Y}$ are finite-dimensional. It remains for us an open question whether Lipschitz continuity is sufficient in the general case. On the other hand, not even simple continuity at q_0 is necessary, even if both K and $\mathcal{Y}$ are finite-dimensional. An example of a function (ϕ, Φ) discontinuous at q_0 which admits a directional derivate container is provided by

$$K = \{(x, y) \in \mathbb{R}^2 \mid 0 \leqq x \leqq y \leqq x^{1/2} \leqq 1\}, \qquad q_0 = (0, 0), \quad q = (x, y),$$

$$\phi = \Phi, \qquad \phi(x, y) = x^{-2} y^4 \quad \text{if } x \neq 0, \quad \phi(0, 0) = 0.$$

(This choice of ϕ was obtained by covering K with the segments

$$\{\alpha(x, x^{1/2}) \in \mathbb{R}^2 \mid 0 \leqq \alpha \leqq 1\} \quad \text{for } 0 \leqq x \leqq 1$$

and setting $\phi(\alpha x, \alpha x^{1/2}) = \alpha^2$ for all x.) We construct a corresponding directional derivate container $\Lambda^e(\phi, \Phi)(q_0)$ in § 3.10.

THEOREM 2.2. *Let $q_0 \in K$, $\mathcal{L}\Lambda(\phi, \Phi, C)(q_0)$ be a scalar directional derivate container for (ϕ, Φ, C) at q_0 and let $\mathcal{U} \subset K$ have the property that for every choice of $N \in \{1, 2, \cdots\}$, $Q \triangleq (q^1, \cdots, q^N) \in K^N$ and $\theta \triangleq (\theta^1, \cdots, \theta^N) \in \mathcal{T}_N$ there exists a sequence $(u_n^Q(\theta))$ in $\mathcal{U}$ such that*

$$\lim_n u_n^Q(\theta) = q_0 + \sum_{j=1}^{N} \theta^j (q^j - q_0) \quad \textit{uniformly for all } \theta \in \mathcal{T}_N$$

and the function $\theta \to u_n^Q(\theta) : \mathcal{T}_N \to K$ is continuous for each $n = 1, 2, \cdots$. Let $|\cdot|$ denote an arbitrary norm on $\mathbb{R}^m$ and $S^F(a, \kappa)$ be accordingly defined if $a \in \mathbb{R}^m$. Then either

(a) *there exists $(l_1, l_2, \lambda) \in \mathcal{L}\Lambda(\phi, \Phi, C)(q_0)$ such that*

$$\lambda q_0 = \min_{q \in K} \lambda q$$

or

(b) *there exist $n, N \in \{1, 2, \cdots\}$, $Q \triangleq (q^1, \cdots, q^N) \in K^N$ and $\kappa, \delta > 0$ such that*

(i) $$S^F(\phi(q_0), \kappa) \subset \{\phi(q) \mid q \in Q_\delta, \Phi(q) + S^F(0, \kappa) \subset C\}$$

and, if (ϕ, Φ) *is continuous,*

(ii) $$S^F(\phi(q_0), \kappa) \subset \{\phi(u_n^Q(\theta)) | \theta \in \delta\mathcal{T}_N, \Phi(u_n^Q(\theta)) + S^F(0, \kappa) \subset C\}$$
$$\subset \{\phi(u) | u \in \mathcal{U}, \Phi(u) + S^F(0, \kappa) \subset C\}.$$

The proof of Theorem 2.2 is largely based on

THEOREM 2.3. *Let* $0 < \gamma, \delta \leqq 1$, $N \in \{1, 2, \cdots\}$, $\mathcal{T} \triangleq \delta\mathcal{T}_N$, *and* $(f, g): \mathcal{T} \to \mathbb{R}^m \times \mathcal{Y}$ *be* C^1 *and such that* f' *is Lipschitzian,* $g(0) \in C$ *and, for each* $\theta \in \mathcal{T}$,

$$S^F(0, \gamma) \subset \{f'(\theta)\omega | \omega \in \mathcal{T}_N, g'(\theta)\omega + S^F(0, \gamma) \subset C - g(0)\}.$$

Then there exists a set $W \subset \mathcal{T}_N$ *such that*

$$g(\theta) + S^F(0, \tfrac{1}{32}\gamma\delta) \subset C \quad (\theta \in W)$$

and $f: W \to S^F(f(0), \frac{1}{8}\gamma\delta)$ *is a homeomorphism. Furthermore, for every continuous* $(F, G): \mathcal{T} \to \mathbb{R}^m \times \mathcal{Y}$ *with*

$$|f - F|_{\sup} \leqq \tfrac{1}{32}\gamma\delta, \qquad |g - G|_{\sup} \leqq \tfrac{1}{64}\gamma\delta,$$

we have

$$S^F(F(0), \tfrac{1}{16}\gamma\delta) \subset \{F(\theta) | \theta \in \mathcal{T}, G(\theta) + S^F(0, \tfrac{1}{64}\gamma\delta) \subset C\}.$$

Theorems 2.4 and 2.5 below are essentially corollaries of Theorem 2.2.

THEOREM 2.4. *Let* q_0 *yield the minimum of* ϕ^0 *on the set* $\mathcal{A}(K) \triangleq \{q \in K | \phi^1(q) = 0, \Phi(q) \in C\}$, *and let* $\mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(q_0)$ *be a scalar directional derivate container for* $((\phi^0, \phi^1), \Phi, C)$ *at* q_0. *Then there exists* $((l_0, l_1), l_2, \lambda) \in \mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(q_0)$ *such that*

$$l_0 \geqq 0, \qquad \lambda q_0 = \min_{q \in K} \lambda q.$$

THEOREM 2.5. *Let* $\mathcal{U} \subset K$ *have the property assumed in Theorem* 2.2, (ϕ, Φ) *be continuous,* $\mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(q_0)$ *be a scalar directional derivate container for* $((\phi^0, \phi^1), \varphi, C)$ *at* q_0, *and assume that* q_0 *minimizes* ϕ^0 *on the closure* $\overline{\mathcal{A}(\mathcal{U})}$ *of the set*

$$\mathcal{A}(\mathcal{U}) \triangleq \{q \in \mathcal{U} | \phi^1(q) = 0, \Phi(q) \in C\}.$$

Then there exists $((l_0, l_1), l_2, \lambda) \in \mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(q_0)$ *such that*

$$l_0 \geqq 0, \qquad \lambda q_0 = \min_{q \in K} \lambda q.$$

Furthermore, if $p_0 \in \mathcal{A}(K)$, $\phi^0(p_0) < \phi^0(q_0)$, *and* $\mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(p_0)$ *is a scalar directional derivate container for* $((\phi^0, \phi^1), \Phi, C)$ *at* p_0 *then there exists* $((\bar{l}_0, \bar{l}_1), \bar{l}_2, \bar{\lambda}) \in \mathcal{L}\Lambda((\phi^0, \phi^1), \Phi, C)(p_0)$ *such that*

$$\bar{l}_0 = 0, \qquad \bar{\lambda} p_0 = \min_{q \in K} \bar{\lambda} q.$$

By analogy with C^1 problems, we shall refer to $(\tilde{q}, \tilde{l}_1, \tilde{l}_2, \tilde{\lambda})$ as *an extremal* (for given ϕ, Φ, C and $\Lambda^e(\phi, \Phi, C)(q)$) if

$$\tilde{q} \in \mathcal{A}(K), \quad (\tilde{l}_1, \tilde{l}_2, \tilde{\lambda}) \in \mathcal{L}\Lambda(\phi, \Phi, C)(\tilde{q}), \quad \tilde{\lambda}\tilde{q} = \min_{q \in K} \tilde{\lambda} q.$$

The extremal is *abnormal* if $\phi = (\phi^0, \phi^1): K \to \mathbb{R} \times \mathbb{R}^{m_1}$, $\tilde{l}_1 = (\tilde{l}_1^0, \tilde{l}_1^1)$ and $\tilde{l}_1^0 = 0$. The point $\tilde{q} \in \mathcal{A}(K)$ is *extremal* (*abnormal*) if there exists $((\tilde{l}_1^0, \tilde{l}_1^1), \tilde{l}_2, \tilde{\lambda})$ such that $(\tilde{q}, (\tilde{l}_1, \tilde{l}_1), \tilde{l}_2, \tilde{\lambda})$ is an extremal (abnormal extremal). Theorem 2.6 below states that the set of extremals respectively of abnormal extremals is, in a certain sense, sequen-

tially compact if the directional derivate containers defined over K have certain upper semicontinuity properties. Theorem 2.7 describes certain directional derivate containers with these upper semicontinuity properties. Finally, Theorem 2.8 shows that, with "upper semicontinuous" directional derivate containers, a point q_0 is abnormal if q_0 minimizes ϕ^0 on $\mathscr{A}(\mathscr{U})$ respectively $\overline{\mathscr{A}(\mathscr{U})}$ but not on $\mathscr{A}(K)$ near q_0.

THEOREM 2.6. *Let $q_0 = \lim_i q_i$ in K and, for each $i = 0, 1, 2, \cdots$, let $\Lambda^\varepsilon(\phi, \Phi)(q_i)$ be a directional derivate container for (ϕ, Φ) at q_i, with the corresponding scalar directional derivate container $\mathscr{L}\Lambda(\phi, \Phi, C)(q_i)$. For $i = 1, 2, \cdots$, let $(l_1^i, l_2^i, \lambda^i) \in \mathscr{L}\Lambda(\phi, \Phi, C)(q_i)$ and*

$$\lambda^i q_i = \min_{q \in K} \lambda^i q.$$

Finally, assume that Φ is continuous and $\mathscr{Y}$ separable and for each $\varepsilon > 0$ there exist $\delta(\varepsilon) > 0$ and $\rho(\varepsilon) > 0$ such that

$$\Lambda^{\delta(\varepsilon)}(\phi, \Phi)(q_i) \subset \Lambda^\varepsilon(\phi, \Phi)(q_0) \quad \text{if } d(q_i, q_0) \leqq \rho(\varepsilon).$$

Then there exist $(l_1, l_2, \lambda) \in \mathscr{L}\Lambda(\phi, \Phi, C)(q_0)$ and $J \subset (1, 2, \cdots)$ such that

$$\lim_{i \in J} l_1^i = l_1, \lim_{i \in J} l_2^i y = l_2 y \quad (y \in \mathscr{Y}), \qquad \lim_{i \in J} \lambda_i q = \lambda q \quad (q \in K)$$

and

$$\lambda q_0 = \min_{q \in K} \lambda q.$$

The following theorem is a direct consequence of Definition 2.1.

THEOREM 2.7. *Let $m \in \{1, 2, \cdots\}$ and $(\phi, \Phi): K \to \mathbb{R}^m \times \mathscr{Y}$. Assume that there exist functions $(\phi_i, \Phi_i): K \to \mathbb{R}^m \times \mathscr{Y}$ $(i = 1, 2, \cdots)$ such that*

(i) $\lim_i (\phi_i, \Phi_i) = (\phi, \Phi)$ *uniformly on K,*

(ii) *for each $\bar{q} \in K$ and $i = 1, 2, \cdots$, the function (ϕ_i, Φ_i) has a Gâteaux derivative $D(\phi_i, \Phi_i)(\bar{q})$ relative to K and $D(\phi_i, \Phi_i)(\bar{q})$ restricted to K are equicontinuous and uniformly bounded,*

(iii) *for each choice of $(\bar{q}, q^1, \cdots, q^N) \in K^{N+1}$ there exists $\delta > 0$ such that the functions*

$$\omega \triangleq (\omega^1, \cdots, \omega^N) \to (\phi_i, \Phi_i)\Big(\bar{q} + \sum_{j=1}^N \omega^j (q^j - \bar{q})\Big) : \delta\mathscr{T}_N \to \mathbb{R}^M \times \mathscr{Y}$$

are continuously differentiable.

Let $\Lambda^\varepsilon(\phi, \Phi)(\bar{q}) \triangleq \{D(\phi_i, \Phi_i)(q') \,|\, |q' - \bar{q}| \leqq \varepsilon, i \geqq 1/\varepsilon\}$. Then, for each $\bar{q} \in K$, $\Lambda^\varepsilon(\phi, \Phi)(\bar{q})$ determines a directional derivate container for (ϕ, Φ) at $\bar{q}$ and

$$\Lambda^\eta(\phi, \Phi)(\bar{q}) \subset \Lambda^\varepsilon(\phi, \Phi)(q_0) \quad \text{if } 0 < \eta \leqq \varepsilon - |\bar{q} - q_0|.$$

THEOREM 2.8. *Let $\mathscr{U} \subset K$ have the property assumed in Theorem 2.2 and (ϕ, Φ) be continuous. Let $\Lambda^\varepsilon(\phi, \Phi)(q)$ be defined as a directional derivate container for (ϕ, Φ) at every $q \in K$ in such a manner that for every $\varepsilon > 0$ there exist $\delta(\varepsilon)$ and $\rho(\varepsilon)$ such that*

$$\Lambda^{\delta(\varepsilon)}(\phi, \Phi)(q) \subset \Lambda^\varepsilon(\phi, \Phi)(\tilde{q}) \quad \text{if } q, \tilde{q} \in K \text{ and } d(q, \tilde{q}) \leqq \rho(\varepsilon).$$

Assume that q_0 minimizes ϕ^0 on $\overline{\mathscr{A}(\mathscr{U})}$. Then either there exists a relatively open subset G of K such that q_0 minimizes ϕ^0 on $\mathscr{A}(K) \cap G$ or q_0 is abnormal.

3. Proofs.

LEMMA 3.1. *Assume that $0 < \beta \leqq \alpha$, $x \in \mathscr{Y}$ and $S^F(x, \alpha) \subset S^F(C, \beta)$. Then $S^F(x, \alpha - \beta) \subset C$.*

Proof. Let $B = S^F(0, 1) \subset \mathcal{Y}$ and let $l \in \mathcal{Y}^*$. Then

$$x + \alpha B = S^F(x, \alpha) \subset S^F(C, \beta) = C + \beta B$$

implies

$$\inf (lC + \beta l B) = \inf lC - \beta |l| \leqq \inf (lx + \alpha l B) = lx - \alpha |l|;$$

hence

$$\inf lC \leqq lx - (\alpha - \beta)|l| = \inf lS^F(x, \alpha - \beta).$$

Since this is true for all $l \in \mathcal{Y}^*$, our conclusion follows from the classical theorem about the separation of convex bodies. Q.E.D.

LEMMA 3.2. *Let $\mathcal{P}$ be a metric space, $\beta_1, \cdots, \beta_1 > 0$, and $S_a \triangleq S(s_\alpha, \beta_\alpha)$ $(\alpha = 1, \cdots, a)$ a covering of $\mathcal{P}$. Then there exist Lipschitzian functions $h_\alpha : \mathcal{P} \to [0, 1]$ such that*

$$\sum_{\alpha=1}^{a} h_\alpha(p) = 1 \quad (p \in \mathcal{P}), \qquad h_\alpha(p) = 0 \quad (p \notin S(s_\alpha, 2\beta_\alpha)).$$

Proof. Let $R_\alpha \triangleq S(s_\alpha, 2\beta_\alpha)$. If $R_i = \mathcal{P}$ for some i, we set $h_i(p) = 1$ $(p \in \mathcal{P})$ and $h_\alpha(p) = 0$ $(\alpha \neq i, p \in \mathcal{P})$. Otherwise, we set

$$g_\alpha(p) = \frac{d[p, \mathcal{P} \sim R_\alpha]}{d[p, \mathcal{P} \sim R_\alpha] + d[p, S_\alpha]} \qquad (\alpha = 1, \cdots, a, p \in \mathcal{P}).$$

Then $g_\alpha(p) = 1$ $(p \in S_\alpha)$, $g_\alpha(p) = 0$ $(p \notin R_\alpha)$ and each g_α is Lipschitzian (because $p \to d[p, A]$ has a Lipschitz constant 1 for all nonempty $A \subset \mathcal{P}$ and the denominator of $g_\alpha(p)$ is at least β_α). We set

$$h_\alpha(p) = g_\alpha(p) \Big/ \sum_{i=1}^{a} g_i(p) \qquad (p \in \mathcal{P})$$

(observing that $\sum_{i=1}^{a} g_i(p) \geqq 1$ for all $p \in \mathcal{P}$). Thus the h_α have the desired properties. Q.E.D.

We shall require the following additional notation and conventions. We endow $\mathbb{R}^m$ with the norm $|\cdot|$ referred to in Theorem 2.2 and define $S^F(a, r)$ accordingly if $a \in \mathbb{R}^m$. We write $|x|_1 \triangleq \sum_{j=1}^{n} |x^j|$ if $x = (x^1, \cdots, x^n) \in \mathbb{R}^n$.

3.3. *Proof of Theorem* 2.3.

Step 1. We shall first assume that $(f, g)(0) = (0, 0)$. Let $A \triangleq \{x \in \mathbb{R}^m \,|\, |x| = 1\}$ and $\mathcal{P} \triangleq \mathcal{T} \times A$. For each $\tilde{p} \triangleq (\tilde{\theta}, \tilde{a}) \in \mathcal{P}$ there exist linearly independent $\xi_1, \cdots, \xi_m \in \mathbb{R}^m$ such that

$$|\xi_i| = \gamma, \qquad \gamma \tilde{a} = 2m^{-1} \sum_{i=1}^{m} \xi_i.$$

Thus there exist $\omega_1, \cdots, \omega_m \in \mathcal{T}_N$ such that

$$f'(\tilde{\theta})\omega_i = \xi_i, \qquad g'(\tilde{\theta})\omega_i + S^F(0, \gamma) \subset C.$$

The matrix $[\xi_1, \cdots, \xi_m]$ with columns $\xi_1, \cdots, \xi_m$ is invertible, and we have

$$f'(\tilde{\theta}) \sum_{i=1}^{m} b^i \omega_i = \sum_{i=1}^{m} b^i \xi_i = \gamma \tilde{a} \text{ for } b^i = 2m^{-1} \qquad (i = 1, \cdots, m).$$

Thus there exists a neighborhood $S(\tilde{p}, 2\tilde{\beta})$ of $\tilde{p}$ in $\mathcal{P}$ such that for all $p \triangleq (\theta, a) \in$

$S^F(\tilde{p}, 2\tilde{\beta})$ the following statements are valid: the matrices $f'(\theta)[\omega_1, \cdots, \omega_m]$ have uniformly bounded inverses;

$$g'(\theta)\omega_i + S^F(0, \gamma/2) \subset C; \tag{1}$$

the equation

$$f'(\theta) \sum_{i=1}^{m} b^i \omega_i = \gamma a$$

has a unique solution $\tilde{b}(p) = (\tilde{b}^1(p), \cdots, \tilde{b}^m(p))$; and

$$m^{-1} \leqq \tilde{b}^i(p) \leqq 4m^{-1}. \tag{2}$$

Furthermore, $p \to \tilde{b}(p) : S(\tilde{p}, 2\tilde{\beta}) \to \mathbb{R}^m$ is Lipschitzian because f' is Lipschitzian and the matrices $[f'(\theta)\omega_1, \cdots, f'(\theta)\omega_m]$ have uniformly bounded inverses.

The compact set $\mathscr{P}$ can be covered by the open sets $S(\tilde{p}, \tilde{\beta})$ and therefore by a finite subcollection $S(p_\alpha, \beta_\alpha)$ $(\alpha = 1, \cdots, \bar{\alpha})$. By Lemma 3.2, there exist Lipschitzian $h_\alpha : \mathscr{P} \to [0, 1]$ such that

$$\sum_{\alpha=1}^{\bar{\alpha}} h_\alpha(p) = 1 \quad (p \in \mathscr{P}), \qquad h_\alpha(p) = 0 \quad (p \notin S(p_\alpha, 2\beta_\alpha)).$$

We denote by $\omega_{i\alpha}$, $b_\alpha(p)$ the vectors $\tilde{\omega}_i$, $\tilde{b}(p)$ corresponding to the neighborhood $S(\tilde{p}, 2\tilde{\beta}) = S(p_\alpha, 2\beta_\alpha)$ and the point $p = (\theta, a)$ in that neighborhood, and set

$$V(\theta, \alpha) \triangleq V(p) \triangleq \gamma^{-1} \Sigma' h_\alpha(p) \sum_{i=1}^{m} b_\alpha^i(p) \omega_{i\alpha},$$

Σ' denoting the sum over all α with $h_\alpha(p) \neq 0$.

The function $p \to V(p)$ is Lipschitzian because each $h_\alpha(\cdot)$ and $b_\alpha(\cdot)$ is Lipschitzian. Furthermore, if $p = (\theta, a) \in S(p_\alpha, 2\beta_\alpha)$ then

$$f'(\theta) \sum_{i=1}^{m} b_\alpha^i(p) \omega_{i,\alpha} = \gamma a;$$

hence

$$f'(\theta) V(\theta, a) = a. \tag{3}$$

Finally, in view of inequalities (2), we have

$$V(p) \in 4\gamma^{-1} \mathscr{T}_N. \tag{4}$$

Step 2. Let $s(x)$ denote the unique point in $\mathscr{T}$ that minimizes the euclidean distance to $x \in \mathbb{R}^N$. For any fixed $a \in A$, we consider the differential equation

$$\frac{du(t)}{dt} = \dot{u}(t) = V(s(u(t)), a), \qquad u(0) = 0.$$

Since $|V(p)|_1 \leqq 4\gamma^{-1}$ for all $p \in \mathscr{P}$ and since $V(\cdot)$ and s$(\cdot)$ are both Lipschitzian, there exists a unique solution $t \to u(t) = u(t; a)$ for all $t \geqq 0$, and the function $(t, a) \to u(t; a) : [0, \infty) \times A \to \mathbb{R}^n$ is continuous. Furthermore, since the components of $V(p)$ are all nonnegative, it follows that, for all $t \in [0, \frac{1}{4}\gamma\delta]$, we have $u(t) \in \mathscr{T} \triangleq \delta \mathscr{T}_N$; hence $s(u(t)) = u(t)$ and, by (3),

$$\frac{d}{dt} f(u(t; a)) = f'(u(t))\dot{u}(t) = f'(u(t)) V(u(t); a) = a.$$

Thus

$$f(u(t;a))=ta \qquad (0\leqq t\leqq \tfrac{1}{4}\gamma\delta,\ a\in A). \tag{5}$$

Let $S_1 \triangleq S^F(0,\frac{1}{4}\gamma\delta)\subset\mathbb{R}^m$. We shall define a mapping $U:S_1\to\mathscr{T}$ by

$$U(x)=u(|x|;x/|x|) \qquad (x\in S_1\sim\{0\}),\quad U(0)=0.$$

Since $u(\cdot;\cdot)$ is continuous, U is continuous at every $x\neq 0$. Furthermore, since $|u(t;a)|_1\leqq 4\gamma^{-1}t$, we have $|U(x)|_1\leqq 4\gamma^{-1}|x|_1$, and thus U is continuous everywhere on S_1. Moreover, U is one-to-one because, by (5), $U(x)=U(\tilde{x})$ implies

$$x=f(U(x))=f(U(\tilde{x}))=\tilde{x}.$$

Thus U is a homeomorphism of S_1 onto $U(S_1)$ with the inverse $f|U(S_1)$.

Step 3. We next study the behavior of g on $U(S_1)$. We first observe that if

$$y_i+S^F(0,\kappa_i)\subset\lambda C, \qquad \phi(t)+S^F(0,\kappa)\subset\lambda C, \qquad \lambda,\kappa_i,\kappa\geqq 0$$

then

$$\sum_{i=1}^{k} y_i+S^F\Big(0,\sum_{i=1}^{k}\kappa_i\Big)\subset k\lambda C, \qquad \int_0^t\phi(\tau)\,d\tau+S^F(0,\kappa t)\subset\lambda tC.$$

Let $a\in A$ be fixed, let $0\leqq t\leqq\frac{1}{4}\gamma\delta$, and let $\bar{h}_\alpha$, $\bar{b}^i_\alpha$ denote h_α, b^i_α evaluated at $(u(t),a)$. We have

$$g'(u(t))\dot{u}(t)=g'(u(t))V(u(t),a)=\gamma^{-1}g'(u(t))\Sigma'\bar{h}_\alpha\sum_{i=1}^{m}\bar{b}^i_\alpha\omega_{i\alpha},$$

and thus it follows from (1) that

$$g'(u(t))\dot{u}(t)+S^F\Big(0,\tfrac{1}{2}\Sigma'\bar{h}_\alpha\sum_{i=1}^{m}\bar{b}^i_\alpha\Big)\subset\gamma^{-1}\Sigma'\bar{h}_\alpha\sum_{i=1}^{m}\bar{b}^i_\alpha C.$$

Since $0\in C$ and $\Sigma'\bar{h}_\alpha=1$, the inequalities (2) imply that

$$g'(u(t))\dot{u}(t)+S^F(0,\tfrac{1}{2})\subset 4\gamma^{-1}C;$$

hence, integrating this inclusion between 0 and $t\leqq\frac{1}{4}\gamma\delta$, we obtain

$$g(u(t))+S^F(0,\tfrac{1}{2}t)\subset 4\gamma^{-1}tC. \tag{6}$$

Step 4. We now drop the assumption that $(f,g)(0)=(0,0)$. Applying relation (6) to $(f,g)-(f,g)(0)$ and $C-g(0)$ yields, for any fixed $a\in A$ and $0\leq t\leq\frac{1}{4}\gamma\delta$,

$$g(u(t))-g(0)+S^F(0,\tfrac{1}{2}t)\subset 4\gamma^{-1}t(C-g(0));$$

hence

$$g(u(t))-[1-4\gamma^{-1}t]g(0)+S^F(0,\tfrac{1}{2}t)\subset 4\gamma^{-1}tC.$$

Let $\alpha_0\in[0,\frac{1}{8}\gamma]$ be such that

$$g(0)+S^F(0,\alpha_0)\subset C.$$

Since $0\leqq 4\gamma^{-1}t\leqq\delta\leqq 1$, we may multiply the above inclusion by $1-4\gamma^{-1}t$ and add to the previous one to obtain

$$g(u(t))+S^F(0,[1-4\gamma^{-1}t]\alpha_0+4\gamma^{-1}t\cdot\tfrac{1}{8}\gamma)\subset 4\gamma^{-1}tC+[1-4\gamma^{-1}t]C=C; \tag{7}$$

hence

$$g(u(t))+S^F(0,\alpha_0)\subset C. \tag{8}$$

Furthermore, relations (4) and (5) applied to $f-f(0)$ yield

$$f(u(t;a))=f(U(ta))=f(0)+ta, \qquad |u(t;a)|_1\leqq 4\gamma^{-1}t. \tag{9}$$

Now let $t_1\triangleq\frac{1}{16}\gamma\delta$, $a_1\triangleq(1,0,\cdots,0)$ and $\theta_1\triangleq u(t_1;a_1)$. Then, by (7) (with $\alpha_0=0$) and (9),

$$\begin{aligned} &f(\theta_1)=f(0)+t_1a_1,\\ &g(\theta_1)+S^F(0,\tfrac{1}{32}\delta\gamma)\subset C,\\ &\theta_1\in\tfrac{1}{4}\delta\mathscr{T}_N. \end{aligned}$$

Next we set

$$(f_1,g_1)(\theta)\triangleq(f,g)(\theta+\theta_1) \qquad (\theta\in\tfrac{3}{4}\delta\mathscr{T}_N)$$

and verify that (f_1,g_1) has all the properties assumed for (f,g) except that its range is restricted to $\frac{3}{4}\delta\mathscr{T}_N$. If we denote by u_1 the function u corresponding to (f_1,g_1) then we have

$$\begin{aligned} &f_1(0)=f(\theta_1)=f(0)+t_1a_1,\\ &g_1(0)+S^F(0,\tfrac{1}{32}\delta\gamma)\subset C,\\ &|u_1(t;a)|_1\leqq 4\gamma^{-1}t. \end{aligned} \tag{10}$$

We apply relations (8) and (9) with f_1, g_1, u_1 replacing f, g, u and with $t=t_1$, $a=-a_1$. This yields, setting $\theta_2\triangleq u_1(t_1;-a_1)$ and $\bar\theta\triangleq\theta_1+\theta_2$ and taking account of (10),

$$\begin{aligned} &f(\bar\theta)=f_1(\theta_2)=f_1(0)-t_1a_1=f(0),\\ &g(\bar\theta)+S^F(0,\tfrac{1}{32}\delta\gamma)\subset C,\\ &\bar\theta\in\tfrac{1}{2}\delta\mathscr{T}_N. \end{aligned} \tag{11}$$

Step 5. We shall henceforth assume, without loss of generality, that $(f,g)(0)=(0,0)$. We consider the function

$$\theta\to(f_2,g_2)(\theta)\triangleq(f,g)(\theta+\bar\theta):\tfrac{1}{2}\delta\mathscr{T}_N\to\mathbb{R}^m\times\mathscr{Y}$$

and the corresponding homeomorphism $x\to U_2(x)$ of $S_2\triangleq S^F(0,\frac{1}{8}\gamma\delta)\subset\mathbb{R}^m$ onto a subset W_1 of $\frac{1}{2}\delta\mathscr{T}_N$ with the inverse $f_2|W_1$. (The radius of S_2 is $\frac{1}{8}\gamma\delta$ instead of $\frac{1}{4}\gamma\delta$ of S_1 because δ is replaced by $\frac{1}{2}\delta$.) Thus $x\to\bar\theta+U_2(x)$ is a homeomorphism of S_2 onto $W\triangleq W_1+\bar\theta\subset\mathscr{T}_N$ with the inverse $f|_W$. By (8) and (11), we have

$$g(\bar\theta+U_2(x))+S^F(0,\tfrac{1}{32}\gamma\delta)=g_2(U_2(x))+S^F(0,\tfrac{1}{32}\gamma\delta)\subset C \qquad (x\in S^F(0,\tfrac{1}{8}\gamma\delta)). \tag{12}$$

This proves the first assertion of the theorem.

Now let $(F,G):\delta\mathscr{T}_N\to\mathbb{R}^m\times\mathscr{Y}$ be continuous and

$$|F-f|_{\sup}\leqq\tfrac{1}{32}\gamma\delta, \qquad |G-g|_{\sup}\leqq\tfrac{1}{64}\gamma\delta.$$

For an arbitrary $z\in\mathbb{R}^m$ with $|z|\leqq\frac{1}{16}\gamma\delta$ we consider the continuous function

$$x\to(f-F)(\bar\theta+U_2(x))+z-(f-F)(0):S^F(0,\tfrac{1}{8}\gamma\delta)\to\mathbb{R}^m$$

which maps $S^F(0,\frac{1}{8}\gamma\delta)$ into itself. By (9), (11) and Brouwer's fixed point theorem,

there exists a fixed point $\bar{x}$ satisfying

$$\begin{aligned}\bar{x} &= f_2(U_2(\bar{x})) - F(\bar{\theta} + U_2(\bar{x})) + z - f(0) + F(0)\\ &= \bar{x} - F(\bar{\theta} + U_2(\bar{x})) + z + F(0);\end{aligned}$$

hence

$$F(\bar{\theta} + U_2(\bar{x})) = F(0) + z.$$

We have $g(\bar{\theta} + U_2(\bar{x})) = g_2(U_2(\bar{x}))$ and therefore, by (12),

$$G(\bar{\theta} + U_2(\bar{x})) + S^F(0, \tfrac{1}{64}\gamma\delta) \subset g(\bar{\theta} + U_2(\bar{x})) + S^F(0, \tfrac{1}{32}\gamma\delta) \subset C.$$

Thus, for any $z \in S^F(0, \frac{1}{16}\gamma\delta) \subset \mathbb{R}^m$ there exists a corresponding $\theta_z (= \bar{\theta} + U_2(\bar{x})) \in \mathcal{T}$ such that

$$F(\theta_z) = F(0) + z, \qquad G(\theta_z) + S^F(0, \tfrac{1}{64}\gamma\delta) \subset C. \quad \text{Q.E.D.}$$

THEOREM 3.4. *Let* $0 \leqq \gamma, \delta \leqq 1$, $N \in \{1, 2, \cdots\}$, $\mathcal{T} \triangleq \delta\mathcal{T}_N$, *and* $(f, g): \mathcal{T} \to \mathbb{R}^m \times \mathcal{Y}$ *be* C^1 *and such that* $g(0) \in C$ *and, for all* $\theta \in \mathcal{T}$,

$$S^F(0, \gamma) \subset \{f'(\theta)\omega \mid \omega \in \mathcal{T}_N,\ g'(\theta)\omega + S^F(0, \gamma) \subset C - g(0)\}.$$

Then for every continuous $(F, G): \mathcal{T} \to \mathbb{R} \times \mathcal{Y}$ *with*

$$|f - F|_{\sup} \leqq \tfrac{1}{32}\gamma\delta, \qquad |g - G|_{\sup} \leqq \tfrac{1}{64}\gamma\delta$$

we have

$$S^F(F(0), \tfrac{1}{16}\gamma\delta) \subset \{F(\theta) \mid \theta \in \mathcal{T},\ G(\theta) + S^F(0, \tfrac{1}{64}\gamma\delta) \subset C\}.$$

Proof. Let $0 < \varepsilon < \frac{1}{32}\gamma\delta$, $\gamma_1 \triangleq \gamma - \varepsilon$, $\delta_1 \triangleq \delta - \varepsilon$ and $(F, G): \mathcal{T} \to \mathbb{R}^m \times \mathcal{Y}$ be continuous and such that

$$|f - F|_{\sup} \leqq \tfrac{1}{32}\gamma_1\delta_1 - \varepsilon, \qquad |g - G|_{\sup} \leqq \tfrac{1}{64}\gamma_1\delta_1.$$

The C^1 function f can be approximated by "mollified" functions

$$\phi_i(x) = \int f(x + y)p_i(y)\, dy \qquad (i = 1, 2, \cdots, x \in \delta_1\mathcal{T}_N),$$

where $f(z)$ is defined as 0 for $z \notin \mathcal{T}$, each p_i is C^2, $p_i(y) = p_i(y^1, \cdots, y^N) = 0$ if either some $y^j \leqq 0$ or $|y| \geqq 1/i$ or $|y|_1 \geqq 1/2i$, $p_i(y) \geqq 0$, $\int p_i(y)\, dy = 1$.

We verify that each ϕ_i is C^2 and therefore each ϕ_i' Lipschitzian on the compact set $\delta_1\mathcal{T}_N$, and that

$$\phi_i'(x) = \int f'(x + y)p_i(y)\, dy$$

and

$$\lim_i \phi_i(x) = f(x), \qquad \lim_i \phi_i'(x) = f'(x) \quad \text{uniformly for all } x \in \delta_1\mathcal{T}_N.$$

We now observe that there exists j such that $|\phi_j(x) - F(x)| \leqq \frac{1}{32}\gamma_1\delta_1$ $(x \in \delta_1\mathcal{T}_N)$ and such that (ϕ_j, g) satisfies the conditions imposed on (f, g) in Theorem 2.3 with γ, δ replaced by γ_1, δ_1. It follows, by Theorem 2.3, that

$$S^F(F(0), \tfrac{1}{16}\gamma_1\delta_1) \subset \{F(\theta) \mid \theta \in \delta_1\mathcal{T}_N,\ G(\theta) + S^F(0, \tfrac{1}{64}\gamma_1\delta_1) \subset C\}.$$

Since (F, G) is continuous on the compact set $\delta\mathcal{T}_N$ and the above relation holds for all $\varepsilon \in (0, \frac{1}{32}\gamma\delta)$, the conclusion of the present theorem follows. Q.E.D.

An essential part of the proof of Theorem 2.2 is contained in Theorems 2.3 and 3.4. The remaining part of the proof that follows is based on the proof of [13, Thm. 2.2, pp. 809–811]. In particular, the first three steps of the proof below are almost an exact copy of the corresponding three steps in [13, pp. 809–811].

3.5. *Proof of Theorem* 2.2.

Step 1. Assume that there exists no (l_1, l_2, λ) as described in the theorem, and observe that the convex body C contains some ball $S^F(y_0, r)$ with $r>0$. We shall first prove that there exists $\beta>0$ such that, for every $(f, F)\in\Lambda^{\beta}(\phi, \Phi)(q_0)$, there exists $\hat{q}\in K$ satisfying

$$f(\hat{q}-q_0)=0, \qquad F(\hat{q}-q_0)+S^F(0,\beta)\subset C-\Phi(q_0). \tag{1}$$

Indeed, assume the contrary, and let ε_0 be as defined in Definition 2.1 (2). Then there exists a sequence $((f_i, F_i))_{i\geqq i_0}$, with $i_0>\max(2/r, 1/\varepsilon_0)$ and $(f_i, F_i)\in\Lambda^{1/i}(\phi, \Phi)(q_0)$, such that for each $i\geqq i_0$ the nonempty closed convex set

$$S_i \triangleq \{0\}\times\{y\in\mathcal{Y} \mid y+S^F(0,1/i)\subset C-\Phi(q_0)\}\subset\mathbb{R}^m\times\mathcal{Y}$$

has no points in common with the nonempty compact convex set $W_i \triangleq (f_i, F_i)(K-q_0)$. Thus there exist $l^i=(l_1^i, l_2^i)\in\mathbb{R}^m\times\mathcal{Y}^*$ such that

$$|l_1^i|+|l_2^i|=1, \qquad l^i w\geqq l^i s \qquad (w\in W_i, s\in S_i). \tag{2}$$

Since the set $\{l_2\in\mathcal{Y}^* \mid |l_2|\leqq 1\}$ is compact in the weak* topology of $\mathcal{Y}^*$, the sequence (l_1^i, l_2^i) has a cluster point $(\bar{l}_1, \bar{l}_2)\in\mathbb{R}^m\times\mathcal{Y}^*$ with respect to that topology. Since $0\in W_i$, relation (2) yields

$$l^i s=l_2^i s_2\leqq 0 \qquad (s=(0,s_2)\in S_i) \tag{3}$$

and, since $i_0>2/r$ and $S^F(y_0, r)\subset C$, this implies that

$$l_2^i(y_0-\Phi(q_0)+z)\leqq 0 \qquad (z\in S^F(0, r/2)).$$

Thus

$$\tfrac{1}{2}r|l_2^i|\leqq l_2^i(\Phi(q_0)-y_0)$$

which, together with the first relation of (2), yields

$$\bar{l}_2(\Phi(q_0)-y_0)\geqq\tfrac{1}{2}r(1-|\bar{l}_1|). \tag{4}$$

This shows that $\bar{l}=(\bar{l}_1, \bar{l}_2)\neq 0$.

Since the collections of functions $\{(f_i, F_i)|_K \mid i\geqq i_0\}$ and $\{l^i \mid i\geqq i_0\}$ are equicontinuous and bounded, there exist $J\subset(1, 2, \cdots)$ and a linear functional $\bar{\lambda}$ on the linear hull of K such that

$$\lim_{i\in J} l^i\circ(f_i, F_i)(q)=\bar{\lambda}(q) \qquad (q\in K) \tag{5}$$

and $\bar{\lambda}|_K$ is continuous. We may arbitrarily extend $\bar{\lambda}$ as a linear functional to the entire vector space $\mathcal{X}$ and we may assume that the sequence $((f_i, F_i))$ was chosen so that $J=(1, 2, \cdots)$. Furthermore, since $\Phi(q_0)\in C$ and $C^0\neq\varnothing$, we can select from each set S_i a point s_i so that $\lim_i |s_i|=0$. It follows then from (2) and (5) that

$$\bar{\lambda}(q-q_0)\geqq 0 \qquad (q\in K).$$

This relation and (3) show that $(\bar{l}_1, \bar{l}_2, \bar{\lambda})$ has all the properties of (l_1, l_2, λ), thus contradicting our first assumption.

Step 2. Let β be as defined in Step 1. If there exists no $\alpha \in (0, \beta]$ such that, for each $(f, F) \in \Lambda^{\alpha}(\phi, \Phi)(q_0)$,

$$S^F(0, \alpha) \subset f(K - q_0) \subset \mathbb{R}^m \tag{6}$$

then there exists a sequence of $(f_i, F_i) \in \Lambda^{1/i}(\phi, \Phi)(q_0)$ such that each convex and compact set $f_i(K - q_0)$ contains a boundary point w_i with $\lim_i w_i = 0$. There exist, therefore, $l_1^i \in \mathbb{R}^m$ such that $|l_1^i| = 1$ and

$$l_1^i f_i(q - q_0) \geqq l_1^i w_i \qquad (q \in K).$$

A simplified version of the argument in Step 1 shows that there exists $(\bar{l}_1, 0, \bar{\lambda}) \in \mathscr{L}\Lambda(\phi, \Phi, C)(q_0)$ such $\bar{\lambda}(q - q_0) \geqq 0$ $(q \in K)$. Thus $(\bar{l}_1, 0, \bar{\lambda})$ has all the properties of (l_1, l_2, λ), again contrary to assumption. Therefore there exists $\alpha \in (0, \beta]$ satisfying relation (6).

Step 3. Since K is compact, there exists $c' \in \mathbb{R}$ such that $|F(q' - q'')| \leqq c'$ for all $(f, F) \in \Lambda^{\alpha}(\phi, \Phi)(q_0)$ and $q', q'' \in K$. Let (f, F) be an arbitrary element of $\Lambda^{\alpha}(\phi, \Phi)(q_0)$ and $\hat{q}$ correspondingly chosen to satisfy relation (1). We set $\beta' \triangleq \frac{1}{2}(c' + \alpha)^{-1}\alpha$ and observe that if $z \in S^F(0, \alpha) \subset \mathbb{R}^m$ then there exists $\bar{q} \in K$ such that $f(\bar{q} - q_0) = z$. We then let $q' \triangleq \beta'\bar{q} + (1 - \beta')\hat{q}$ and observe that

$$f(q' - q_0) = f(\beta'[\bar{q} - q_0] + (1 - \beta')[\hat{q} - q_0]) = \beta' f(\bar{q} - q_0) = \beta' z$$

and, by (1),

$$\begin{aligned} F(q' - q_0) + S^F(0, \alpha/2) &= F(\hat{q} - q_0) + \beta' F(\bar{q} - \hat{q}) + S^F(0, \alpha/2) \\ &\subset F(\hat{q} - q_0) + S^F\left(0, \beta' c' + \frac{\alpha}{2}\right) \\ &\subset F(\hat{q} - q_0) + S^F(0, \alpha) \subset C - \Phi(q_0). \end{aligned}$$

This shows that

$$S^F(0, \beta'\alpha) \subset \{f(q - q_0) | q \in K, F(q - q_0) + S^F(0, \alpha/2) \subset C - \Phi(q_0)\} \tag{7}$$
$$[(f, F) \in \Lambda^{\alpha}(\phi, \Phi)(q_0)].$$

Step 4. Let $\gamma \triangleq \min(\frac{1}{2}\beta'\alpha, \frac{1}{8}\alpha)$. We can determine a finite collection $\{b_1, \cdots, b_n\} \subset S^F(0, \beta'\alpha)$ and a number of $\varepsilon_1 > 0$ such that

$$S^F(0, \gamma) \subset \operatorname{co}\{b_1', \cdots, b_n'\}$$

whenever $|b_j' - b_j| \leqq \varepsilon_1$ $(j = 1, \cdots, n)$. Furthermore, by the definition of the directional derivate container, all the elements of $\Lambda^{\alpha}(\phi, \Phi)(q_0)$ are equicontinuous when restricted to K. It follows therefore from (7) that there exists a finite subset $\{q^1, \cdots, q^N\}$ of the compact set K such that, for every $(f, F) \in \Lambda^{\alpha}(\phi, \Phi)(q_0)$ and every $j \in \{1, \cdots, n\}$, there exists $\hat{q} \in \{q^1, \cdots, q^N\}$ satisfying

$$|f(\hat{q} - q_0) - b_j| \leqq \varepsilon_1, \qquad F(\hat{q} - q_0) + S^F(0, \alpha/4) \subset C - \Phi(q_0).$$

This implies that

$$S^F(0, \gamma) \subset \operatorname{co}\{f(q^j - q_0) | j \in \{1, \cdots, N\}, F(q^j - q_0) + S^F(0, \alpha/4) \subset C - \Phi(q_0)\}$$

and therefore, setting $\hat{Q} \triangleq \operatorname{co}\{q_0, q^1, \cdots, q^N\}$,

$$S^F(0, \gamma) \subset \{f(q - q_0) | q \in \hat{Q}, F(q - q_0) + S^F(0, 2\gamma) \subset C - \Phi(q_0)\} \tag{8}$$
$$[(f, F) \in \Lambda^{\alpha}(\phi, \Phi)(q_0)].$$

Now let $\varepsilon = \alpha$ and $Q \triangleq (q^1, \cdots, q^N)$, and let (ϕ_i, Φ_i), $\delta = \delta(Q)$, $i^* = i^*(Q, \varepsilon)$ and $h^Q(\omega)$ be correspondingly defined as in Definition 2.1. We set

$$(\bar{\phi}_i, \bar{\Phi}_i)(\theta) \triangleq (\phi_i, \Phi_i)(h^Q(\theta)),$$
$$(\bar{\phi}, \bar{\Phi})(\theta) \triangleq (\phi, \Phi)(h^Q(\theta)),$$
$$q' \triangleq h^Q(\theta), \qquad q = h^Q(\omega) \qquad (\theta \in \delta\mathcal{T}_N),$$

and observe that

$$D(\phi_i, \Phi_i)(q'; q - q') = (\bar{\phi}_i, \bar{\Phi}_i)'(\theta)(\omega - \theta).$$

Therefore, by 2.1(4), for each $\theta \in \delta\mathcal{T}_N$ and $i \geqq i^*$ there exists $(\bar{f}, \bar{F}) \in \Lambda^\varepsilon(\phi, \Phi)(q_0)$ satisfying

$$(\bar{\phi}_i, \bar{\Phi}_i)'(\theta)\omega = (\bar{f}, \bar{F})(q - q_0).$$

It follows then from (8) that

$$S^F(0, \gamma) \subset \{\bar{\phi}_i'(\theta)\omega \mid \omega \in \mathcal{T}_N, \bar{\Phi}_i'(\theta)\omega + S^F(0, 2\gamma) \subset C - \Phi(q_0)\}$$
$$\subset \{\bar{\phi}_i'(\theta)\omega \mid \omega \in \mathcal{T}_N, \bar{\Phi}_i'(\theta)\omega + S^F(0, \gamma) \subset C - \bar{\Phi}_i(0)\}$$

for all $\theta \in \delta\mathcal{T}_N$ and all sufficiently large i, say $i \geqq i_1 \geqq i^*$, for which

$$|\bar{\Phi}_i(0) - \Phi(q_0)| = |\Phi_i(q_0) - \Phi(q_0)| \leqq \gamma.$$

Thus each of the functions $(f, g) = (\bar{\phi}_i, \bar{\Phi}_i)$ $(i \geqq i_1)$ satisfies the assumptions of Theorem 3.4 provided we replace C by $C_i \triangleq S^F(C, d_i)$, where

$$d_i \triangleq |\Phi_i(q_0) - \Phi(q_0)| = |\bar{\Phi}_i(0) - \Phi(q_0)|,$$

so that $\bar{\Phi}_i(0) \in C_i$. We choose a positive integer i sufficiently large so that

$$(9) \qquad |\bar{\phi}_i(\theta) - \bar{\phi}(\theta)| \leqq \tfrac{1}{32}\gamma\delta, \qquad |\bar{\Phi}_i(\theta) - \bar{\Phi}(\theta)| \leqq \tfrac{1}{128}\gamma\delta \qquad (\theta \in \delta\mathcal{T}_N).$$

Then Theorem 3.4 and Lemma 3.1 yield, for $(F, G) = (\bar{\phi}, \bar{\Phi})$,

$$S^F(\phi(q_0), \tfrac{1}{16}\gamma\delta) = S^F(\bar{\phi}(0), \tfrac{1}{16}\gamma\delta)$$
$$\subset \{\bar{\phi}(\theta) \mid \theta \in \delta\mathcal{T}_N, \bar{\Phi}(\theta) + S^F(0, \tfrac{1}{64}\gamma\delta) \subset C_i\}$$
$$= \{\phi(q) \mid q \in Q_\delta, \Phi(q) + S^F(0, \tfrac{1}{64}\gamma\delta) \subset S^F(C, d_i)\}$$
$$\subset \{\phi(q) \mid q \in Q_\delta, \Phi(q) + S^F(0, \tfrac{1}{128}\gamma\delta) \subset C\}$$

which shows that relation (i) of alternative (b) of the theorem is valid with $\kappa = \frac{1}{128}\gamma\delta$.

Finally, assume that (ϕ, Φ) is continuous. Let $u_n = u_n^Q$ and let i and n be sufficiently large so that

$$|\bar{\phi}_i(\theta) - \phi(u_n(\theta))| \leqq \tfrac{1}{32}\gamma\delta, \qquad |\bar{\Phi}_i(\theta) - \Phi(u_n(\theta))| \leqq \tfrac{1}{128}\gamma\delta \qquad (\theta \in \delta\mathcal{T}_N).$$

We set $(F, G)(\theta) \triangleq (\phi, \Phi)(u_n(\theta))$ $(\theta \in \delta\mathcal{T}_N)$ and again apply Theorem 3.4 and Lemma 3.1 to obtain the inclusions

$$S^F(\phi(q_0), \tfrac{1}{16}\gamma\delta) \subset \{\phi(u_n(\theta)) \mid \theta \in \delta\mathcal{T}_N, \Phi(u_n(\theta)) + S^F(0, \tfrac{1}{128}\gamma\delta) \subset C\}$$
$$\subset \{\phi(u) \mid u \in \mathcal{U}, \Phi(u) + S^F(0, \tfrac{1}{128}\gamma\delta) \subset C\}.$$

This completes the proof of the theorem. Q.E.D.

3.6. *Proof of Theorem* 2.4. We observe that q_0 minimizes $\phi^0(q)$ on the set

$$\{q \in K \mid \phi^1(q) = 0, \Phi(q) \in C\}$$

if and only if $(q_0, 0)$ minimizes $\phi^0(q)+a$ on the set

$$\{(q, a)\in K\times[0, 1]|\phi^1(q)=0, \Phi(q)\in C\}.$$

We let

$$\hat{\mathscr{X}}\triangleq\mathscr{X}\times\mathbb{R},\quad \hat{K}\triangleq K\times[0, 1],\quad \hat{q}\triangleq(q, a),\quad \hat{q}_0\triangleq(q_0, 0),$$
$$(\hat{\phi}^1, \hat{\Phi})(\hat{q})\triangleq(\phi^1, \Phi)(q),\qquad \hat{\phi}^0(\hat{q})\triangleq\phi^0(q)+a\qquad(\hat{q}\in\hat{K}),$$

and verify that we can define a directional derivate container for $((\hat{\phi}^0, \hat{\phi}^1), \hat{\Phi})$ at $\hat{q}_0$ by

$$\Lambda^e((\hat{\phi}^0, \hat{\phi}^1), \hat{\Phi})(\hat{q}_0)\triangleq\{(h^0, h^1, H)|(h^1, H)(x, \alpha)=(g^1, G)(x),$$
$$h^0(x, \alpha)=g^0(x)+\alpha(x\in\mathscr{X}, \alpha\in\mathbb{R}),$$
$$((g^0, g^1), G)\in\Lambda^e((\phi^0, \phi^1), \Phi)(q_0)\}.$$

We may apply Theorem 2.2 to the problem in which ϕ, Φ, K and q_0 are replaced by $(\hat{\phi}^0, \hat{\phi}^1)$, $\hat{\Phi}$, $\hat{K}$ and $\hat{q}_0$, respectively. Then alternative (b) of Theorem 2.2 is invalid because $\hat{q}_0$ minimizes $\hat{\phi}_0$ on the set

$$\{\hat{q}\in\hat{K}|\hat{\phi}^1(\hat{q})=0, \hat{\Phi}(\hat{q})\in C\}.$$

Therefore there exists $((l_0, l_1), l_2, \lambda)\in\mathscr{L}\Lambda((\phi^0, \phi^1), \Phi, C)(q_0)$ such that

$$\lambda q_0=\min_{q\in K}\lambda q\quad\text{and}\quad 0=\min_{a\in[0,1]} l_0 a\leqq l_0.\qquad\text{Q.E.D.}$$

3.7. *Proof of Theorem* 2.5. The proof of the first part of the theorem is identical with that of Theorem 2.4 except that we also set $\hat{\mathscr{U}}\triangleq\mathscr{U}\times[0, 1]$ and observe that $\hat{q}_0$ cannot satisfy the corresponding relations (ii) of alternative (b) of Theorem 2.2 because it minimizes $\hat{\phi}^0$ on the closure of the set

$$\{\hat{q}\in\hat{\mathscr{U}}|\hat{\phi}^1(\hat{q})=0, \hat{\Phi}(\hat{q})\in C\}.$$

To prove the second part of the theorem, we proceed again as in the proof of Theorem 2.4 except that we set

$$\beta\triangleq\tfrac{1}{2}[\phi^0(q_0)-\phi^0(p_0)],\quad \hat{K}\triangleq K\times[-\beta, \beta],\quad \hat{\mathscr{U}}\triangleq\mathscr{U}\times[-\beta, \beta],\quad \hat{p}_0\triangleq(p_0, 0)$$

and apply Theorem 2.2 to $(\hat{\phi}^0, \hat{\phi}^1)$, $\hat{\Phi}$, C at $\hat{p}_0$. Then relation (ii) of alternative (b) cannot apply, and there exists $((\bar{l}_0, \bar{l}_1), \bar{l}_2, \bar{\lambda})\in\mathscr{L}\Lambda((\phi^0, \phi^1), \Phi, C)(p_0)$ such that

$$\bar{\lambda}p_0=\min_{q\in K}\bar{\lambda}q\quad\text{and}\quad 0=\min_{a\in[-\beta,\beta]}\bar{l}_0 a\leqq\min\{\bar{l}_0\beta, -\bar{l}_0\beta\}.$$

Thus $\bar{l}_0=0$. Q.E.D.

3.8. *Proof of Theorem* 2.6. Since $\mathscr{Y}$ is assumed separable, the weak* topology of $\mathscr{Y}^*$ is metric. Therefore, by the definition of $\mathscr{L}\Lambda(\phi, \Phi, C)(q_i)$, for each i there exist sequences $(l^{i,j})_j=(l_1^{i,j}, l_2^{i,j}))_j$ in $\mathbb{R}^m\times\mathscr{Y}^*$ and $(M^{i,j})_j$ with $M^{i,j}\in\Lambda^{1/j}(\phi, \phi)(q_i)$ such that

$$\lim_j l_1^{i,j}=l_1^i,\quad \lim_j l_2^{i,j}y=l_2^i y\quad(y\in\mathscr{Y}),\quad |l_1^{i,j}|+|l_2^{i,j}|=1,$$

$$\lim_j l^{i,j}M^{i,j}q=\lambda^i q\qquad(q\in K),$$

$$l_2^{i,j}y\leqq 0\text{ if }S^F(y, 1/j)\subset C-\Phi(q_i).$$

For the same reason, there exists [8, Thm. I.3.12, p. 41] a norm $|\cdot|_w$ (weak norm) on $\mathscr{Y}^*$ such that the strong unit ball $\mathscr{U}^*$ in $\mathscr{Y}^*$ is $|\cdot|_w$-sequentially compact and, for

$f, f^i \in \mathcal{U}^*$, we have

$$\lim_i f^i y = fy \quad (y \in \mathcal{Y}) \quad \text{if and only if} \quad \lim_i |f^i - f|_w = 0.$$

Thus for each $i \in \{1, 2, \cdots\}$ we can determine j_1 such that

$$j_1 \geqq 2i, \quad |l_1^{i,j_i} - l_1^i| \leqq \frac{1}{i}, \quad |l_2^{i,j_i} - l_2^i|_w \leqq \frac{1}{i},$$

$$|l^{i,j_i} M^{i,j_i} q - \lambda^i q| \leqq \frac{1}{i} \qquad (q \in K).$$

We set

$$\bar{l}_1^i \triangleq l_1^{i,j_i}, \quad \bar{l}_2^i \triangleq l_2^{i,j_i}, \quad \bar{\lambda}^i \triangleq l^{i,j_i} M^{i,j_i}$$

and observe that $|\bar{l}_1^i| + |\bar{l}_2^i| = 1$ and the functions $\bar{\lambda}^i : K \to \mathbb{R}$ are uniformly bounded and equicontinuous. Thus there exist $J \subset (1, 2, \cdots)$ and $l_1 \in \mathbb{R}^m$, $l_2 \in \mathcal{Y}^*$ and $\lambda : K \to \mathbb{R}$ such that

$$\lim_{i \in J} \bar{l}_1^i = l_1, \quad \lim_{i \in J} |\bar{l}_2^i - l_2|_w = 0, \quad \lim_{i \in J} \bar{\lambda}^i = \lambda \text{ uniformly on } K. \tag{1}$$

Next we observe that

$$\bar{l}_2^i y \leqq 0 \text{ if } S^F(y, 1/j_i) \subset C - \Phi(q_i) \qquad (i \in J). \tag{2}$$

For each $n \in \{1, 2, \cdots\}$ we determine $i_n \in J$ such that

$$|\Phi(q_{i_n}) - \Phi(q_0)| \leqq \frac{1}{2n}, j_{i_n} \geqq 2i_n \geqq 2n, d(q_{i_n}, q_0) \leqq \rho\left(\frac{1}{n}\right), j_{i_n} \geqq 2i_n \geqq 1/\delta(1/n),$$

set

$$\hat{l}_1^n \triangleq \bar{l}_1^{i_n}, \quad \hat{l}_2^n \triangleq \bar{l}_2^{i_n}, \quad \bar{\lambda}^n \triangleq \bar{\lambda}^{i_n},$$

and replace i by i_n in (2). We obtain

$$\hat{l}_2^n y \leqq 0 \quad \text{if } S^F(y + \Phi(q_{i_n}) - \Phi(q_0), 1/j_{i_n}) \subset C - \Phi(q_0)$$

and, a fortiori,

$$\hat{l}_2^n y \leqq 0 \quad \text{if } S^F\left(y, \frac{1}{n}\right) \subset C - \Phi(q_0).$$

Furthermore, we observe that, by (1),

$$\lim_n \hat{\lambda}^n = \lim_n \bar{\lambda}^{i_n} = \lambda \quad \text{uniformly on } K.$$

Finally, we recall that

$$M^{i_n, j_{i_n}} \in \Lambda^{1/j_{i_n}}(\phi, \Phi)(q_{i_n}), \quad j_{i_n} \geqq 1/\delta\left(\frac{1}{n}\right), \quad d(q_{i_n}, q_0) \leqq \rho\left(\frac{1}{n}\right)$$

and therefore, for $\hat{M}^n \triangleq M^{i_n, j_{i_n}}$, we have

$$\hat{M}_n \in \Lambda^{1/n}(\phi, \Phi)(q_0).$$

This shows that the sequences $\hat{l}^n \triangleq (\hat{l}_1^n, \hat{l}_2^n)$ and $(\hat{M}^n)$ correspond to (l_1, l_2, λ) as in the definition of $\mathcal{L}\Lambda(\phi, \Phi, C)(q_0)$. Q.E.D.

3.9. *Proof of Theorem* 2.8. If the first alternative of the theorem (concerning the existence of G) is not valid then there exists a sequence (q_i) in $\mathcal{A}(K)$ converging to q_0 and such that $\phi^0(q_i) < \phi^0(q_0)$. Then, by Theorem 2.5, each q_i is abnormal and therefore, by Theorem 2.6, q_0 is itself abnormal. Q.E.D.

3.10. *A discontinuous function with a directional derivate container.* Let K and $(\phi, \Phi): K \to \mathbb{R}^2$ be defined as in Remark 3 following Definition 2.1. We set

$$\Lambda^\varepsilon(\phi, \Phi)(q_0) = \{(z^T, z^T) | z^T = t(-2\xi^{-3}\eta^4, 4\xi^{-2}\eta^3), |z| \leqq 1, (\xi, \eta) \in K\} \cup \{(0,0)\},$$

where z^T denotes a row vector viewed as an element of $(\mathbb{R}^N)^*$, (z^T, z^T) is the 2×2 matrix with rows z^T, z^T, and superscripts of ξ, η (and below $\bar{\xi}, \bar{\eta}$) denote powers. To show that $\Lambda^\varepsilon(\phi, \Phi)(q_0)$ satisfies Definition 2.1 we set $\phi_i = \Phi_i = \phi$ and observe that, for

$$Q = (q^1, \cdots, q^N), \quad q^j = (x_j, y_j) \neq (0,0), \quad \psi(\omega) = \phi\left(\sum_{j=1}^N \omega^j(x_j, y_j)\right) \qquad (\omega \in \mathcal{T}_N),$$

$$\omega = t\theta, \quad \text{where } t = \sum_{j=1}^N \omega^j, \ \sum_{j=1}^N \theta^j = 1,$$

$$(\bar{\xi}, \bar{\eta}) \triangleq \sum_{j=1}^N \theta^j(x_j, y_j), \quad X \triangleq (x_1, \cdots, x_n), \quad Y \triangleq (y_1, \cdots, y_n),$$

we have

$$\psi'(\omega) = tw^T(X^T, Y^T), \quad \text{where } w^T = (-2\bar{\xi}^{-3}\bar{\eta}^{-4}, 4\bar{\xi}^{-2}\bar{\eta}^{-3}) = \phi'(\bar{\xi}, \bar{\eta})$$

and (X^T, Y^T) is the matrix with rows X^T and Y^T. Since

$$\bar{\xi} \geqq \min x_j > 0, \qquad \bar{\eta} \geqq \min y_j > 0, \qquad 0 < x_j \leqq 1, \quad 0 < y_i \leqq 1,$$

it follows that w^T are bounded for any fixed Q and we can choose $\delta = \delta(Q)$ positive and small enough so that $t|w^T| \leqq 1$ when $0 \leqq t \leqq \delta$ i.e. when $\omega \in \delta\mathcal{T}_N$.

Since $\Lambda^\varepsilon(\phi, \Phi)(q_0)$ is independent of ε and contains elements whose norms are at most 1 and since $\phi_i = \Phi_i = \phi$, conditions 2.1(1), 2.1(2) and 2.1(5) are satisfied. Condition 2.1(3) is satisfied because $\omega \to \psi'(\omega)$ is continuous. If $q' \in \delta(Q)\mathcal{I}_N$ and $q \in K$ then q' is of the form $t(\bar{\xi}, \bar{\eta})$ for $0 \leqq t \leqq \delta(Q)$ and

$$D\phi(q'; q - q) = M(q - q'), \quad \text{where } M = t(-2\bar{\xi}^{-3}\bar{\eta}^4, 4\bar{\xi}^{-2}\bar{\eta}^3).$$

Thus $D(\phi, \Phi)(q'; q - q') = (M, M)(q - q')$, with $(M, M) \in \Lambda^\varepsilon(\phi, \Phi)(q_0)$, and condition 2.1(4) holds.

Acknowledgment. I am indebted to A. D. Ioffe for pointing out that my original assumption of separability of $\mathcal{Y}$ was not needed in the proof of Theorem 2.2.

REFERENCES

[1] J. P. Aubin, *Contingent derivatives of set-valued maps and existence of solutions to differential inclusions*, MRS Technical Summary Report, University of Wisconsin, Madison, 1979.

[2] F. H. Clarke, *Generalized gradients of Lipschitz functionals*, Adv. Math., 40 (1981), pp. 52–67.

[3] S. Dolecki and S. Rolewicz, *Exact penalty for local minima*, this Journal, 17 (1979), pp. 596–606.

[4] H. Halkin, *Mathematical programming without differentiability*, Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976.

[5] A. D. Ioffe, *Nonsmooth analysis: differential calculus of nondifferentiable functions*, Trans. Amer. Math. Soc., 266 (1981), pp. 1–56.

[6] B. S. Mordukhovich, *The maximum principle in the problem of optimal action speed in the non-smooth constraints*, Prikl. Mat. Meh. 40, 6 (1976); J. Appl. Math. Mech., 40 (1976), pp. 960–969.

[7] T. R. ROCKAFELLAR, *Generalized directional derivatives and subgradients of nonconvex functions*, Canad. J. Math., 32 (1980), pp. 257–280.
[8] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[9] ———, *Necessary conditions without differentiability assumptions in optimal control*, J. Differential Equations, 18 (1975), pp. 41–62.
[10] ———, *Controllability and necessary conditions in unilateral problems without differentiability assumptions*, this Journal, 14 (1976), pp. 546–572.
[11] ———, *Derivate containers, inverse functions, and controllability*, Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976.
[12] ———, *Chapter* XI, appended to the Russian translation of Optimal Control of Differential and Functional Equations, Nauka, Moscow, 1977.
[13] ———, *Controllability and a multiplier rule for nondifferentiable optimization problems*, this Journal, 16 (1978), pp. 803–812.
[14] ———, *Controllability of nondifferentiable hereditary processes*, this Journal, 16 (1978), pp. 813–831.
[15] ———, *Fat homeomorphisms and unbounded derivate containers*, J. Math. Anal. Appl., 81 (1981), pp. 545–560.
[16] ———, *Controllability, extremality and abnormality in nonsmooth optimal control*, Cesari Festschrift, J. Optim. Theory and Applic., 41 (1983), to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0003 $01.25/0

LOCAL OPTIMALITY CONDITIONS AND LIPSCHITZIAN SOLUTIONS TO THE HAMILTON-JACOBI EQUATION*

FRANK H. CLARKE† AND RICHARD B. VINTER‡

Abstract. We consider an optimal control problem with end-constraints formulated in terms of a differential inclusion. A sufficient condition for local optimality of a trajectory is given, involving a Lipschitzian function ϕ which is a generalized solution to the Hamilton–Jacobi equation. It is shown that the weakest hypothesis under which the condition is also necessary is that the problem be locally calm. It is further proved that local calmness is implied by strong normality. We thereby establish that the Carathéodory approach, modified to permit Lipschitzian functions ϕ, is applicable in principle when the first order optimality conditions yield nontrivial information.

Key words. calculus of variations and optimal control, necessary and sufficient conditions for optimality, Hamilton–Jacobi equation, nonsmooth analysis

1. Preliminaries.

The definitions of generalized gradient, generalized directional derivative and normal cone used in this paper are those of [2].

Let $\mathcal{O}$ be an open subset of $\mathbb{R}^n$ and suppose that $g(\cdot): \mathcal{O} \to \mathbb{R}^n$ is a locally Lipschitzian function. The generalized gradient of $g(\cdot)$ at a point $x \in \mathcal{O}$ is written $\partial g(x)$, and the generalized directional derivative of $g(\cdot)$ at a point $x \in \mathcal{O}$ in a direction $v \in \mathbb{R}^n$ is written $g^0(x; v)$.

For C a closed subset of $\mathbb{R}^n$, the normal cone to C at a point $x \in C$ is written $N_C(x)$.

Euclidean distance is written $|\cdot|$.

We denote by dist (A_1, A_2) the Hausdorff distance between two sets $A_1, A_2 \subset \mathbb{R}^n$. The modulus $\sup\{|x|: x \in A\}$ of a set $A \subset \mathbb{R}^n$ is written $|A|$, and the Euclidean distance between a point $x \in \mathbb{R}^n$ and a set A is written $d_A(\cdot)$.

The only measure we shall consider is Lebesgue measure. "a.e." signifies "almost everywhere with respect to Lebesgue measure".

2. Introduction.

We study the optimal control problem

$$\text{(P)} \qquad \begin{aligned} &\text{Minimize } f(x(1)) && (2.1)\\ &\text{subject to} \\ &\dot{x}(t) \in F(t, x(t)) \text{ a.e.}, && (2.2)\\ &x(0) = x_0, \quad x(1) \in C_1 && (2.3) \end{aligned}$$

expressed in terms of:

a nonempty subset $\Omega \subset [0, 1] \times \mathbb{R}^n$,

a function $F(\cdot, \cdot)$ with domain Ω which takes as values subsets of $\mathbb{R}^n$,

a point $x_0 \in \mathbb{R}^n$ and a nonempty subset $C_1 \subset \mathbb{R}^n$ and

a function $f(\cdot): \{x: (1, x) \in \Omega\} \to \mathbb{R}$.

An absolutely continuous function $x(\cdot): I \to \mathbb{R}^n$ (where I is a sub-interval of $[0, 1]$) which satisfies (2.2) and has its graph in Ω is called a *trajectory*. If furthermore $I = [0, 1]$ and $x(\cdot)$ satisfies (2.3), it is called an *admissible* trajectory.

A *tube* about some function $x(\cdot): I \to \mathbb{R}^n$ is a set of the form

$$\{(t, y): t \in I, |y - x(t)| < \varepsilon\} \qquad (2.4)$$

* Received by the editors August 2, 1982. This research was performed while the authors were visiting Mathématiques de La Decision, Université de Paris (Dauphine), Paris, France.

† Department of Mathematics, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Y4.

‡ Department of Electrical Engineering, Imperial College, London, England SW7 2BT.

for some $\varepsilon > 0$. More specifically, we refer to (2.4) as the ε-tube about $x(\cdot)$.

A function $x(\cdot)$ is said to be *interior* if some tube about $x(\cdot)$ is contained in Ω.

An admissible trajectory $z(\cdot)$ will be termed *locally optimal* if it achieves the minimum of the functional (2.1) over admissible trajectories $x(\cdot)$ having graphs in some tube about $z(\cdot)$.

Let $z(\cdot)$ be an admissible, interior trajectory. A result customarily associated with the name of Carathéodory, but which has appeared in a variety of guises virtually from the inception of the calculus of variations (see [18, Chapt. 1] or [10, Chapt. 7]), provides a sufficient condition that $z(\cdot)$ be locally optimal, expressed in terms of a solution to the Hamilton–Jacobi equation

$$\phi_t(t, x) + \min_{e \in F(t,x)} \{\phi_x(t, x) \cdot e\} = 0 \tag{2.5}$$

with boundary condition

$$\phi(1, x) = f(x). \tag{2.6}$$

The sufficient condition (which applies under suitable conditions on $F(\cdot, \cdot)$) is that there exist a continuously differentiable function $\phi(\cdot, \cdot)$ which satisfies (2.5) for all (t, x) in the interior of a tube about $z(\cdot)$ and (2.6) for all x in the intersection of C_1 and a neighbourhood of $z(1)$, and is such that

$$\phi(0, x_0) = f(z(1)).$$

The question arises: how widely applicable is the Carathéodory condition? If we consider the condition essentially as stated above then the answer is disappointing. It is easy to construct counter-examples of problems in modern control theory, of a nature such as we would not like to exclude from consideration, whose solutions cannot be characterized in this way. These observations have prompted a number of authors to search for modifications of the condition which are, in principle, more widely applicable. We refer to [1], [9], [14], [15], [16] and [17]. Such a quest is implicit also in [11]. All these papers, in one way or another, deal with the main stumbling block in application of the original condition, namely the requirement that it involve a continuously differential function $\phi(\cdot, \cdot)$ is excessively stringent.

The results presented here are in the tradition of such modifications. We show that the Carathéodory condition can be modified to supply an optimality condition involving a Lipschitzian function $\phi(\cdot, \cdot)$ which is a generalized solution to the Hamilton–Jacobi equation (2.5), and that the modified condition is applicable under a reasonable hypothesis.

As a first step it is necessary to interpret Lipschitzian functions $\phi(\cdot, \cdot)$ which are "generalized solutions" of (2.5). We shall follow Offin [12] and replace (2.5), when $\phi(\cdot, \cdot)$ is Lipschitzian, by

$$\min_{(\alpha,\beta) \in \partial\phi(t,x)} \left\{\alpha + \min_{e \in F(t,x)} \{e \cdot \beta\}\right\} = 0 \tag{2.7}$$

in which $\partial\phi$ denotes the generalized gradient of ϕ. (Notice that, when $\phi(\cdot, \cdot)$ is continuously differentiable, $(\alpha, \beta) \in \partial\phi(t, x)$ implies $(\alpha, \beta) = (\partial\phi/\partial t(t, x), \partial\phi/\partial x(t, x))$ and so, in this case, we recover (2.5) from (2.7)). We mention that generalized gradients and their properties play an important part, not merely in our interpretation of generalized solutions to (2.5), but in the arguments employed throughout this paper.

What is the "reasonable hypothesis" under which the modified Carathéodory condition applies? It is that the first order conditions for local optimality, applied at

the locally optimal trajectory $z(\cdot)$ of interest, are nontrivial in the sense that they must take a form in which the multipler associated with the cost is nonzero. Experience shows that, typically, the first order conditions are nontrivial (in the above sense) or they can be made so by elimination of redundant constraints. Indeed the hypothesis is implicit in many of the algorithms found in the optimal control literature where it is judged sufficient to prove that sequences of trajectories generated by algorithms converge to a trajectory satisfying merely first order conditions. To this extent then our hypothesis is reasonable.

Our results may be seen as confirming a conjecture of Young [18, p. 264] that nontriviality of the first order conditions is the weakest condition under which we can expect to derive sufficient conditions for local optimality.

Actually rather more is proved in this paper than the foregoing remarks suggest. Firstly, as an intermediate result, a weakest hypothesis is identified under which the modified Carathéodory condition applies. This hypothesis is local calmness. Local calmness concerns the stability of the minimum cost under data perturbations. Secondly it is shown that local calmness is implied by nontriviality of the first order optimality conditions.

The intermediate result is interesting in its own right as relating two desirable properties, namely applicability of the modified Carathéodory condition and stability of the minimum cost under data perturbations; the result is also significant because it is easy to deduce from it that, if the modified Carathéodory condition fails to apply, then it can be made to do so by an arbitrarily small perturbation of the endpoint data.

The main difficulty in proving our results is the presence of the terminal constraint $x(1) \in C_1$. When this is absent it is easy to show that local optimality implies existence of a suitable function $\phi(\cdot, \cdot)$; we construct it as the "value function" of the problem (P) (see [12] or, for such results in the setting of differential equations with control term, [8]). In the constrained case (under the local calmness hypothesis) we can still prove the result by construction; $\phi(\cdot, \cdot)$ is now, not the value function for (P) which may be discontinuous, but that of some auxiliary problem involving no terminal constraint.

The present paper appears to be the first which connects sufficient conditions involving Lipschitzian functions $\phi(\cdot, \cdot)$ and local calmness in a differential inclusions setting, and which connects calmness and nontriviality of the first order conditions, for fully nonlinear problems, in any setting.

Our methods can be readily adapted to apply in the framework of differential equations with control term. We thereby obtain a direct proof of results similar to those in [15], where it is also shown that local calmness is the weakest hypothesis under which a modified Carathéodory condition involving Lipschitzian $\phi(\cdot, \cdot)$'s applies, but by use of a very different, nonconstructive approach. By adapting our methods we can also show that nontriviality of the first order conditions, in the form of the Pontryagin maximum principle, implies local calmness in the framework of differential equations with control term. The result however takes a more natural form in the differential inclusions context of this paper, for reasons connected with the "intrinsic" nature of the first order optimality conditions which are appropriate here (see [5, § 5]).

A proof, along somewhat different lines to the self-contained proof presented in this paper of the result "nontriviality of the first order conditions implies that the modified Carathéodory condition applies", will appear in [7] as part of a general development of optimality conditions.

3. The hypotheses. We hypothesize:

(a) $f(\cdot)$ is locally Lipschitzian;

(b) $F(\cdot, \cdot)$ takes as values nonempty compact convex subsets of $\mathbb{R}^n$ and is continuous in the sense that

$$\text{dist}\,(F(t', x'), F(t, x)) \to 0 \text{ if } (t', x') \to (t, x) \text{ in } \Omega;$$

(c) there exists a constant k such that for any t, x, x'

$$\text{dist}\,(F(t, x'), F(t, x)) \leqq k|x' - x| \tag{3.1}$$

whenever the left-hand side is defined;

(d) there exists a constant r such that

$$|F(t, x)| \leqq r \quad \text{for all } (t, x) \in \Omega; \tag{3.2}$$

(e) C_1 is closed.

In fact the values of the functions $f(\cdot)$ and $F(\cdot, \cdot)$ on only some fixed bounded subsets of their domains will be relevant to what follows. We can assume therefore without loss of generality (and this we do for convenience) that Ω is a bounded set. $f(\cdot)$ is thereby rendered a Lipschitzian function taking values in some fixed bounded set.

4. Normality and calmness. We define the Hamiltonian function $H(\cdot, \cdot, \cdot): \Omega \times \mathbb{R}^n \to \mathbb{R}$:

$$H(t, x, p) = \max\{p \cdot v : v \in F(t, x)\}. \tag{4.1}$$

Under our hypotheses $(x, p) \to H(t, x, p)$ is locally Lipschitzian, for each t, on its domain of definition.

In order to motivate the definitions to follow, we state here the conditions of local optimality in [5, Thm. 2] as they bear on problem (P):

THEOREM 4.1. *Let $z(\cdot)$ be an interior, admissible trajectory which is locally optimal. Then there exists an absolutely continuous function $p(\cdot): [0, 1] \to \mathbb{R}^n$ and a number λ equal to 0 or 1, such that*

$$\begin{aligned} &(-\dot{p}(t), \dot{z}(t)) \in \partial H(t, z(t), p(t)) \quad \text{a.e.,} \\ &-p(1) \in N_{C_1}(z(1)) + \lambda\, \partial f(z(1)), \qquad \lambda + |p(1)| \text{ is nonzero,} \end{aligned} \tag{4.2}$$

where ∂H refers to the generalized gradient of $(x, p) \to H(t, x, p)$ for each fixed t.

While the theorem provides merely necessary conditions for local optimality, its significance resides in the fact that, under favourable circumstances, a trajectory $z(\cdot)$ satisfying (4.2) for some λ, $p(\cdot)$ is a likely candidate as a local solution to (P). $z(\cdot)$ is certainly not a likely candidate when λ may be taken as zero, for then all trace of the cost functional $f(\cdot)$ vanishes from (4.2). The condition which excludes this property (the "triviality of the first order conditions" referred to in § 2) is strong normality.

DEFINITION 4.1. Let $z(\cdot)$ be an interior, admissible trajectory which is locally optimal. We say (P) is *strongly normal* at $z(\cdot)$ if the only absolutely continuous function $p(\cdot)$ which satisfies

$$(-\dot{p}(t), \dot{z}(t)) \in \partial H(t, z(t), p(t)), \qquad -p(1) \in N_{C_1}(z(1))$$

is the zero function.

We remark that (P) is automatically strongly normal at $z(\cdot)$ if $z(1) \in \text{int}\,\{C_1\}$.

A weaker condition merely requires that the multipliers $p(\cdot)$, λ can be chosen with $\lambda = 1$. This is "normality".

DEFINITION 4.2. Let $z(\cdot)$ be an interior, admissible trajectory which is locally optimal. We say (P) is *normal* at $z(\cdot)$ if there exists an absolutely continuous function $p(\cdot)$ such that

$$(-\dot{p}(t), \dot{z}(t)) \in \partial H(t, z(t), p(t)), \qquad -p(1) \in N_{C_1}(z(1)) + \partial f(z(1)).$$

Let $z(\cdot):[0, 1] \to \mathbb{R}^n$ be an interior trajectory and ε a positive number. We define the function $q_{z(\cdot),\varepsilon}(\cdot): \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$

$$q_{z(\cdot),\varepsilon}(u) = \inf\{f(x(1))\}, \qquad u \in \mathbb{R}^n,$$

where the infimum is taken over trajectories $x(\cdot):[0, 1] \to \mathbb{R}^n$ which have graphs in the ε-tube about $z(\cdot)$ and which satisfy

$$x(0) = x_0, \qquad x(1) \in C_1 + \{u\}.$$

If no such trajectories exist then the infimum is interpreted as $+\infty$. The definition of local calmness, which we now give, involves the function $q_{z(\cdot),\varepsilon}(\cdot)$.

DEFINITION 4.3. Let $z(\cdot)$ be an interior, admissible trajectory which is locally optimal. We say (P) is *locally calm* at $z(\cdot)$ if there exists positive numbers K and ε such that

$$q_{z(\cdot),\varepsilon}(u) - f(z(1)) \geqq -K|u|, \quad \text{all } u \in \mathbb{R}^n.$$

The significance of such a condition was apparently first noted by Clarke. In earlier literature local calmness (often called simply "calmness") has appeared as a hypothesis assuring normality in a variety of settings (see, e.g., [4]). In the present setting too, if $z(\cdot)$ is an interior, locally optimal trajectory, then local calmness at $z(\cdot)$ implies normality at $z(\cdot)$.

In this paper, the traditional roles of local calmness and normality (in the strengthened form of strong normality) are reversed. Local calmness is the property of primary interest here since, as we shall show, it is the weakest condition under which our modified Carathéodory conditions apply. Strong normality will be presented then as a sufficient condition for local calmness.

5. Main results.

THEOREM 5.1. *Let $z(\cdot)$ be an admissible, interior trajectory.*

(a) *Suppose that there exists some $\delta > 0$ and a Lipschitzian function $\phi(\cdot, \cdot)$ defined on the δ-tube about $z(\cdot)$ such that*

$$\min_{(\alpha,\beta) \in \partial\phi(t, x)} \{\alpha - H(t, x, -\beta)\} = 0 \quad \text{for all } (t, x) \in \{(\tau, \xi): 0 < \tau < 1, |\xi - z(\tau)| < \delta\}, \tag{5.1}$$

$$\phi(1, x) = f(x) \quad \text{for all } x \in \{\xi: |\xi - z(1)| < \delta\} \cap C_1, \tag{5.2}$$

and

$$\phi(0, x_0) = f(z(1)). \tag{5.3}$$

Then $z(\cdot)$ is locally optimal, and (P) *is locally calm at $z(\cdot)$. Conversely:*

(b) *Suppose that $z(\cdot)$ is locally optimal and that* (P) *is locally calm at $z(\cdot)$. Then there exists $\delta > 0$, and a Lipschitzian function $\phi(\cdot, \cdot)$ defined on the δ-tube about $z(\cdot)$ such that* (5.1)–(5.3) *are satisfied.*

As remarked in § 2, (5.1) (which may alternatively be written as (2.7)) reduces to the Hamilton–Jacobi equation (2.5) when $\phi(\cdot, \cdot)$ is a continuously differentiable function. Thus the theorem does indeed provide a modified Carathéodory condition expressed in terms of a Lipschitzian function which is a generalized solution to the

Hamilton–Jacobi equation. It also establishes local calmness as the weakest hypothesis under which such a modified condition applies.

Local calmness may not hold, and then the modified Carathéodory condition fails to apply. However in this case one can show that local calmness can be induced by an arbitrarily small perturbation of the set C_1 (c.f., [15, § 7]).

And now our sufficient condition for local calmness:

THEOREM 5.2. *Let $z(\cdot)$ be an interior trajectory which is locally optimal at $z(\cdot)$ and let* (P) *be strongly normal at $z(\cdot)$. Then* (P) *is locally calm at $z(\cdot)$.*

In light of Theorem 5.2 and remarks in § 4, local calmness (at a locally optimal, interior trajectory) is a condition intermediate in strength between strong normality and normality. It is very easy to construct examples of problems illustrating that local calmness is strictly weaker than strong normality.

For problems (P) in which

(i) $f(\cdot)$ is convex and

(ii) $F(\cdot, \cdot)$ takes the special form

$$F(t, x) = \{A(t)x\} + G(t), \quad \text{all } (t, x)$$

for some continuous $n \times n$-matrix valued function A and some continuous function G taking values compact convex subsets of $\mathbb{R}^n$, it can be shown that local calmness is equivalent to normality. It remains a promising, but as yet unverified, conjecture that local calmness is equivalent to normality for a large class of nonlinear problems.

Recalling that (P) is always strongly normal at $z(\cdot)$ if $z(1) \in \operatorname{int}\{C_1\}$, we deduce from Theorems 5.1 and 5.2 the following local optimality condition for "free endpoint" problems:

COROLLARY 5.1. *Let $z(\cdot)$ be an interior, admissible trajectory and suppose that $z(1) \in \operatorname{int}\{C_1\}$. Then $z(\cdot)$ is locally optimal if and only if there exists some $\delta > 0$ and a Lipschitz continuous function $\phi(\cdot, \cdot)$ defined on the δ-tube about $z(\cdot)$ such that* (5.1)–(5.3) *are satisfied.*

As we have observed in § 1, a simple, direct proof of this last result can be given.

Finally we mention that results somewhat similar to Theorem 5.2, but developed in the context of mathematical programming, are implicit in Robinson's work [13].

6. Proof of Theorem 5.1.

Part (a). Let δ be a positive number and $\phi(\cdot, \cdot)$ a Lipschitzian function which satisfies (5.1), (5.2) and (5.3).

Suppose that $x(\cdot)$ is a trajectory having graph in the $\delta/2$-tube about $z(\cdot)$, and such that $x(0) = x_0$. Consider the function $t \to \phi(t, x(t))$. This is Lipschitzian since it is formed by the composition of Lipschitzian functions. It is therefore certainly absolutely continuous and

$$\phi(1, x(1)) = \phi(0, x_0) + \int_0^1 \frac{d}{dt}\phi(t, x(t))\, dt.$$

By (5.3)

$$\phi(1, x(1)) = f(z(1)) + \int_0^1 \frac{d}{dt}\phi(t, x(t))\, dt. \tag{6.1}$$

Lemma 6.1. $(d/dt)\phi(t, x(t)) \geqq 0$, *a.e.*

Proof. Let $s \in (0, 1)$ be a point with the following properties.

$$t \to \phi(t, x(t)) \text{ is differentiable at } s, \tag{6.2}$$

(6.3) $$\dot{x}(s) \in F(s, x(s)),$$

(6.4) s is a Lebesgue point of $t \to x(t)$.

By (6.2)

$$\frac{d}{dt}\phi(t, x(t))\Big|_{t=s} = \lim_{h \downarrow 0} h^{-1}\left[\phi\left(s+h, x(s) + \int_s^{s+h} \dot{x}(t)\, dt\right) - \phi(s, x(s))\right].$$

Using the Lipschitz continuity of $\phi(\cdot, \cdot)$ and (6.4), we can readily justify writing this limit as

$$\lim_{h \downarrow 0} h^{-1}[\phi(s+h, x(s) + h\dot{x}(s)) - \phi(s, x(s))].$$

This in turn is bounded below by the following expression involving the generalized directional derivative $(-\phi)^0(\cdot; \cdot)$ of $(-\phi)$:

$$\begin{aligned} &-(-\phi)^0((s, x(s)); (1, \dot{x}(s))) \\ &\quad = \max_{(\alpha,\beta) \in \partial(-\phi)(s,x(s))} \{\alpha + \beta\dot{x}(s)\} \\ &\quad = \min_{(\alpha,\beta) \in \partial\phi(s,x(s))} \{\alpha + \beta\dot{x}(s)\} \quad \text{(we have used the fact that } \partial(-\phi) = -\partial\phi\text{)}, \\ &\quad \geqq \min_{(\alpha,\beta) \in \partial\phi(s,x(s))} \min_{e \in F(s,x(s))} \{\alpha + \beta e\} \quad \text{(by (6.3))}, \\ &\quad = \min_{(\alpha,\beta) \in \partial\phi(s,x(s))} \{\alpha - H(s), -\beta\} = 0 \end{aligned}$$

by (5.1). We have shown that $(d/dt)\phi(t, x(t))|_{t=s} \geqq 0$.

Since points s having properties (6.2), (6.3) and (6.4) constitute a subset of full measure, the lemma is proved. □

From (6.1) and Lemma 6.1 we deduce

(6.5) $$\phi(1, x(1)) \geqq f(z(1)).$$

Now let e be a closest point in C_1 to $x(1)$. Since $|x(1) - z(1)| < \delta/2$, the point e is in $\{\xi : |\xi - z(1)| < \delta\} \cap C_1$. By (5.2)

$$f(x(1)) - \phi(1, x(1)) > -K|x(1) - e| = -Kd_{C_1}(x(1)),$$

where K is the Lipschitz constant of the function $x \to f(x) - \phi(1, x)$. By (6.5) then

$$f(x(1)) \geqq f(z(1)) - Kd_{C_1}(x(1)).$$

We conclude from this inequality that $z(\cdot)$ is locally optimal, and (P) is locally calm at $z(\cdot)$.

Part (b). We first state a lemma, an obvious variant of [3, Prop. 2], which provides an estimate of the "distance" of a function from the set of trajectories satisfying a specified boundary condition.

LEMMA 6.2. *There exists a constant k_1 with the following properties: given $\alpha > 0$, a sub-interval $[s_1, s_2] \subset [0, 1]$, $s \in [s_1, s_2]$ and an absolutely continuous function $x(\cdot) : [s_1, s_2] \to \mathbb{R}^n$ such that*

(i) *the α-tube about $x(\cdot)$ is contained in Ω, and*

(ii) $k_1 \Delta_F(x(\cdot)) < \alpha$, *where*

$$\Delta_F(x(\cdot)) = \int_{s_1}^{s_2} d_{F(t,x(t))}(\dot{x}(t))\, dt,$$

then there exists a trajectory $y(\cdot):[s_1, s_2]\to\mathbb{R}^n$ *such that* $y(s)=x(s)$ *and*

$$\int_{s_1}^{s_2}|\dot{y}(t)-\dot{x}(t)|\,dt\leqq k_1\Delta_F(x(\cdot)).$$

We shall denote by T_α the α-tube about $z(\cdot)$. Let ε, $K>0$ be numbers having the properties

$$\varepsilon<2r\ (r \text{ as in (3.2)}),\qquad T_{3\varepsilon}\subset\Omega$$

and

(6.6) $f(z(1))\leqq f(x(1))+Kd_{C_1}(x(1))$ for all trajectories $x(\cdot):[0,1]\to\mathbb{R}^n$ such that $x(0)=x_0$ and $x(\cdot)\in$ closure $\{T_{2\varepsilon}\}$.

Such numbers exist since $z(\cdot)$ is an interior arc and since (P) is locally calm at $z(\cdot)$.

Define the function $g(\cdot,\cdot):\Omega\to\mathbb{R}^n$ by

$$g(t,x)=\max\{|x(t)-z(t)|-\varepsilon,\,0\},\qquad (x,t)\in\Omega.$$

With each point $(s,u)\in\Omega$ we associate a problem $(\mathrm{P}_{s,u})$:

Minimize the functional

$$(\mathrm{P}_{s,u})\qquad \eta_{s,u}(x(\cdot))=n\int_s^1 g(t,x(t))\,dt+Kd_{C_1}(x(1))+f(x(1))$$

over trajectories $x(\cdot):[s,1]\to\mathbb{R}^n$ which satisfy $x(s)=u$.

Here n is some positive number.

If the set of such trajectories is nonempty, we denote the infimum of the values of $\eta_{s,u}(\cdot)$ by $\inf\{\mathrm{P}_{s,u}\}$.

LEMMA 6.3. *Numbers* $\delta\in(0,\varepsilon)$ *and* $n>0$ *can be chosen with the following properties*:

for each $(s,u)\in T_\delta$, $(\mathrm{P}_{s,u})$ *has a solution,*

and all solutions to $(\mathrm{P}_{s,u})$ *have graphs in closure* $\{T_{2\varepsilon}\}$.

Proof. Choose $\gamma>0$. Now select $\delta\in(0,\varepsilon)$ and $n>0$ to satisfy

$$(6.7)\qquad \delta<\frac{\varepsilon}{k_1k+1}\quad\text{and}\quad n>\frac{4r}{\varepsilon^2}(K\varepsilon+2L+\gamma)$$

(here k_1 is the constant of Lemma 6.2, k that of (3.1), r that of (3.2) and L is a bound on the values of $|f(\cdot)|$).

Take $(s,u)\in T_\delta$. Recall that $\delta\in(0,\varepsilon/(k_1k+1))$ and $T_\varepsilon\subset\Omega$. By Lemma 6.2 then, applied to the function $z(\cdot)$ restricted to $[s,1]$ and translated by $u-z(s)$, and (3.1), there exists a trajectory $y(\cdot):[s,1]\to\mathbb{R}^n$ in T_ε satisfying $y(s)=u$. It follows that there exists a minimizing sequence $\{x_1(\cdot)\}$ for $(\mathrm{P}_{s,u})$.

We note in passing that $\inf\{\mathrm{P}_{s,u}\}$ is bounded above by $\eta_{s,u}(y(\cdot))$, whence

$$(6.8)\qquad \inf\{\mathrm{P}_{s,u}\}\leqq Kd_{C_1}(y(1))+f(y(1))\leqq K\varepsilon+L.$$

We shall show presently that for every trajectory $\bar{x}(\cdot):[s,1]\to\mathbb{R}^n$ not in closure $\{T_{2\varepsilon}\}$ and such that $\bar{x}(s)=u$ we have

$$(6.9)\qquad \eta_{s,u}(\bar{x}(\cdot))\geqq\inf\{\mathrm{P}_{s,u}\}+\gamma.$$

This will in effect complete the proof, because (6.9) means that the elements $x_i(\cdot)$ in the minimizing sequence must eventually have graphs in the closed set closure $\{T_{2\varepsilon}\} \subset \Omega$; bearing in mind our hypotheses on the data, notably that $F(\cdot,\cdot)$ has as values compact, convex sets, we deduce, by means of standard compactness arguments, the existence of a cluster point (in an appropriate sense) of $\{x_i(\cdot)\}$ with graph in closure $\{T_{2\varepsilon}\}$, which solves $(P_{s,u})$.

For a trajectory $\bar{x}(\cdot)$ with the stated properties, we must have

$$(1-s) \geqq \frac{\varepsilon}{r}$$

by (3.2), and since $|\bar{x}(s)-z(s)|<\varepsilon$. But then, because $t \to |x(t)-z(t)|$ has Lipschitz constant at most $2r$ and $\varepsilon \in (0, 2r)$, we can easily show that

$$\eta_{s,u}(\bar{x}(\cdot)) = n\int_s^1 g(t, \bar{x}(t))\,dt + Kd_{C_1}(\bar{x}(1)) + f(\bar{x}(1)) \geqq \frac{n\varepsilon^2}{4r} + 0 - L$$

$$\geqq K\varepsilon + L + \gamma \geqq \inf\{P_{s,u}\} + \gamma,$$

by (6.7) and (6.8). We have shown (6.9). □

In view of Lemma 6.3, we may define a function $\phi(\cdot,\cdot): T_\delta \to \mathbb{R}$:

$$\phi(s,u) = \inf\{P_{s,u}\}, \qquad (s,u) \in T_\delta.$$

LEMMA 6.4. (i) $f(z(1)) = \phi(0, x_0)$; *and* (ii) $\phi(1,x) = f(x)$ *for* $x \in \{y : |y - z(1)| < \delta\} \cap C_1$.

Proof. (i) Let $x(\cdot)$ be any solution to (P_{0,x_0}). By Lemma 6.3, $x(\cdot)$ has its graph in closure $\{T_{2\varepsilon}\}$. We have, bearing in mind (6.6),

$$\inf\{P_{0,x_0}\} = n\int_0^1 g(t, x(t))\,dt + Kd_{C_1}(x(1)) + f(x(1))$$

$$\geqq Kd_{C_1}(x(1)) + f(x(1)) \geqq f(z(1)) = \eta_{0,x_0}(z(\cdot)) \geqq \inf\{P_{0,x_0}\}.$$

Hence $f(z(1)) = \inf\{P_{0,x_0}\} = \phi(0, x_0)$.

(ii) This follows directly from the definition of $\phi(\cdot,\cdot)$. □

LEMMA 6.5. $\phi(\cdot,\cdot): T_\delta \to \mathbb{R}^n$ *is Lipschitzian.*

Proof. We show first that

$$|\phi(s,\bar{u}) - \phi(s,u)| \leqq (n + K + c)(1 + kk_1)|\bar{u} - u| \tag{6.10}$$

for $s \in [0,1]$, and points $u, \bar{u}$ in the δ-ball about $z(s)$ such that $|u - \bar{u}| \leqq \varepsilon/(k_1 k + 1)$.

Here c is the Lipschitz constant of $f(\cdot)$.

Consider such s, u and $\bar{u}$. Let $x(\cdot)$ solve $(P_{s,u})$. Recall that $x(\cdot)$ has its graph in closure $\{T_{2\varepsilon}\}$. Since $T_{3\varepsilon} \subset \Omega$ we may apply Lemma 6.2 to $x(\cdot)$ translated by $\bar{u} - u$ and conclude that there exists a trajectory $\bar{x}(\cdot)$ with $\bar{x}(s) = \bar{u}$ such that

$$\max_{t \in [s,1]} |\bar{x}(t) - x(t)| \leqq (k_1 k + 1)|\bar{u} - u|.$$

But then,

$$\phi(s,\bar{u}) \leqq n\int_s^1 g(t, \bar{x}(t))\,dt + Kd_{C_1}(\bar{x}(1)) + f(\bar{x}(1))$$

$$\leqq n\int_s^1 g(t, x(t))\,dt + Kd_{C_1}(x(1)) + f(x(1)) + (n + K + c)(k_1 k + 1)|\bar{u} - u|$$

$$= \phi(s,u) + (n + K + c)(k_1 k + 1)|\bar{u} - u|.$$

Since the roles of u and $\bar{u}$ may be reversed, we deduce (6.10).

Now let (s, u) be an arbitrary point in T_δ.

We next show that

$$|\phi(s,u)-\phi(\bar{s},\bar{u})| \leqq (n+K+c)(kk_1+1)(|u-\bar{u}|+r|s-\bar{s}|) \tag{6.11}$$

for any $(\bar{s}, \bar{u}) \in T_\delta$ sufficiently close to (s, u).

This property implies the Lipschitz continuity of $\phi(\cdot, \cdot)$.

Take $(\bar{s}, \bar{u}) \in T_\delta$. We shall need to consider two possible cases: (i) $\bar{s} \leqq s$ and (ii) $\bar{s} > s$. In the first case, let $\bar{y}(\cdot)$ be a solution to $(P_{\bar{s},\bar{u}})$ and define $w = \bar{y}(s)$. Since $\bar{y}(\cdot)$ has Lipschitz constant at most r,

$$|w-\bar{u}| \leqq r|s-\bar{s}|. \tag{6.12}$$

We may also arrange, by requiring that $(\bar{s}, \bar{u})$ be sufficiently close to (s, u), that the restriction of $\bar{y}(\cdot)$ to $[\bar{s}, s]$ has graph in T_δ. Since the restriction of $\bar{y}(\cdot)$ to [s, 1] solves $(P_{s,w})$ (the "principle of optimality") and $t \to g(t, x(t))$ is zero on $[\bar{s}, s]$,

$$\phi(s, w) = \phi(\bar{s}, \bar{u}). \tag{6.13}$$

Equations (6.10), (6.12), (6.13) and the triangle inequality imply (6.11) (for $(\bar{s}, \bar{u})$ sufficiently close to $(\bar{s}, \bar{u})$).

In the second case, take $y(\cdot)$ to be a solution to $(P_{s,u})$ and define now $w = y(\bar{s})$. Arguing as before, we deduce that, for $(\bar{s}, \bar{u})$ sufficiently close to (s, u),

$$|w-u| \leqq r|s-\bar{s}| \tag{6.14}$$

and

$$\phi(\bar{s}, w) = \phi(s, u). \tag{6.15}$$

Equations (6.10), (6.14) and (6.15) also imply (6.11). In either case then (6.11) is true. □

LEMMA 6.6. *For every* $(t, x) \in \operatorname{int}\{T_\delta\}$

$$\min_{(\alpha,\beta)\in\partial\phi(t,x)} \{\alpha - H(t, x, -\beta)\} = 0.$$

Proof. Since the mapping $(\alpha, \beta, t, x) \to \alpha - H(t, x, -\beta)$ is continuous, and since the set-valued function $(t, x) \to \partial\phi(t, x)$ has closed graph and values contained in some compact set, it suffices to prove:

(i) For all points $(t, x) \in \operatorname{int}\{T_\delta\}$

$$\min_{(\alpha,\beta)\in\partial\phi(t,x)} \{\alpha - H(t, x, -\beta)\} \geqq 0; \tag{6.16}$$

and

(ii) For all points $(\bar{s}, \bar{u})$ in a dense subset of $\operatorname{int}\{T_\delta\}$

$$\min_{(\alpha,\beta)\in\partial\phi(\bar{s},\bar{u})} \{\alpha - H(\bar{s}, \bar{u}, -\beta)\} \leqq 0. \tag{6.17}$$

Consider (i). Let $(s, u) \in \operatorname{int}\{T_\delta\}$ be an arbitrary point at which $\phi(\cdot, \cdot)$ is differentiable, and e an arbitrary point in $F(s, u)$.

Note that (6.16) can alternatively be expressed as

$$\alpha + \beta d \geqq 0 \quad \text{for all } (\alpha, \beta) \in \partial\phi(t, x), \quad d \in F(t, x).$$

In view of the continuity of $F(\cdot, \cdot)$ and the characterization of $\partial\phi(\cdot, \cdot)$ provided in

[2, Prop. 1.11], it follows that (i) will be true if we can show that

$$\phi'(s, u) \cdot (1, e) \geqq 0. \tag{6.18}$$

Here $\phi'(s, u)$ is the gradient of $(t, x) \to \phi(t, x)$ at (s, u). Define $y : [0, s] \to \mathbb{R}^n$ by

$$y(t) = -(s-t)e + u, \qquad t \in [0, s].$$

Since $e \in F(s, u)$, $d_{F(t, y(t))}(\dot{y}(t)) \to 0$ as $t \uparrow s$ and therefore

$$\lim_{\tau \downarrow 0} \tau^{-1} \int_{s-\tau}^{s} d_{F(t, y(t))}(\dot{y}(t))\, dt = 0. \tag{6.19}$$

Now let $\{s_i\}$ be a sequence of numbers such that $s_i \uparrow s$. From (6.19) and Lemma 6.2 we deduce that, for $|s_i - s|$ sufficiently small, there exist trajectories $y_i(\cdot) : [s_i, s] \to \mathbb{R}^n$ in $\operatorname{int}\{T_\delta\}$ such that $y_i(s) = u$ and

$$(s - s_i)^{-1} \int_{s_i}^{s} |\dot{y}_i(t) - e|\, dt \to 0 \quad \text{as } i \to \infty. \tag{6.20}$$

In view of the definition of $\phi(\cdot, \cdot)$ and the fact that $g(\cdot, \cdot)$ is zero on T_δ

$$\phi(s, u) - \phi(s_i, y_i(s_i)) \geqq 0$$

for i sufficiently large.

Now using (6.20) and the facts that $\phi(\cdot, \cdot)$ is differentiable at (s, u) and Lipschitzian, we deduce that

$$\lim (s - s_i)^{-1} \{\phi(s, u) - \phi(s_i, y_i(s_i))$$

exists and

$$\begin{aligned} 0 &\leq \lim_i (s - s_i)^{-1} [\phi(s, u) - \phi(s_i, y_i(s_i))] \\ &= \lim_i (s - s_i) \left[\phi(s, u) - \phi\left(s_i, u - \int_{s_i}^{s} \dot{y}_i(t)\, dt \right) \right] \\ &= \lim_i (s - s_i)^{-1} [\phi(s, u) - \phi(s_i, u - (s - s_i)e)] \\ &= \phi'(s, u) \cdot (1, e). \end{aligned}$$

We have arrived at inequality (6.18).

Consider now (ii). Take any point $(s, u) \in \operatorname{int}\{T_\delta\}$. Let $y(\cdot)$ be a solution to $(P_{s,u})$ and let $\bar{s} > s$ be such that $(\bar{s}, y(\bar{s})) \in \text{interior}\{T_\delta\}$, $\dot{y}(\bar{s}) \in F(s, \bar{y}(s))$, $t \to \phi(t, y(t))$ is differentiable at $\bar{s}$, and $\bar{s}$ is a Lebesgue point of $t \to \dot{y}(t)$. We define $\bar{u} = y(\bar{s})$ and $e = \dot{y}(\bar{s})$.

The point $(\bar{s}, \bar{u})$ can be chosen arbitrarily close to (s, u) in this manner. For $h > 0$ sufficiently small we have from the definition of $\phi(\cdot, \cdot)$ and the fact that $g(\cdot, \cdot)$ is zero on T_δ

$$\phi(\bar{s} + h, y(\bar{s} + h)) - \phi(\bar{s}, \bar{u}) = 0.$$

Our assumptions about the point $\bar{s}$ permit us to write

$$\begin{aligned} 0 &= \lim_{h \downarrow 0} h^{-1} [\phi(\bar{s} + h, y(\bar{s} + h)) - \phi(\bar{s}, \bar{u})] \\ &= \lim_{h \downarrow 0} h^{-1} \left[\phi\left(\bar{s} + h, \bar{u} + \int_{\bar{s}}^{\bar{s}+h} \dot{y}(t)\, dt \right) - \phi(\bar{s}, \bar{u}) \right] \end{aligned}$$

$$= \lim_{h \downarrow 0} h^{-1}[\phi(\bar{s}+h, \bar{u}+he) - \phi(\bar{s}, \bar{u})]$$

$$\geqq -(-\phi)^0((\bar{s}, \bar{u}); (1, e))$$

$$= \max_{(\alpha,\beta)\in\partial(-\phi)(\bar{s},\bar{u})} -\{\alpha+\beta e\} = \min_{(\alpha,\beta)\in\partial\phi(\bar{s},\bar{u})} \{\alpha+\beta e\}$$

$$\geqq \min_{(\alpha,\beta)\in\partial\phi(\bar{s},\bar{u})} \min_{e\in F(\bar{s},\bar{u})} \{\alpha+\beta e\} = \min_{(\alpha,\beta)\in\partial\phi(\bar{s},\bar{u})} \{\alpha - H(\bar{s}, \bar{u}, -\beta)\}.$$

We have proved (6.17). □

Reviewing Lemmas 6.4, 6.5 and 6.6, we see that δ and the Lipschitzian function $\phi(\cdot, \cdot)$ defined on the δ-tube about $z(\cdot)$ satisfy (5.1), (5.2) and (5.3). The proof of part (b) is now complete.

7. Proof of Theorem 5.2. As previously, we denote by T_α the α-tube about $z(\cdot)$. Suppose (P) is not locally calm at $z(\cdot)$. We shall show it is not strongly normal at $z(\cdot)$. Let $\varepsilon > 0$ be such that

$$T_\varepsilon \subset \Omega$$

and such that $z(\cdot)$ is minimizing with respect to all trajectories $x(\cdot)$ with $x(0) = x_0$, $x(1) \in C_1$ and having graph in T_ε. Such an ε exists since $z(\cdot)$ is interior and locally optimal.

Let $\{\alpha_i\}$, $\{\beta_i\}$, $\{\varepsilon_i\}$ be sequences of positive numbers converging to zero. For each i define $g_i(\cdot, \cdot): \Omega \to \mathbb{R}^n$ by

$$g_i(t, x) = \max\{|x(t) - z(t)| - \varepsilon_i, 0\}, \qquad (x, t) \in \Omega.$$

With each i, we associate a problem (P_i):

Minimize the functional

$$(P_i) \qquad \eta_i(x(\cdot)) = \alpha_i \int_0^1 g_i(t, x(t))\, dt + d_{C_1}(x(1)) + \beta_i f(x(1))$$

over trajectories $x(\cdot)$ satisfying $x(0) = x_0$.

If the set of such trajectories is nonempty, we denote the infimum of the values of $\eta_i(\cdot)$ by $\inf\{P_i\}$.

LEMMA 7.1. *We may assume that the sequences $\{\alpha_i\}$, $\{\beta_i\}$ and $\{\varepsilon_i\}$ have been so chosen that, for each i, (P_i) has a solution (this we shall write $x_i(\cdot)$) and*

$$x_i(\cdot) \to z(\cdot) \text{ uniformly.}$$

Proof. Let $\{\gamma_i\}$ be any sequence of positive numbers converging to zero. We now choose the sequences $\{\alpha_i\}$, $\{\beta_i\}$ and $\{\varepsilon_i\}$ which in addition satisfy closure $\{T_{2\varepsilon_i}\} \subset \Omega$, $i = 1, 2, \cdots$

$$(7.1) \qquad \frac{\varepsilon_i}{2r} < 1, \qquad i = 1, 2, \cdots,$$

$$(7.2) \qquad \alpha_i \varepsilon_i^2 > 4r(2\beta_i L + \gamma_i).$$

Here, r is the constant of (3.2) and L is a bound on the values of $|f(\cdot)|$.

Consider i fixed. Let $\{x_{i_j}(\cdot)\}$ be a minimizing sequence for $\{P_i\}$ (such exists, since $z(\cdot)$ is a trajectory with $z(0) = x_0$). Suppose $x_{i_j}(\cdot)$ does not have its graph in closure $\{T_{2\varepsilon}\}$. Then we can easily show, using (7.1) and the fact that $t \to |x_{i_j}(t) - z(t)|$ is

Lipschitzian with constant at most $2r$, that

$$\eta_i(x_{i_j}(\cdot)) \geqq \frac{\alpha_i \varepsilon_i^2}{4r} - \beta_i L.$$

By (7.2), this last number is bounded below by

$$\beta_i L + \gamma_i \geqq \beta_i f(z(1)) + \gamma_i = \eta_i(z(\cdot)) + \gamma_i \geqq \inf\{\mathrm{P}_i\} + \gamma_i.$$

This means that $\{x_{i_j}(\cdot)\}$ must have its graph in the set closure $\{T_{2\varepsilon}\} \subset \Omega$ for i sufficiently large. Again we appeal to standard compactness arguments to deduce that (P_i) has a solution, $x_i(\cdot)$, and

$$\sup_{t\in[0,1]} |x_i(t) - z(t)| \leqq 2\varepsilon_i. \qquad \square$$

LEMMA 7.2. *$x_i(1) \notin C_1$ for all i sufficiently large.*

Proof. Fix i. Since (P) is not locally calm, there exists a trajectory $\bar{x}_i(\cdot)$ having graph in T_{ε_i} such that

$$d_{C_1}(\bar{x}_i(1)) + \beta_i f(\bar{x}_i(1)) < \beta_i f(z(1)). \tag{7.3}$$

Suppose in contradiction $x_i(1) \in C_1$. Then, since $x_i(\cdot)$ has its graph in T_ε for i sufficiently large (see Lemma 7.1), it follows from the local optimality of $z(\cdot)$ that

$$f(z(1)) \leqq f(x_i(1)) \tag{7.4}$$

for i sufficiently large. But then, by (7.3) and (7.4),

$$\begin{aligned}\inf\{\mathrm{P}_i\} &= \alpha_i \int_0^1 g_i(t, x_i(t))\, dt + d_{C_1}(x_i(1)) + \beta_i f(x_i(1)) \\ &\geqq 0 + 0 + \beta_i f(z(1)) > d_{C_1}(\bar{x}_i(1)) + \beta_i f(\bar{x}_i(1)) \geqq \inf\{\mathrm{P}_i\}.\end{aligned}$$

From this contradiction we deduce that $x_i(1) \notin C_1$. $\square$

Next we note:

LEMMA 7.3. *There exists a positive number ν and, for each i, both an absolutely continuous function $p_i(\cdot) : [0, 1] \to \mathbb{R}^n$ and a vector $q_i \in \mathbb{R}^n$ with the properties:*

$$(-\dot{p}_i(t), \dot{x}_i(t)) \in \alpha_i \partial g_i(t, x_i(t)) \times \{0\} + \partial H(t, x_i(t), p_i(t)), \tag{7.5}$$

$$-p_i(1) = \{q_i\} + \beta_i \partial f(x_i(1)), \qquad q_i \in \nu d_{C_1}(e_i), \quad |q_i| = 1, \tag{7.6}$$

where e_i is a closest point in C_1 to $x_i(1)$.

Proof. Consider the optimal control problem in which the state vector, written (x^0, x^1, x), is a point in $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$:

Minimize $x^0(1) + \beta_i f(x(1)) + |x(1) - x^1(1)|$

over trajectories $(x^0(\cdot), x^1(\cdot), x(\cdot)) : [0, 1] \to \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n$ such that

$$\frac{d}{dt}(x^0(t), x^1(t), x(t)) \in \{\alpha_i g_i(t, x(t))\} \times \{0\} \times F(t, x(t)),$$

$(x^0(0), x^1(0), x(0)) \in \{0\} \times C_1 \times \{x_0\}$, and

$x(\cdot)$ has graph in Ω.

We see that

$$t \to \left(\alpha_i \int_0^t g_i(s, x_i(s))\, ds, e_i, x_i(t)\right) \tag{7.7}$$

is a solution to this problem, for each i. In light of Lemma 7.2

$$x_i^1(1) \neq e_i. \tag{7.8}$$

Now we apply to (7.7) the necessary conditions of [5, Thm. 2] in a slightly strengthened form (but one which can be established by minor modifications to the arguments in [5]), in which inclusion [2, (6)] is replaced by

$$p(0) \in \lambda\theta(z(0)) + l d_{C_0}(z(0)).$$

Here l is some positive number which depends only on the Lipschitz constants of the data.

It is an easy exercise in calculating generalized gradients to deduce the lemma from the necessary conditions and (7.8). □

We are now ready to complete the proof of the theorem. It is not difficult to show that (7.5) and (7.6) imply

$$|p_i(t)| \leqq (1 + \beta_i c + \alpha_i) + k \int_t^1 |p_i(s)|\, ds, \qquad t \in [0, 1], \quad i = 1, 2, \cdots. \tag{7.9}$$

Here c is the Lipschitz constant of $f(\cdot)$ and k is as in (3.1). Application of Gronwall's lemma to (7.9) yields the information that the family of functions $\{p_i(\cdot)\}$ is bounded with respect to the supremum norm. By (7.9) then, the family is uniformly Lipschitzian. The family $\{x_i(\cdot)\}$, too, is uniformly Lipschitzian by hypothesis (3.2). It follows that after extraction of subsequences,

$$p_i(\cdot) \to p(\cdot) \quad (\text{and } x_i(\cdot) \to z(\cdot)) \quad \text{uniformly,} \tag{7.10}$$

and

$$p_i(\cdot) \to \dot{p}_i(\cdot) \quad (\text{and } \dot{x}_i(\cdot) \to \dot{z}(\cdot)) \quad \text{weakly in } L^2 \tag{7.11}$$

for some Lipschitzian function $p(\cdot)$.

By Lemma (7.1),

$$e_i \to z(1). \tag{7.12}$$

After further extraction of subsequences

$$q_i \to q \tag{7.13}$$

for some $q \in \mathbb{R}^n$.

Consider now the differential inclusion (7.5). The set $\alpha_i \partial g_i(t, x_i(t))$ is contained in the closed ball of radius α_i. Recall $\alpha_i \to 0$. Using these facts, the upper semicontinuity of $(x, p) \to \partial H(t, x, p)$ and properties (7.10) and (7.11) we can show by familliar arguments (c.f. proof of [6, Lemma 5]) that the limiting functions $p(\cdot)$, $z(\cdot)$ satisfy

$$(-\dot{p}(t), \dot{z}(t)) \in \partial H(t, z(t), p(t)) \quad \text{a.e.} \tag{7.14}$$

We examine finally the boundary conditions (7.6). The set $\beta_i \partial f(x_i(1))$ is contained in a ball whose radius tends to zero as $i \to \infty$. It follows from (7.6), (7.10), (7.12), (7.13) and the upper semicontinuity of $x \to \partial d_{C_1}(x)$ that

$$p(1) \in \nu \partial d_{C_1}(z(1)) \tag{7.15}$$

and

$$|p(1)| = 1. \tag{7.16}$$

Equation (7.15) implies

$$p(1) \in N_{C_1}(z(1)). \tag{7.17}$$

The existence of a function $p(\cdot)$ satisfying (7.14), (7.16) and (7.17) establishes that (P) is not strongly normal.

REFERENCES

[1] V. G. BOLTYANSKI, *Sufficient conditions for optimality and the justification of the dynamic programming method*, this Journal, 4 (1966), pp. 326–361.

[2] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[3] ———, *Admissible relaxation in variational and control problems*, J. Math. Anal. Appl., 51 (1975), pp. 557–576.

[4] ———, *The generalized problem of Bolza*, this Journal, 14 (1976), pp. 682–699.

[5] ———, *Necessary conditions for a general control problem*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Mathematics Research Center, University of Wisconsin–Madison; Pub. No. 36, Academic Press, New York, 1976.

[6] ———, *The maximum principle under minimal hypotheses*, this Journal, 14 (1976), pp. 1078–1091.

[7] ———, *Optimization and Nonsmooth Analysis*, book, in preparation.

[8] M. R. GONZALES, *Sur l'existence d'une solution maximale de l'équation de Hamilton–Jacobi*, C. R. Acad. Sci., 282 (1976), pp. 1287–1290.

[9] A. D. IOFFE, *Convex functions occurring in variational problems and the absolute minimum problem*, Mat. Sb., 88 (1972), pp. 194–210; Math. USSR-Sb. (1972), pp. 191–208.

[10] A. D. IOFFE AND V. M. TIKHOMIROV, *Theory of Extremal Problems*, North-Holland, Amsterdam, 1979.

[11] V. L. LEVIN AND A. A. MILYUTIN, *The problem of mass transfer with a discontinuous cost function, and a mass statement of the duality theorem for convex extremal problems*, Uspekhi Mat. Nauk, 34 (1979), pp. 3–68; Russian Math. Surveys, 34 (1978), pp. 1–78.

[12] D. C. OFFIN, *A Hamilton–Jacobi approach to the differential inclusion problem*, Master's thesis, University of British Columbia, Vancouver, 1978.

[13] S. M. ROBINSON, *Stability theory for systems of inequalities* (Part II), SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[14] R. B. VINTER, *Weakest conditions for existence of Lipschitz continuous Krotov functions in optimal control theory*, this Journal, 21 (1983), pp. 215–234.

[15] ———, *The equivalence of "strong calmness" and "calmness" in optimal control theory*, J. Math. Anal. Appl., to appear.

[16] ———, *New global optimality conditions in optimal control theory*, this Journal, 21 (1983), pp. 235–245.

[17] R. B. VINTER AND R. M. LEWIS, *Necessary and sufficient conditions for optimality of dynamic programming type, making no a priori assumptions on the controls*, this Journal, 16 (1978), pp. 571–583.

[18] L. C. YOUNG, *Lectures on the Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0004 $01.25/0

OPTIMAL FEEDBACK CONTROLS FOR A CLASS OF NONLINEAR DISTRIBUTED PARAMETER SYSTEMS*

VIOREL BARBU†

Abstract. It is shown that the optimal controls for a distributed control problem governed by semilinear parabolic equations can be expressed in feedback form. The corresponding Hamilton–Jacobi equation is studied and applications to infinite time horizon and time optimal control problems are given. Some extensions to control problems governed by parabolic variational inequalities are indicated also.

Key words. optimal feedback control, optimal value function, Hamilton-Jacobi equation, control problem with infinite time horizon, time optimal control, approximating feedback control

1. Introduction. We consider the class of semilinear diffusion control processes described by the generalized heat equation

$$\begin{aligned} &y_t + Ay + \beta(y) = Bu \quad \text{in } Q = \Omega \times]0, T[, \\ &y(x, 0) = y_0(x), \qquad x \in \Omega, \end{aligned} \tag{1.1}$$

with the cost functional

$$\int_0^T (\Phi(y(t)) + h(u(t)))\, dt + \psi(y(T)). \tag{1.2}$$

Here the subscript t denotes the partial differentiation of $y : Q \to R =]-\infty, +\infty[$ with respect to t and Ω is a bounded and open subset of the Euclidean space R^N which has a sufficiently smooth boundary Γ. As regards the operators $A : D(A) \subset L^2(\Omega) \to L^2(\Omega)$, $B : U \to L^2(\Omega)$ and the functions $\beta : R \to R$, Φ, $\psi : L^2(\Omega) \to R$, $h : U \to \bar{R} =]-\infty, +\infty]$ the following conditions will be assumed:

(i) A is a linear, self-adjoint and positive operator in $L^2(\Omega)$ having the property that for every λ the subset $\{y \in D(A); (Ay, y) + |y|^2 \leqq \lambda\}$ is compact in $L^2(\Omega)$ and for every Lipschitz increasing function $\delta : R \to R$ with $\delta(0) = 0$, one has

$$(Ay, \delta(y)) \geqq -C|\delta(y)|(1 + |y|) \quad \text{for all } y \text{ in } D(A). \tag{1.3}$$

Here C is some positive constant which depends on δ and $(\cdot, \cdot)$, $|\cdot|$ are the scalar product and the norm, respectively, in $L^2(\Omega)$.

(ii) B is a linear continuous operator from a real Hilbert space U to $L^2(\Omega)$.

(iii) $\beta : R \to R$ is a locally Lipschitz monotonically increasing function such that $\beta(0) = 0$.

(iv) The functions $\Phi : L^2(\Omega) \to R$ and $\psi : L^2(\Omega) \to R$ are locally Lipschitz and nonnegative. The function $h : U \to \bar{R}$ is convex, lower semicontinuous (l.s.c.), nonidentically $+\infty$ and satisfies the growth condition

$$h(u) \geqq C_1 \|u\|^2 + C_2 \quad \text{for all } u \text{ in } U. \tag{1.4}$$

Here $\|\cdot\|$ denotes the norm of the control space U. The scalar product in U will be denoted by $\langle \cdot, \cdot \rangle$. From now on the space $L^2(\Omega)$ with the usual norm $|\cdot|$ and the scalar product $(\cdot, \cdot)$ will be denoted by H. The scalar product and the norm of the Euclidean space R^n will be denoted by $(\cdot, \cdot)_n$ and $|\cdot|_n$, respectively. By $D(A)$ we have denoted the domain of A endowed with the graph norm.

Let X be a real Banach space with the dual X^* and let $\varphi : X \to R$ be a locally Lipschitz function on X (by a locally Lipschitz function on X we mean a function

* Received by the editors March 23, 1982, and in revised form September 28, 1982.
† University Iasi, Iasi 6600, Romania.

which is Lipschitz on every bounded subset of X). Denote by $\varphi^0 : X \times X \to R$ the function

$$\varphi^0(y, v) = \limsup_{\substack{\lambda \downarrow 0 \\ z \to y}} (\varphi(z + \lambda v) - \varphi(z))\lambda^{-1} \tag{1.5}$$

and by $\partial\varphi : X \to 2^{X^*}$ the generalized gradient of φ ([9], [10], [16]), i.e.,

$$\partial\varphi(y) = \{y^* \in X^*; (y^*, v) \leqq \varphi^0(y, v) \text{ for all } v \text{ in } X\}. \tag{1.6}$$

(Here $(\cdot, \cdot)$ denotes the pairing between X and X^*.) If φ is convex then $\partial\varphi$ is just the subdifferential of φ in the sense of convex analysis (see e.g. [5, p. 89]). If φ admits a continuous Gâteaux derivative $\nabla\varphi$ then $\partial\varphi = \nabla\varphi$. Given a compact interval $[a, b]$ we shall denote by $C([a, b]; X)$ the space of all continuous X-valued functions on $[a, b]$ by $BV([a, b]; X)$ the space of all X-valued functions of bounded variation on $[a, b]$ and by $AC([a, b]; X)$ the space of all absolutely continuous functions $y : [a, b] \to X$. We shall denote by $W^{1,2}(a, b; X)$ the space $\{y \in L^2(a, b; X); y' \in L^2(a, b; X)\}$ where y' is the derivative of y in the sense of distributions. It is useful to recall that $W^{1,2}(a, b; X) \subset AC([a, b]; X)$ and for any $y \in W^{1,2}(a, b; X)$, y' is just the ordinary derivative of y (which exists almost everywhere). By $AC_{\text{loc}}(R^+; X)$ we shall denote the space of all functions $y : R^+ = [0, +\infty[\to X$ which are absolutely continuous on every compact interval $[0, T]$. The spaces $L^2_{\text{loc}}(R^+; X)$ and $W^{1,2}_{\text{loc}}(R^+; X)$ are similarly defined.

Let $F; H \to H$ be the nonlinear operator given by

$$Fy = Ay + \beta(y) \quad \text{for } y \in D(F), \tag{1.7}$$

where

$$D(F) = \{y \in D(A); \beta(y) \in H\}. \tag{1.8}$$

By (1.3) and assumption (iii) it follows that the operator F is maximal monotone in $H \times H$ (see for instance [1, p. 83] and [8]). As a matter of fact F is the subdifferential of the convex and l.s.c. function $l : H \to \bar{R}$,

$$l(y) = \frac{1}{2}|A^{1/2}y|^2 + \int_\Omega j(y(x))\, dx, \qquad y \in H, \tag{1.9}$$

where $j(r) = \int_0^r \beta(s)\, ds$ for $r \in R$. Also $\overline{D(F)} = H$.

In terms of F, the Cauchy problem (1.1) can be written as

$$\begin{aligned} &y'(t) + Fy(t) = Bu(t), \quad \text{a.e. } t \in]0, T[, \\ &y(0) = y_0. \end{aligned} \tag{1.1$'$}$$

According to the general existence theory of nonlinear evolution equations (see [8] and [1, p. 202]) for each $y_0 \in H$ and $u \in L^2(t, T; U)$, $0 \leqq t \leqq T$ the Cauchy problem

$$\begin{aligned} &y'(s) + Fy(s) = Bu(s), \quad \text{a.e. } s \in]t, T[, \\ &y(t) = y_0, \end{aligned} \tag{1.10}$$

has a unique solution, denoted by $y(s, t, y_0, u)$, satisfying

$$y(s, t, y_0, u) \in C([t, T]; H), \qquad l(y(s, t, y_0, u)) \in L^1(t, T), \tag{1.11}$$

$$(s - t)^{1/2} y(s, t, y_0, u) \in L^2(t, T; D(A)), \tag{1.12}$$

$$(s - t)^{1/2} y_s(s, t, y_0, u) \in L^2(t, T; H). \tag{1.13}$$

If $l(y_0)<+\infty$ then $y(\cdot, t, y_0, u)\in W^{1,2}(t, T; H)\cap L^2(t, T; D(A))$. Along with the Arzela–Ascoli theorem the latter implies that for every $y_0\in D(l)$ the map $u\to y(\cdot, t, y_0, u)$ is compact from $L^2(t, T; U)$ to $C([t, T]; H)$. In general, for $y_0\in H$ we see by (1.12) and (1.13) that this map is compact from $L^2(t, T; U)$ to $C(]t, T]; H)\cap L^2(t, T; H)$. In (1.13) y_s denotes the derivative y' of $y:[t, T]\to H$.

We shall denote by $\varphi:[0, T]\times H\to R$ the *optimal value function* of the control system (1.1) with the cost criterion (1.2), i.e.,

$$\varphi(t, y_0)=\inf\left\{\int_t^T (\Phi(y(s, t, y_0, u))+h(u(s)))\,ds \right. \\ \left. +\psi(y(T, t, y_0, u));\ u\in L^2(t, T; U)\right\}. \tag{1.14}$$

The contents of the paper are outlined below.

In § 2 necessary conditions of optimality for problem (1.1), (1.2) are derived in terms of the generalized gradients of the functions β, Φ, h and ψ.

Proposition 1 and 2 given here for the purposes of §§ 3 and 4 have an intrinsic interest and extend in certain directions some results in [3] (see also [17]). In § 3 it is proved that every optimal control u for problem (1.1), (1.2) is a *feedback optimal control* expressed by the formula

$$u(t)\in -\partial\varphi(t, y(t)), \qquad t\in[0, T], \tag{1.15}$$

where $\partial\varphi$ is the generalized gradient of the function $y\to\varphi(t, y)$. Furthermore, it will be shown that the function φ is the solution in a certain generalized sense to a partial differential equation of Hamilton–Jacobi type. In § 4 similar results are obtained for control problems of the form (1.1), (1.2) with infinite time horizon, i.e., $T=+\infty$. In § 5 some applications to optimal time control problems associated with system (1.1) are given. In § 6 are indicated some extensions to control problems governed by parabolic variational inequalities.

To conclude this section it is worth noting that a typical example of operator A satisfying assumption (i) is the following

$$Ay=-\sum_{i,j=1}^N (a_{ij}(x)y_{x_i})_{x_j}+A_0y, \\ D(A)=\left\{y\in H^2(\Omega);\ \alpha_1y+\alpha_2\frac{\partial y}{\partial\nu}=0, \text{ a.e. on } \Gamma\right\}. \tag{1.16}$$

Here A_0 is a linear continuous, positive, self-adjoint operator from H to itself, $\alpha_1+\alpha_2>0$, $\alpha_i\geqq 0$ for $i=1, 2$ and $\partial/\partial\nu$ is the outward normal derivative corresponding to a_{ij}. We must assume also that $a_{ij}\in C^1(\bar{\Omega})$, $a_{ij}=a_{ji}$ for all $i, j=1, \cdots, N$ and for some $\omega>0$,

$$\sum_{i,j=1}^N a_{ij}(x)\xi_i\xi_j\geqq\omega|\xi|_N^2, \quad \text{a.e. } x\in\Omega, \quad \xi\in R^N. \tag{1.17}$$

Nonlinear systems of the form (1.1) occur for instance in the temperature control of a heat conductor Ω (see [11, p. 28]). In this case the function β has the following

form:

$$\beta(r)=\begin{cases} g_1, & \text{if } r\leqq h_1+g_1k_1^{-1},\\ k_1(r-h_1), & \text{if } h_1+g_1k_1^{-1}\leqq r\leqq h_1,\\ 0, & \text{if } h_1\leqq r\leqq h_2,\\ k_2(r-h_2), & \text{if } h_2<r\leqq h_2+g_2k_2^{-1},\\ g_2, & \text{if } r>h_2+g_2k_2^{-1}, \end{cases}$$

where g_1, g_2 and h_1, h_2, k_1, $k_2\geqq 0$ are real constants.

2. The generalized Euler–Lagrange equations. We shall study here the optimal control problem with state system (1.1) and cost (1.2), i.e.,

$$\text{(2.1)}\quad \inf\left\{\int_0^T (\Phi(y(s,0,y_0,u))+h(u(s)))\,ds+\psi(y(T,0,y_0,u));\ u\in L^2(0,T;U)\right\},$$

where $y_0\in D(l)$, i.e.,

$$\text{(2.2)}\qquad y_0\in D(A^{1/2}),\qquad j(y_0)\in L^1(\Omega)$$

and the functions $\Phi: H\to R$, $\psi: H\to R$, $h: U\to\bar{R}$ and A, β satisfy assumptions (i)–(iv). Since, as remarked earlier, the map $u\to y(t,0,y_0,u)$ is compact from $L^2(0,T;U)$ to $C([0,T];H]$ it follows by (1.4) that problem (2.1) has at least one optimal control (see Lemma 3 below). Let u^* be a such optimal control and let $y^*=y(\cdot,0,y_0,u^*)$ be the corresponding state. We notice that $y^*\in W^{1,2}(0,T;H)\cap L^2(0,T;D(A))$. For any $\varepsilon>0$ consider the control problem: minimize

$$\text{(2.3)}\qquad \int_0^T \left(\Phi^\varepsilon(y(t))+h_\varepsilon(u(t))+\frac{1}{2}\|u(t)-u^*(t)\|^2\right)dt+\psi^\varepsilon(y(T))$$

over all $u\in L^2(0,T;U)$ and $y\in W^{1,2}(0,T;H)$ subject to

$$\text{(2.4)}\qquad \begin{aligned} &y_t+Ay+\beta^\varepsilon(y)=Bu,\quad \text{a.e. } t\in\,]0,T[,\\ &y(0)=y_0, \end{aligned}$$

where

$$\text{(2.5)}\qquad h_\varepsilon(u)=\inf\{(2\varepsilon)^{-1}\|u-v\|^2+h(v);\ v\in U\},$$

$$\text{(2.6)}\qquad \beta^\varepsilon(r)=\int_{-\infty}^{+\infty}\beta_\varepsilon(r-\varepsilon\tau)\rho(\tau)\,d\tau.$$

Here $\beta_\varepsilon=\varepsilon^{-1}(1-(1+\varepsilon\beta)^{-1})$ and ρ is a C_0^∞-"mollifier" on R. The functions Φ^ε and ψ^ε are defined as follows. Let $\{e_i\}$ be an orthonormal basis in H and let X_n be the linear space generated by $\{e_i\}_{i=1}^n$. For $n=[\varepsilon^{-1}]$ we define

$$\text{(2.7)}\qquad \begin{aligned} \Phi^\varepsilon(y)&=\int_{R^n}\Phi(P_ny-\varepsilon\Lambda_n\tau)\rho_n(\tau)\,d\tau,\qquad y\in H,\\ &=\int_{R^n}\Phi(\Lambda_n\theta)\rho_n\left(\frac{\Lambda_n^{-1}P_ny-\theta}{\varepsilon}\right)d\theta, \end{aligned}$$

and

$$\text{(2.8)}\qquad \psi^\varepsilon(y)=\int_{R^n}\psi(P_ny-\varepsilon\Lambda_n\tau)\rho_n(\tau)\,d\tau,\qquad y\in H,$$

where $P_n : H \to X_n$ is the projection operator on X_n, ρ_n is a C_0^∞-"mollifier" in R^n and $\Lambda_n : R^n \to X_n$ is the operator

$$\Lambda_n(\tau) = \sum_{i=1}^{n} \tau_i e_i, \qquad \tau = (\tau_1, \cdots, \tau_n),$$

Clearly the functions Φ^ε, ψ^ε, h_ε are Lipschitz, Fréchet differentiable and with the Fréchet differentials $\nabla\Phi^\varepsilon$, $\nabla\psi^\varepsilon$, ∇h_ε Lipschitzian on H and U, respectively. As a matter of fact since Φ and ψ are locally Lipschitzian we see that $\nabla\Phi^\varepsilon$ and $\nabla\psi^\varepsilon$ are bounded (uniformly with respect to ε) on every bounded subset. In other words, for every $r > 0$ there exists $C_r > 0$ independent of ε such that

$$|\nabla\Phi^\varepsilon(y)| + |\nabla\psi^\varepsilon(y)| \leqq C_r \quad \text{for } |y| \leqq r. \tag{2.9}$$

Let $(y_\varepsilon, u_\varepsilon)$ be an optimal pair for problem (2.3).

LEMMA 1. *For $\varepsilon \to 0$ one has*

$$u_\varepsilon \to u^* \quad \textit{strongly in } L^2(0, T; U), \tag{2.10}$$

$$y_\varepsilon \to y^* \quad \textit{strongly in } C([0, T]; H) \textit{ and weakly in } W^{1,2}(0, T; H) \cap L^2(0, T; D(A)). \tag{2.11}$$

Proof. We have

$$\begin{aligned} &\int_0^T \left(\Phi^\varepsilon(y_\varepsilon) + h_\varepsilon(u_\varepsilon) + \frac{1}{2}\|u_\varepsilon - u^*\|^2\right) dt + \psi^\varepsilon(y_\varepsilon(T)) \\ &\qquad \leqq \int_0^T (\Phi^\varepsilon(z_\varepsilon) + h(u^*))\, dt + \psi^\varepsilon(z_\varepsilon(T)), \end{aligned} \tag{2.12}$$

where $z_\varepsilon \in W^{1,2}(0, T; H)$ is the solution to

$$\begin{aligned} &(z_\varepsilon)_t + Az_\varepsilon + \beta^\varepsilon(z_\varepsilon) = Bu^*, \quad \text{a.e. } t \in]0, T[, \\ &z_\varepsilon(0) = y_0. \end{aligned}$$

Using assumption (i) it follows by a standard argument (see [3]) that $z_\varepsilon \to y^*$ in $C([0, T]; H)$. Hence

$$\lim_{\varepsilon \to 0} \int_0^T dt \int_{R^n} \Phi(P_n z_\varepsilon - \varepsilon \Lambda_n \tau)\rho_n(\tau)\, d\tau = \int_0^T \Phi(y^*)\, dt \tag{2.13}$$

because $\int_{R^n} \rho_n(\tau)\, d\tau = 1$, support $\rho_n \subset \{\tau \in R^n;\ |\tau|_n \leqq 1\}$ and

$$\begin{aligned} &|\Phi(P_n z_\varepsilon(t) - \varepsilon \Lambda_n \tau) - \Phi(y^*(t))| \\ &\qquad \leqq L(|z_\varepsilon(t) - y^*(t)| + n^{-1}|\tau|_n + |P_n y^*(t) - y^*(t)|) \quad \text{for } t \in [0, T]. \end{aligned}$$

Similarly, we have

$$\lim_{\varepsilon \to 0} \psi^\varepsilon(z_\varepsilon(T)) = \psi(y^*(T)). \tag{2.14}$$

On the other hand, it follows by (2.12) that $\{u_\varepsilon\}$ is bounded in $L^2(0, T; U)$. Hence one some subsequence again denoted $\{\varepsilon\}$

$$u_\varepsilon \to u_1 \quad \text{weakly in } L^2(0, T; U),$$

$$y_\varepsilon \to y_1 \quad \text{weakly in } W^{1,2}(0, T; H) \text{ and strongly in } C([0, T]; H),$$

and therefore

$$\lim_{\varepsilon\to 0}\int_0^T \Phi^\varepsilon(y_\varepsilon)\,dt = \int_0^T \Phi(y_1)\,dt, \tag{2.15}$$

$$\lim_{\varepsilon\to 0}\psi^\varepsilon(y_\varepsilon(T)) = \psi(y_1(T)). \tag{2.16}$$

Finally, by the Fatou lemma,

$$\liminf_{\varepsilon\to 0}\int_0^T h_\varepsilon(u_\varepsilon)\,dt \geqq \int_0^T h(u_1)\,dt. \tag{2.17}$$

Along with (2.12), (2.13), (2.14) formulas (2.15), (2.16), (2.17) imply (2.10). As regards (2.11) it follows by (2.10).

Since the functions Φ^ε, ψ^ε, h_ε and β_ε are differentiable it follows by a standard device that there exists $p_\varepsilon \in W^{1,2}(0, T; H)\cap L^2(0, T; D(A))$ which satisfies along with y_ε and u_ε the system

$$p_\varepsilon' - Ap_\varepsilon - p_\varepsilon\nabla\beta^\varepsilon(y_\varepsilon) = \nabla\Phi^\varepsilon(y_\varepsilon), \quad \text{a.e. } t\in\,]0, T[, \tag{2.18}$$

$$p_\varepsilon(T) + \nabla\psi^\varepsilon(y_\varepsilon(T)) = 0, \tag{2.19}$$

$$B^*p_\varepsilon(t) = \nabla h_\varepsilon(u_\varepsilon(t)) + u_\varepsilon(t) - u^*(t), \quad \text{a.e. } t\in\,]0, T[. \tag{2.20}$$

We take the scalar product of (2.18) (in the space H) by $p_\varepsilon(t)$ and integrate on $[t, T]$. We get

$$|p_\varepsilon(t)|^2 + \int_0^T (Ap_\varepsilon(t), p_\varepsilon(t))\,dt \leqq C, \qquad t\in[0, T] \tag{2.21}$$

because $(\beta^\varepsilon)' \geqq 0$ on R and by (2.9) $\{\nabla\Phi^\varepsilon(y_\varepsilon)\}$ is bounded in $C([0, T]; H)$. (We shall denote by C several positive constants independent of ε.) Now we take the scalar product of (2.18) with $\zeta(p_\varepsilon)$ where ζ is a C^∞-increasing function on R which approximates the signum function. Integrating on $[0, T]$ and using condition (1.3) we see that

$$\int_Q p_\varepsilon\zeta(p_\varepsilon)\,\nabla\beta^\varepsilon(y_\varepsilon)\,dx\,dt \leqq \int_Q \nabla\Phi^\varepsilon(y_\varepsilon)\zeta(p_\varepsilon)\,dx\,dt + C\int_Q |\zeta(p_\varepsilon)|^2\,dx\,dt. \tag{2.22}$$

Then letting ζ tend to sgn we get the estimate

$$\|\nabla\beta^\varepsilon(y_\varepsilon)p_\varepsilon\|_{L^1(Q)} \leqq C \quad \text{for all } \varepsilon > 0. \tag{2.23}$$

Let us denote by V the space $D(A^{1/2})$ endowed with the norm

$$\|y\|_V^2 = |A^{1/2}y|^2 + |y|^2$$

and with the usual Hilbert structure. Denote by V^* the dual space of V. We have $V\subset H\subset V^*$ and notice that by assumption (i) the injection of V into H is compact and A is continuous from V to V^*. Then by (2.22) it follows that $\{Ap_\varepsilon\}$ is bounded in $L^2(0, T; V^*)$. Hence $\{p_\varepsilon'\}$ is bounded in the space $L^1(0, T; L^1(\Omega) + V^*)$. Since by the Sobolev embedding theorem $L^1(\Omega)\subset (H^s(\Omega))^*$ for $s > N/2$ we may conclude that $\{p_\varepsilon'\}$ is bounded in the space $L^1(0, T; (H^s(\Omega))^* + V^*)$. We shall denote by Y^* the space $(H^s(\Omega))^* + V^*$ which can be viewed as the dual space of $Y = H^s(\Omega)\cap V$. Since the injection of H into Y^* is compact (because the injection of Y into H is compact) by the Helly theorem in infinite dimensional spaces we may infer that there exists a subsequence of $\{p_\varepsilon\}$ which converges pointwise to a function $p\in BV([0, T]; Y^*)$ in

the strong topology of Y^*. In other words,

$$p_\varepsilon(t) \to p(t) \text{ strongly in } Y^* \quad \text{for } t \in [0, T]. \tag{2.24}$$

On the other hand, for every $\delta > 0$ there is $C(\delta) > 0$ such that (see [14, Chapt. I, Lemma 5.1])

$$|p_\varepsilon(t) - p(t)| \leqq \delta \|p_\varepsilon(t) - p(t)\|_V + C(\delta)\|p_\varepsilon(t) - p(t)\|_{Y^*}, \qquad t \in [0, T].$$

Along with (2.21) and (2.24) the latter implies that

$$p_\varepsilon \to p \quad \text{strongly in } L^2(Q) \text{ and weakly in } L^2(0, T; V) \tag{2.25}$$

and

$$p_\varepsilon(t) \to p(t) \quad \text{weakly in } H \text{ for all } t \in [0, T]. \tag{2.26}$$

By (2.23) we see that on some subsequence we have

$$\mu_p = \lim_{\varepsilon \to 0} \nabla\beta^\varepsilon(y_\varepsilon) p_\varepsilon \quad \text{weak star in } M(\bar{Q}), \tag{2.27}$$

where $M(\bar{Q})$ is the space of all bounded Radon measures on $\bar{Q}$. Summarizing at this point, we have shown that there exists $p \in L^2(0, T; V) \cap L^\infty(0, T; H) \cap BV([0, T]; Y^*)$ and $\mu_p \in M(\bar{Q})$ which are the limit in the sense of (2.24), (2.25), (2.26) and (2.27) on some subsequence and satisfy the system

$$\begin{aligned} &p' - Ap - \mu_p = \nu && \text{in } Q, \\ &p(T) = \eta && \text{in } \Omega, \end{aligned} \tag{2.28}$$

$$B^* p(t) \in \partial h(u^*(t)), \quad \text{a.e. } t \in]0, T[. \tag{2.29}$$

Here $\nu \in L^2(Q)$ and $\eta \in H$ are the weak limits of $\nabla\Phi^\varepsilon(y_\varepsilon)$ and $\nabla\psi^\varepsilon(y_\varepsilon(T))$ in $L^2(Q)$ and H, respectively and p' is the derivative of p in the sense of vectorial distributions.

We need the following lemma.

LEMMA 2. *Let $\{y_n\}$ be a sequence strongly convergent to y in H and such that*

$$\nabla\Phi^\varepsilon(y_n) \to \pi \quad \textit{weakly in } H \textit{ for } \varepsilon = n^{-1} \to 0. \tag{2.30}$$

Then $\pi \in \partial\Phi(y)$.

Proof. By the theorem of the mean and formula (2.7) we see that

$$\begin{aligned} &\lambda^{-1}(\Phi^\varepsilon(y_n + \lambda z) - \Phi^\varepsilon(y_n)) \\ &\quad = \lambda^{-1}(\Phi(P_n(y_n + \lambda z) - \varepsilon \Lambda_n \tau_{n,\lambda}) - \Phi(P_n y_n - \varepsilon \Lambda_n \tau_{n,\lambda})), \end{aligned}$$

where $|\tau_{n,\lambda}|_n \leqq 1$. On some subsequence $\lambda \to 0$ we have $\tau_{n,\lambda} \to \tau_n$ and therefore

$$(\nabla\Phi^\varepsilon(y_n), P_n z) \leqq \Phi^0(P_n y_n - \varepsilon \Lambda_n \tau_n, P_n z) \quad \text{for all } z \text{ in } H. \tag{2.31}$$

Inasmuch as the function Φ^0 is upper semicontinuous on $H \times H$ (see e.g. [10]) the latter yields $(\pi, z) \leqq \Phi^0(y, z)$ for all $z \in H$. Hence $\pi \in \partial\Phi(y)$ as claimed.

Coming back to the system (2.28) we may conclude by (2.30) and Lemma 2 that $\eta \in \partial\psi(y^*(T))$.

On the other hand, since by (2.9) and (2.11) $\{\nabla\Phi^\varepsilon(y_\varepsilon)\}$ is bounded in $C([0, T]; H)$ we may infer that

$$\nabla\Phi^\varepsilon(y_\varepsilon) \to \nu \quad \text{weak star in } L^\infty(0, T; H).$$

Then by (2.31) we see that

$$\int_0^T (\nu(t), z(t))\, dt \leqq \int_0^T \Phi^0(y^*(t); z(t)) \quad \forall z \in L^\infty(0, T; H),$$

and this yields

$$\nu(t) \in \partial\Phi(y^*(t)), \quad \text{a.e. } t \in]0, T[.$$

We have therefore proved:

PROPOSITION 1. *Let* $(y^*, u^*) \in W^{1,2}(0, T; H) \cap L^2(0, T; U)$ *be an optimal pair for problem* (2.1). *Then there exists* $p \in BV([0, T]; Y^*) \cap L^\infty(0, T; H) \cap L^2(0, T; V)$ *and* $\mu_p \in M(\bar{Q})$ *such that*

$$p' - Ap - \mu_p \in L^\infty(0, T; H), \tag{2.32}$$

$$p' - Ap - \mu_p \in \partial\Phi(y^*), \quad \textit{a.e. on }]0, T[, \tag{2.33}$$

$$p(T) + \partial\psi(y^*(T)) \ni 0, \tag{2.34}$$

$$B^* p(t) \in \partial h(u^*(t)), \quad \textit{a.e. } t \in]0, T[. \tag{2.35}$$

The function p is called the *dual extremal arc* associated with the optimal pair (y^*, u^*).

Let us assume now that β satisfies the following condition:

(a) β is a monotonically increasing, locally Lipschitz function on R such that $\beta(0) = 0$ and

$$\beta'(r) \leqq C(|\beta(r)| + |r| + 1), \quad \text{a.e. } r \in R. \tag{2.36}$$

PROPOSITION 2. If β satisfies assumption (a) then $\mu_p \in L^1(Q)$, $p \in AC([0, T]; Y^*) \cap C_w([0, T]; H)$ for $s > N/2$, and

$$\mu_p(x, t) \in p(x, t)\, \partial\beta(t^*(x, t)), \quad \text{a.e. } (x, t) \in Q. \tag{2.37}$$

Here $C_w([0, T]; H)$ is the space of all weakly continuous functions from $[0, T]$ to H.

Proof. By (2.36) we see after some elementary calculations that

$$|\nabla\beta^\varepsilon(y)| \leqq C(|\beta^\varepsilon(y)| + |y| + 1) \quad \forall y \in R$$

and therefore

$$\begin{aligned} \int_E |p_\varepsilon|\, \nabla\beta^\varepsilon(y_\varepsilon)\, dx\, dt \leqq C\Big(& \int_E |p_\varepsilon|\, |\beta^\varepsilon(y_\varepsilon)|\, dx\, dt \\ & + \int_E |p_\varepsilon|\, |y_\varepsilon|\, dx\, dt + \int_E |p_\varepsilon|\, dx\, dt \Big), \end{aligned} \tag{2.38}$$

where E is an arbitrary measurable subset of Q. By Lemma 1 and (2.25) we see that $\{p_\varepsilon \beta^\varepsilon(y_\varepsilon)\}$ is weakly convergent in $L^1(Q)$ while $\{p_\varepsilon\}$ and $\{y_\varepsilon\}$ are strongly convergent in $L^2(Q)$. Along with (2.38) these imply that the family $\{\int_E p_\varepsilon\, \nabla\beta^\varepsilon(y_\varepsilon)\, dx\, dt;\ E \subset Q\}$ is uniformly absolutely continuous and therefore by the Dunford–Pettis criterion $\{p_\varepsilon\, \nabla\beta^\varepsilon(y_\varepsilon)\}$ is weakly compact in $L^1(Q) \subset L^1(0, T; (H^s(\Omega))^*)$. Hence $\mu_p \in L^1(Q)$ and since $p' \in L^2(0, T; V) + L^1(0, T; (H^s(\Omega))^*) \subset L^1(0, T; Y^*)$, $p \in AC([0, T]; Y^*)$.

Since $p \in L^\infty(0, T; H) \cap C([0, T]; Y^*)$ we see that for every t_0 and each sequence $\{t_n\} \to t_0$ the weak limit in H of $p(t_n)$ exists and equals $p(t_0)$. Hence $p : [0, T] \to H$ is

weakly continuous. Relation (2.37) has been proved in [3]. However, for the reader's convenience we outline its proof. By Egorov's theorem, for each $\eta > 0$ there exists a measurable subset $E_\eta \subset Q$ such that $\{y_\varepsilon\}$ is bounded on E_η and $y_\varepsilon \to y^*$ uniformly on E_η for $\varepsilon \to 0$. Hence $\nabla\beta^\varepsilon(y_\varepsilon)$ are uniformly bounded on E_η so that without loss of generality we may assume that

$$\nabla\beta^\varepsilon(y_\varepsilon) \to g \quad \text{weak star in } L^\infty(E_\eta).$$

Then by [4, Lemma 3] we may infer that $g(x,t) \in \partial\beta(y^*(x,t))$ a.e. $(x,t) \in E_\eta$. Since $\{p_\varepsilon\}$ is strongly convergent to p in $L^2(Q)$ we may conclude that

$$\mu_p(x,t) = p(x,t)g(x,t) \in p(x,t)\,\partial\beta(y^*(x,t)), \quad \text{a.e. } (x,t) \in E_\eta.$$

Since η is arbitrarily small (2.37) follows.

PROPOSITION 3. *In addition to the assumptions of Proposition* 2 *suppose that either* β *is globally Lipschitz or* $D(A) \subset H^2(\Omega)$ *and* $N = 1$. *Then* $\mu_p \in L^2(Q)$, $p \in C([0,T];H)$ *and for every* $0 < \delta < T$,

$$p \in L^2(\delta, T; D(A)), \qquad p' \in L^2(\delta, T; H). \tag{2.39}$$

If $\psi \equiv 0$ *then* $p \in L^2(0,T;D(A)) \cap W^{1,2}(0,T;H)$.

Proof. If β is globally Lipschitz then $\{\nabla\beta^\varepsilon\}$ is uniformly bounded on R and therefore $\{p_\varepsilon\, \nabla\beta^\varepsilon(y_\varepsilon)\}$ is weakly compact in $L^2(Q)$. Hence $\mu_p \in L^2(Q)$ and (2.39) follows by (1.12), (1.13). If $\psi \equiv 0$ then $p(T) = 0 \in D(A)$ and therefore $p \in L^2(0,T;D(A)) \cap W^{1,2}(0,T;H)$. If $D(A) \subset H^2(\Omega)$ and $N = 1$ then

$$W^{1,2}(0,T;H) \cap L^2(0,T;D(A)) \subset C(\bar{Q})$$

($C(\bar{Q})$ is the space of all continuous functions on $\bar{Q}$) and by (2.11) it follows that $\{y_\varepsilon\}$ is a bounded subset of $C(\bar{Q})$. Thus $\nabla\beta^\varepsilon(y_\varepsilon)$ are uniformly bounded in $\bar{Q}$ and the conclusion follows as in the first case.

Remark 1. Condition $D(A) \subset H^2(\Omega)$ is in particular satisfied by the operator A defined by (1.16). As regards assumption (a) it is satisfied by a large class of monotone nonlinear functions growing not faster than exponentially at $\pm\infty$. In particular it is satisfied by any polynomial monotonically increasing function β.

3. Optimal feedback controls for problem (2.1). Let $\varphi : [0,T] \times H \to R$ be the optimal value function defined by (1.14). Observe that φ is everywhere finite on $[0,T] \times H$. Other elementary properties of φ are indicated in the lemmas which follow.

LEMMA 3. *For every* $(t, y_0) \in [0,T] \times H$ *the infimum defining* $\varphi(t, y_0)$ *is attained. For every* $t \in [0,T]$ *the function* $\varphi(t,\cdot)$ *is locally Lipschitz and for each* $y_0 \in D(F)$ *the function* $t \to \varphi(t, y_0)$ *is Lipschitz on* $[0,T]$.

Proof. As remarked earlier the map $u \to y(\cdot, t, y_0, u)$ is compact from $L^2(t,T;U)$ to $C([t,T];H) \cap L^2(t,T;H)$. Since the function $u \to \int_t^T h(u(s))\,ds$ is weakly lower semicontinuous it follows by condition (1.4) via a standard device that the infimum in (1.14) is attained. Let $t \in [0,T]$ be arbitrary but fixed. Since the operator F defined by (1.7) and (1.8) is monotone we have

$$|y(s,t,y_0,u) - y(s,t,\tilde{y}_0,u)| \leqq |y_0 - \tilde{y}_0|, \qquad t \leqq s \leqq T, \tag{3.1}$$

$$|y(s,t,y_0,u)| \leqq |y_0| + \|B\| \int_t^s \|u(\tau)\|\,d\tau, \qquad t \leqq s \leqq T, \tag{3.2}$$

where $\|B\|$ is the norm of B. Let $y_0 \in H$ be such that $|y_0| \leqq r$. We have (we may assume $h(0) < +\infty$)

$$\varphi(t, y_0) \leqq \int_t^T (\Phi(y(s, t, y_0, 0) + h(0))\, ds + \psi(y(T, t, y_0, 0)) \leqq C_r. \tag{3.3}$$

Along with (1.14) the latter implies that in (1.14) we may restrict the infimum to the class $\mathcal{M}_r = \{u \in L^2(t, T; U);\ \int_t^T \|u(\tau)\|^2\, d\tau \leqq C_r^1\}$ where C_r^1 is independent of t and y_0. Let u^t be an optimal control for problem (1.14) and $p^t \in L^\infty(t, T; H)$ be a corresponding dual arc. By (2.18), (2.19) we see that for $u^t \in \mathcal{M}_r$ $|p^t(s)| \leqq C_r^2$ a.e. $s \in]t, T[$ and so (2.35) implies that $u^t \in L^\infty(t, T; U)$ and $\|u^t(s)\| \leqq C_r^3$ a.e. $s \in]t, T[$ (because by condition (1.4) $(\partial h)^{-1}$ is bounded on bounded sets). Hence without loss of generality we may take $\mathcal{M}_r = \{u \in L^\infty(t, T; U);\ \|u^t(s)\| \leqq C_r^3\}$, a.e. $s \in]t, T[$.

By (3.1) and (3.2) we see that for every $u \in \mathcal{M}_r$ the function $y_0 \to \Phi(y(s, t, y_0, u))$ is locally Lipschitz with Lipschitz constant independent of u. Hence $y \to \varphi(t, y)$ is locally Lipschitz on H. Next for every $y_0 \in D(F)$ we have

$$|y(s, \tilde{t}, y_0, u) - y(s, t, y_0, u)| \leqq |y(t, \tilde{t}, y_0, u) - y_0| \leqq (|Fy_0| + C_r^3)|t - \tilde{t}|. \tag{3.4}$$

Let $u_t \in L^2(t, T; U)$ and $y_t = y(s, t, y_0, u_t)$ be such that

$$\varphi(t, y_0) = \int_t^T (\Phi(y_t(s)) + h(u_t(s)))\, ds + \psi(y_t(T))$$

and let $v(s) = u_0$ for $\tilde{t} \leqq s \leqq t$, $v(s) = u_t(s)$ for $t \leqq s \leqq T$ where u_0 is such that $h(u_0) < +\infty$. We have

$$\begin{aligned}\varphi(\tilde{t}, y_0) - \varphi(t, y_0) \leqq &\int_{\tilde{t}}^t (\Phi(y(s, \tilde{t}, y_0, v) + h(u_0)))\, ds \\ &+ \int_t^T (\Phi(y(s, \tilde{t}, y_0, v)) - \Phi(y(s, t, y_0, v)))\, ds \\ &+ \psi(y(T, \tilde{t}, y_0, v)) - \psi(y(T, t, y_0, v)).\end{aligned}$$

Along with (3.4) the latter implies that

$$|\varphi(\tilde{t}, y_0) - \varphi(t, y_0)| \leqq L|\tilde{t} - t| \quad \text{for all } \tilde{t}, t \in [0, T].$$

LEMMA 4. *For all $t \in [0, T]$ and $y_0 \in H$ we have*

$$\begin{aligned}\varphi(0, y_0) = \inf \Big\{ &\int_0^t (\Phi(y(s, 0, y_0, u)) + h(u(s)))\, ds \\ &+ \varphi(t, y(t, 0, y_0, u));\ u \in L^2(0, t; U) \Big\}.\end{aligned} \tag{3.5}$$

Proof. Let (y, u) be such that $y(s, 0, y_0, u) = y$ and

$$\begin{aligned}\varphi(0, y_0) = &\int_0^t (\Phi(y(s)) + h(u(s)))\, ds \\ &+ \int_t^T (\Phi(y(s)) + h(u(s)))\, ds + \psi(y(T)).\end{aligned}$$

The latter yields

$$\varphi(0, y_0) \geqq \varphi(t, y(t)) + \int_0^t (\Phi(y(s)) + h(u(s)))\, ds. \tag{3.6}$$

On the other hand, for all $u \in L^2(0, T; U)$ and $y = y(s, 0, y_0, u)$ we have

$$\varphi(0, y_0) \leqq \int_0^t (\Phi(y(s)) + h(u(s)))\, ds + \int_t^T (\Phi(y(s)) + h(u(s)))\, ds + \psi(y(T)).$$

We may choose the pair (y, u) in such a way that

$$\varphi(t, y(t)) = \int_t^T (\Phi(y(s)) + h(u(s)))\, ds + \psi(y(T))$$

and therefore

$$\varphi(0, y_0) \leqq \int_0^t (\Phi(y(s)) + h(u(s)))\, ds + \varphi(t, y(t)).$$

Along with (3.6) the latter inequality implies (3.5) as claimed.

We are now ready to prove the main result of this section.

THEOREM 1. *Let assumptions* (i)–(iv) *and* (a) *be satisfied. Let* (y^*, u^*) *be an optimal pair in problem* (2.1) *where* $y_0 \in D(l)$. *Then*

$$u^*(t) \in \partial h^*(-B^* \partial \varphi(t, y^*(t))), \quad \text{a.e. } t \in]0, T[. \tag{3.7}$$

Proof. By Lemma 4, for every $t \in [0, T]$ the pair (y^*, u^*) restricted to $[0, t]$ is optimal for the control problem

$$\inf\left\{\int_0^t (\Phi(y(s, 0, y_0, u)) + h(u(s)))\, ds + \varphi(t, y(t, 0, y_0, u)); u \in L^2(0, t; U)\right\}. \tag{3.8}$$

Then in virtue of Proposition 2, for every $t \in [0, T]$ there exists $p^t \in AC([0, t]; Y^*) \cap C_w([0, t]; H)$ which satisfies the equation

$$B^* p^t(s) \in \partial h(u^*(s)), \quad \text{a.e. } s \in]0, t[\tag{3.9}$$

and the transversality condition

$$p^t(t) \in -\partial \varphi(t, y^*(t)). \tag{3.10}$$

It is well known that every measurable function is a.e. approximatively continuous on $[0, T]$. Let E be the set of all points $t \in [0, T]$ where u^* is approximatively continuous. This means that for every $t \in E$ there exists a measurable subset $E_t \subset [0, T]$ having the property that t is a density point of E_t and u^* restricted to E_t is continuous at t. Let E^t be the set of all $s \in [0, t]$ which satisfy (3.9) where t is a fixed point in E. Clearly there exists at least one sequence $\{t_n\} \subset E^t \cap E_t$ convergent to t for $n \to \infty$. Hence $u^*(t) = \lim_{t_n \to t} u^*(t_n)$ where $B^* p^t(t_n) \in \partial h(u^*(t_n))$. Since $p^t(t_n)$ is weakly convergent to $p^t(t)$ and ∂h is strongly-weakly closed in $U \times U$ we may infer that

$$B^* p^t(t) \in \partial h(u^*(t)) \quad \text{for all } t \in E.$$

Along with (3.10) the latter implies (3.7) as claimed.

COROLLARY 1. *In Theorem* 1 *assume in addition that h is Gâteaux differentiable on U and the range* $R(B)$ *of B is a dense subset of H. Let p be any dual extremal arc associated with* (y^*, u^*) *by Proposition* 2. *Then*

$$p(t) \in -\partial \varphi(t, y^*(t)) \quad \textit{for all } t \in [0, T]. \tag{3.11}$$

Proof. Let p be a dual extremal arc associated with (y^*, u^*). By (2.35) and (3.9) we see that

$$p^t(s) = p(s) \quad \text{for all } s \in [0, t].$$

Because ∇h is single valued, the kernel $N(B^*) = \{0\}$ and p^t, p are weakly continuous on $[0, t]$, (3.11) follows by (3.10).

Let us assume now that besides (i)–(iv) and (a) the following conditions are satisfied:

(b) The function β is either globally Lipschitz or $D(A) \subset H^2(\Omega)$ and $N = 1$.

(c) $\psi \equiv 0$, h^* is Fréchet differentiable and ∇h^* is locally Lipschitz on U.

Let y_0 be arbitrary in $D(F)$ and let $t \in [0, T]$ be such that $s \to \varphi(s, y_0)$ is differentiable at $s = t$. Let $(y^t, u^t) \in W^{1,2}(t, T; H) \times L^2(t, T, U)$ be such that $y^t(s) = y(s, t, y_0, u^t)$ for $t \leqq s \leqq T$ and

$$\varphi(t, y_0) = \int_t^T (\Phi(y^t(s)) + h(u^t(s)))\, ds. \tag{3.12}$$

By Lemma 3 it follows that

$$\varphi(s, y^t(s)) = \int_s^T (\Phi(y^t(\tau)) + h(u^t(\tau)))\, d\tau, \qquad t \leqq s \leqq T. \tag{3.13}$$

By Proposition 3 the dual extremal arc p^t associated with (y^t, u^t) belongs to $W^{1,2}(t, T; H)$. Recalling that $u^t(s) = \nabla h^*(B^* p^t(s))$ a.e. $s \in]t, T[$ it follows from assumption (c) that $u^t \in W^{1,2}(t, T; U)$. This implies (see for instance [1, p. 133]) that $d^+/ds\, y^t(s)$ exists everywhere on $[t, T[$ and therefore by (3.13)

$$\frac{d^+}{ds}\varphi(s, y^t(s)) + \Phi(y^t(s)) + h(u^t(s)) = 0 \quad \text{for all } s \in [t, T[\tag{3.14}$$

because in virtue of the well-known conjugacy formula

$$h(v) + h^*(v^*) = \langle v, v^* \rangle \quad \text{for } v = \nabla h^*(v^*) \tag{3.15}$$

and assumption (c) the function $s \to h(u^t(s)) = h(\nabla h^*(B^* p^t(s)))$ is continuous on $[t, T]$.

On the other hand, we have

$$\begin{aligned} \frac{d^+}{ds}\varphi(t, y^t(t)) &= \varphi_t(t, y_0) + \lim_{\varepsilon \downarrow 0} (\varphi(t+\varepsilon, y^t(t+\varepsilon)) - \varphi(t+\varepsilon, y^t(t)))/\varepsilon \\ &= \varphi_t(t, y_0) + \lim_{\varepsilon \downarrow 0} (\zeta_\varepsilon, y^t(t+\varepsilon) - y^t(t))/\varepsilon, \end{aligned} \tag{3.16}$$

where $\zeta_\varepsilon(t) \in \partial\varphi(t+\varepsilon, \theta_\varepsilon)$ and θ_ε is a point in the open line segment between $y^t(t)$ and $y^t(t+\varepsilon)$. Here we have used a mean value theorem due to G. Lebourg (see [9]). Then if further we impose the condition

(d) The map $(t, y) \to \partial\varphi(t, y)$ is w^*-upper semicontinuous on $[0, T] \times H$.

It follows by (1.1), (3.14) that

$$\begin{aligned} \varphi_t(t, y_0) + (\zeta(t, y_0), B\, \nabla h^*(-B^* \eta(t, y_0)) - F y_0) + \Phi(y_0) \\ + h(\nabla h^*(-B^* \eta(t, y_0))) = 0, \end{aligned} \tag{3.17}$$

where

$$\zeta(t, y_0) \in \partial\varphi(t, y_0), \quad \eta(t, y_0) \in \partial\varphi(t, y_0). \tag{3.18}$$

Since for $y_0 \in D(F)$ the function $s \to \varphi(s, y_0)$ is a.e. differentiable on $[0, T]$, it follows that (3.17), (3.18) hold for all $y_0 \in D(F)$ and a.e. $t \in [0, T]$.

In this sense we may view φ as a solution to the Hamilton–Jacobi equation

$$\varphi_t(t, y) + h(\nabla h^*(-B^* \partial\varphi(t, y)) + (\partial\varphi(t, y), B \nabla h^*(-B^* \partial\varphi(t, y)) - Fy) + \Phi(y) = 0, \quad y \in D(F), \quad \text{a.e. } t \in [0, T] \tag{3.19}$$

with the Cauchy condition

$$\varphi(T, y) = 0 \quad \text{for all } y \text{ in } H. \tag{3.20}$$

In summary, we have proved:

PROPOSITION 4. *Under assumptions* (i)–(iv) *and* (a), (b), (c), (d) *the value function* φ *is the solution to the Cauchy problem* (3.19), (3.20).

If $\partial\varphi$ happens to be single-valued at (t, y) then by the conjugacy formula (3.15) we see that (3.19) can be written as

$$\varphi_t(t, y) - h^*(-B^* \partial\varphi(t, y)) - (Fy, \partial\varphi(t, y)) + \Phi(y) = 0. \tag{3.21}$$

Remark 2. Condition (d) is in particular satisfied if $\varphi(t, y)$ is convex in y (this is the case if Φ is convex and F linear) or if φ admits a (jointly) continuous derivative φ_y. The latter case occurs when the functions φ, h and β are continuously differentiable and the interval $[0, T]$ is sufficiently small. For a direct treatment of (3.19) in this case we refer the reader to [7]. In the special case when F is linear, and Φ is convex (3.21) has been studied in [5, p. 293]. For classical results on feedback synthesis we refer the reader to [13].

Remark 3. If $\beta \equiv 0$ then in Proposition 4 condition (c) can be weakened to:

(c′) h^* *is Gâteaux differentiable on* U. Indeed in this case $(d^+/ds)y^t(s)$ exists at $s = t$ and the conclusion follows by the same argument.

4. Control problems with infinite time horizon. We shall study here the control problem

$$\inf\left\{\int_0^\infty (\Phi(y(s, 0, y_0, u)) + h(u(s)))\, ds;\ u \in L^2_{\text{loc}}(R^+; U)\right\} = \varphi(y_0), \tag{4.1}$$

where $y(s, 0, y_0, u)$ is the solution to (1.1).

Besides (i)–(iv) and (a) the following assumptions will be imposed:

(j) $\Phi(0) = 0$, h is Gâteaux differentiable on U and $h(0) = 0$, $\nabla h(0) = 0$.

(jj) $\overline{R(B)} = H$.

(jjj) There exists $\omega > 0$ such that

$$(Fy, y) \geqq \omega |y|^2, \qquad y \in D(F). \tag{4.2}$$

LEMMA 5. *The function* φ *is everywhere finite on* H *and for every* $y_0 \in H$ *the infimum defining* $\varphi(y_0)$ *is attained. The function* φ *is locally Lipschitz on* H.

Proof. Let y_0 be fixed but arbitrary in H. By assumption (jjj) it follows that there exists $u \in L^2(R^+; U)$ such that $h(u) \in L^1(R^+)$ and $\Phi(y(t, 0, y_0, u)) \in L^1(R^+)$. Indeed it suffices to take $u = -B^* y$ where y is the solution to

$$y' + Fy + BB^* y = 0, \quad \text{a.e. } t > 0,$$
$$y(0) = y_0.$$

Hence $\varphi(y_0) < +\infty$. Arguing as in the proof of Lemma 3 it follows that the infimum in (4.1) is attained. Now for $|y_0| \leqq r$ it follows by (4.2) that

$$\varphi(y_0) \leqq \int_0^\infty \Phi(y(t, 0, y_0, 0))\, dt \leqq C_r.$$

Hence we may confine ourselves in (4.1) to those $u \in L^2_{\text{loc}}(R^+; U)$ which satisfy

$$\int_0^\infty h(u(t))\, dt \leqq C_r; \tag{4.3}$$

denote by $\mathcal{U}_r$ this subset of $L^2_{\text{loc}}(R^+; U)$. For every $u \in \mathcal{U}_r$, by (1.3), (4.1) and (4.3), we have

$$|y(t, 0, y_0, u)| \leqq e^{-\omega t}|y_0| + \|B\| \int_0^t e^{-\omega(t-s)}\|u(s)\|\, ds$$

$$\leqq e^{-\omega t} r + \|B\|(C_r + C) \leqq C_r^1.$$

Let y_0, $\tilde{y}_0$ be arbitrary in $\Sigma_r = \{y \in H; |y| \leqq r\}$. We have

$$\varphi(y_0) - \varphi(\tilde{y}_0) \leqq \int_0^\infty (\Phi(y(t, 0, y_0, \tilde{u}^*)) - \Phi(y(t, 0, \tilde{y}_0, \tilde{u}^*)))\, dt$$

where $\tilde{u}^*$ is such that

$$\varphi(\tilde{y}_0) = \int_0^\infty (\Phi(y(t, 0, \tilde{y}_0, \tilde{u}^*)) + h(\tilde{u}^*(t)))\, dt.$$

Since Φ is locally Lipschitz, by (4.2),

$$|y(t, 0, y_0, \tilde{u}^*) - y(t, 0, \tilde{y}_0, \tilde{u}^*)| \leqq e^{-\omega t}|y_0 - \tilde{y}_0|, \qquad t \geqq 0$$

and we find that

$$\varphi(y_0) - \varphi(\tilde{y}_0) \leqq L_r|y_0 - \tilde{y}_0|, \qquad y_0, \tilde{y}_0 \in \Sigma_r$$

thereby completing the proof.

In particular it follows by Lemma 5 that for every $y_0 \in H$ the control problem (4.1) has at least one optimal pair $(y^*, u^*) \in C(R^+; H) \times L^2_{\text{loc}}(R^+; U)$. Throughout the sequel we shall assume that $y_0 \in D(l)$, i.e., y_0 satisfies condition (2.2). Then $y^* = y(t, 0, y_0, u^*) \in W^{1,2}_{\text{loc}}(R^+; H) \cap L^2_{\text{loc}}(R^+; D(A))$.

THEOREM 2. *Let assumptions* (i)–(iv), (a) *and* (j)–(jjj) *be satisfied. If* $(y^*, u^*) \in W^{1,2}_{\text{loc}}(R^+; H) \times L^2_{\text{loc}}(R^+; U)$ *is an optimal pair in problem* (4.1) *then there exists* $p \in L^\infty(R^+; H) \cap L^2_{\text{loc}}(R^+; V) \cap AC_{\text{loc}}(R^+; Y^*) \cap C_w(R^+; H)$ *which satisfies the system*

$$p' - Ap - p\, \partial\beta(y^*) - \partial\Phi(y^*) \ni 0, \quad \text{a.e. } t > 0, \tag{4.4}$$

$$B^* p(t) = \partial h(u^*(t)), \qquad \text{a.e. } t > 0, \tag{4.5}$$

$$p(t) + \partial\varphi(y^*(t)) \ni 0 \qquad \text{for all } t \geqq 0. \tag{4.6}$$

Proof. Arguing as in the proof of Lemma 4 it follows that for every $t \geqq 0$,

$$\varphi(y_0) = \inf\left\{\int_0^t (\Phi(y(s, 0, y_0, u)) + h(u(s)))\, ds + \varphi(y(t)); u \in L^2(0, t; U)\right\}. \tag{4.7}$$

From Proposition 2 it follows that there exists

$$p^t \in L^\infty(0, t; H) \cap L^2(0, t; V) \cap AC([0, t]; Y^*) \cap C_w([0, t]; H)$$

such that

$$p_s^t - Ap^t - p^t\,\partial\beta(y^*) - \partial\Phi(y^*) \ni 0, \quad \text{in } Q_t = \Omega\times]0,t[, \tag{4.8}$$

$$p^t(t) + \partial\varphi(y^*(t)) \ni 0, \tag{4.9}$$

$$B^*p^t(s) = \partial h(u^*(s)), \qquad \text{a.e. } s\in]0,t[. \tag{4.10}$$

(Here the subscript s denotes as usual the derivative of p^t as a function from $[0,t]$ to Y^*.) Inasmuch as by assumptions (j) and (jj), $N(B^*)=\{0\}$ and ∂h is single valued, it follows by (4.10) that

$$p^t(s) = p^{\tilde t}(s) \quad \text{for } 0\leqq s\leqq t\leqq \tilde t. \tag{4.11}$$

Let $p: R^+\to H$ be the function defined by

$$p(s) = p^t(s) \quad \text{for } s\in[0,t]. \tag{4.12}$$

Obviously the function p satisfies (4.4), (4.5), (4.6). Since $h(u^*)\in L^1(R^+)$ it follows by condition (1.3) that $u^*\in L^2(R^+;U)+L^\infty(R^+;U)$ and so by (4.2) the function $t\to|y^*(t)|$ is bounded on R^+. Inasmuch as $\partial\varphi$ is locally bounded we see by (4.6) that $p\in L^\infty(R^+;H)$ thereby completing the proof of Theorem 2.

In particular, it follows by Theorem 2 that the set $\{(y,p)\in H\times H;\ p+\partial\varphi(y)\ni 0\}$ is an *invariant manifold* of the Hamiltonian system

$$\begin{aligned} &y_t + Ay + \beta(y) - B\,\partial h^*(B^*p) \ni 0, \qquad t\geqq 0,\\ &p_t - Ap - p\,\partial\beta(y) - \partial\Phi(y) \ni 0, \qquad t\geqq 0. \end{aligned} \tag{4.13}$$

(For related results in the case of linear systems of the form (1.1) and Φ convex we refer to [2].)

By Theorem 2 we may also conclude that every optimal control u^* for problem (4.1) is an optimal feedback control of the form

$$u^*(t) = \partial h^*(-B^*\,\partial\varphi(y^*(t))), \quad \text{a.e. } t>0. \tag{4.14}$$

We notice that by (4.7)

$$\varphi(y^*(t)) = \int_t^\infty (\Phi(y^*(s)) + h(u^*(s)))\,ds \quad \text{for all } t\geqq 0$$

and therefore

$$\frac{d}{dt}\varphi(y^*(t)) + \Phi(y^*(t)) + h(u^*(t)) = 0, \quad \text{a.e. } t>0.$$

Along with (4.5), (4.6) and the mean value theorem the latter yields

$$\begin{aligned} &(\zeta(t), Fy^*(t)) - (B^*\zeta(t), \partial h^*(-B^*\eta(t)))\\ &\qquad = \Phi(y^*(t)) + h(\partial h^*(-B^*\eta(t))), \quad \text{a.e. } t>0, \end{aligned} \tag{4.15}$$

where $\zeta(t)$, $\eta(t)\in\partial\varphi(y^*(t))$, a.e. $t\in R^+$. Thus there exists a dense subset $E\subset D(F)$ such that

$$\begin{aligned} &(\zeta(y_0), Fy_0) - (B^*\zeta(y_0), \partial h^*(-B^*\eta(y_0)))\\ &\qquad = \Phi(y_0) + h(\partial h^*(-B^*\eta(y_0))) \quad \text{for all } y_0 \text{ in } E, \end{aligned} \tag{4.16}$$

where $\zeta(y_0)$, $\eta(y_0)\in\partial\varphi(y_0)$.

Thus φ can be viewed as a generalized solution to the Hamilton–Jacobi equation (the Bellman equation)

$$(\partial\varphi(y), Fy) - (B^*\partial\varphi(y), \partial h^*(-B^*\partial\varphi(y)) = \Phi(y) + h(\partial h^*(-B^*\partial\varphi(y))). \tag{4.17}$$

If $\zeta(y) = \eta(y)$ then in view of the conjugacy formula, (4.17) can be written as

$$(\partial\varphi(y), Fy) + h^*(-B^*\partial\varphi(y)) = \Phi(y). \tag{4.18}$$

For a direct treatment of this equation in the convex case we refer to [6].

Remark 4. Let us assume that Φ is Fréchet differentiable and that condition (b) is satisfied. Then by Proposition 3 the dual extremal arc p arising in system (4.4), (4.6) is locally absolutely continuous as a function from $]0, +\infty[$ to H. Then multiplying (1.1) by p', (4.4) by $-y'$ and adding the results we get

$$\frac{d}{dt}(p(t), Fy^*(t)) = \frac{d}{dt}(h^*(p(t)) - \Phi(y^*(t))), \quad \text{a.e. } t > 0.$$

Hence

$$(p(t), Fy^*(t)) - h^*(p(t)) + \Phi(y^*(t)) = C \quad \text{for all } t \geqq 0. \tag{4.19}$$

5. Approximating feedback controls for the time optimal problem. Consider the time optimal control problem

$$\inf\{T;\ y'(t) + Fy(t) = u(t);\ |u(t)| \leqq 1, \text{ a.e. } t \in [0, T];\ y(0) = y_0,\ y(T) = 0\} = T(y_0), \tag{5.1}$$

where F is defined by (1.7) and (1.8) and y_0 is in H. Throughout this section we shall assume that hypotheses (i), (a) and (jjj) are satisfied.

The value $T(y_0)$ of problem (5.1) is called the *optimal time* corresponding to y_0.

Next we shall prove a null controllability result.

LEMMA 6. *For every $y_0 \in H$ there exists at least one control u^* such that $y(T(y_0), 0, y_0, u^*) = 0$.*

Proof. Consider the Cauchy problem

$$y'(t) + Fy(t) + \operatorname{sgn} y(t) \ni 0, \quad \text{a.e. } t > 0, \qquad y(0) = y_0 \tag{5.2}$$

where

$$\operatorname{sgn} y = y/|y| \quad \text{for } y \neq 0, \qquad \operatorname{sgn} 0 = \{z \in H;\ |z| \leqq 1\}.$$

The operator $y \to Fy + \operatorname{sgn} y$ is maximal monotone in $H \times H$ as a sum of two maximal monotone operators satisfying the condition $D(F) \cap \operatorname{int} D(\operatorname{sgn}) \neq \varnothing$ (see [1, p. 46]). As a matter of fact this operator is the subdifferential of the function $y \to l(y) + |y|$. So, as noticed earlier, for every $y_0 \in H$ the Cauchy problem (5.2) has a unique solution $y \in W^{1,2}_{\text{loc}}(\delta, +\infty; H)$ for every $\delta > 0$ and continuous from $[0, +\infty[$ to H. Taking the scalar product of (5.2) by $y(t)$ we get

$$(|y(t)|^2)' + 2|y(t)| \leqq 0, \quad \text{a.e. } t > 0.$$

Hence

$$|y(t)| \leqq |y_0| - t \quad \text{for all } t \geqq 0.$$

The latter implies that $y(t)=0$ for $t \geqq |y_0|$. We have therefore proved that there exists at least one time $T>0$ and a control $u(t)=-y'(t)-Fy(t) \in \operatorname{sgn} y(t)$, a.e. $t \in]0, T[$ such that $y(T, 0, y_0, u)=0$. Hence $T(y_0)<+\infty$ for all $y_0 \in H$ and there exist the sequences $\{T_n\} \to T(y_0)$ and $u_n \in L^\infty(0, T_n; H)$ such that $|u_n(t)| \leqq 1$ a.e. $t \in [0, T_n]$ and

$$y_n'(t)+Fy_n(t)=u_n(t), \quad \text{a.e. } t \in [0, T_n],$$

$$y_n(0)=y_0, \qquad y_n(T_n)=0.$$

Without loss of generality we may assume that the u_n are defined on the whole positive half axis and

$$u_n \to u^* \quad \text{weak star in } L^\infty(R^+; H).$$

Using hypothesis (i) and condition (4.2) we see that for every $\delta>0$

$$\int_\delta^{T_n} |y_n'(t)|^2\, dt + \|y_n(t)\|_V^2 \leqq C \quad \text{for all } n \text{ and } t \in [\delta, T_n]$$

and

$$\int_0^{T_n} \|y_n(t)\|_V^2\, dt \leqq C \quad \text{for all } n.$$

Thus selecting a further subsequence we may assume that

$$y_n \to y \quad \text{weakly in } L^2(0, T^*; V),$$

$$y_n' \to y' \quad \text{weakly in } L^2(\delta, T^*; H),$$

and by the Arzela–Ascoli theorem

$$y_n \to y \quad \text{in } C([\delta, T^*]; H),$$

where $T^*=T(y_0)$. In particular we infer that $y(T^*)=0$. On the other hand, since the operator $y \to Fy$ is maximal monotone in every $L^2(\delta; T^*; H)$ it is strongly-weakly closed in this space. Hence

$$u^*-y'=Fy \quad \text{a.e. } t \in [\delta, T^*] \quad \text{for all } \delta>0.$$

We have therefore proved that

$$y'+Fy=u^*, \quad \text{a.e. } t \in]0, T^*[,$$

$$y(0)=y_0, \qquad y(T^*)=0,$$

thereby completing the proof of Lemma 6.

The above argument can be used to prove the null controllability of general systems of the form

$$y'+Fy=u,$$

where F is a maximal monotone operator in a Hilbert space H. Such a control is called a *time optimal control* of system (1.1). For linear systems there exists a number of significant results on the time optimal control problem (see for instance [12], [15]). Here we shall use a different approach which relies on the results of § 4.

Let $\pi \in C^\infty(R^+)$ be a given function such that $0 \leqq \pi \leqq 1$, $\pi' \geqq 0$ and

$$\pi(r)=\begin{cases} 1 & \text{for } r \geqq 2, \\ 0 & \text{for } 0 \leqq r \leqq 1. \end{cases} \tag{5.3}$$

Let $g^\varepsilon : H \to R$ be defined by

$$g^\varepsilon(y) = \pi(|y|^2/\varepsilon^2) \quad \text{for } y \in H. \tag{5.4}$$

We set

$$G^\varepsilon = \nabla g^\varepsilon, \qquad h^\varepsilon(u) = (2\varepsilon)^{-1}((|u|-1)^+)^2 \quad \text{for } u \in H \tag{5.5}$$

and define the function $\varphi^\varepsilon : H \to R$,

$$\varphi^\varepsilon(y_0) = \inf\left\{\int_0^\infty (g^\varepsilon(y(s, 0, y_0, u)) + h^\varepsilon(u(s)))\, ds;\ u \in L^2_{\text{loc}}(R^+; H)\right\}. \tag{5.6}$$

Here $y(s, 0, y_0, u)$ denotes the solution to (1.1)′ where $B \equiv I$.

Let us assume that $y_0 \in D(l)$. Since assumptions (i)–(iv), (a) and (j)–(jjj) are trivially satisfied with $\Phi = g^\varepsilon$ and $h = h^\varepsilon$, for every $\varepsilon > 0$ problem (5.6) has at least one solution $(y^\varepsilon, u^\varepsilon) \in W^{1,2}_{\text{loc}}(R^+; H) \times L^2_{\text{loc}}(R^+; H)$ and by Theorem 2 there exists $p^\varepsilon \in L^\infty(R^+; H) \cap L^2_{\text{loc}}(R^+; V) \cap AC_{\text{loc}}(R^+; Y^*) \cap C_w(R^+; H)$ which satisfies, along with y^ε and u^ε, the system

$$\begin{aligned} &y^\varepsilon_t + Fy^\varepsilon(t) = u^\varepsilon, \quad \text{a.e. } t > 0, \\ &p^\varepsilon_t - Ap^\varepsilon - p^\varepsilon\, \partial\beta(y^\varepsilon) \ni G^\varepsilon(y^\varepsilon), \quad \text{a.e. } t > 0, \end{aligned} \tag{5.7}$$

$$p^\varepsilon(t) = \nabla h^\varepsilon(u^\varepsilon(t)) = \begin{cases} \dfrac{u^\varepsilon(t)}{\varepsilon|u^\varepsilon(t)|}(|u^\varepsilon(t)|-1) & \text{if } |u^\varepsilon(t)| \geqq 1, \\ 0 & \text{if } |u^\varepsilon(t)| < 1, \end{cases} \tag{5.8}$$

$$p^\varepsilon(t) + \partial\varphi^\varepsilon(y^\varepsilon(t)) \ni 0 \quad \text{for all } t \geqq 0. \tag{5.9}$$

By (5.5) and (5.8) it follows that

$$(h^\varepsilon)^*(p) = |p| + \frac{\varepsilon}{2}|p|^2 \quad \text{for all } p \in H \tag{5.10}$$

and therefore

$$u^\varepsilon(t) = \operatorname{sgn} p^\varepsilon(t) + \varepsilon p^\varepsilon(t), \quad \text{a.e. } t > 0.$$

Along with (5.9) the latter implies that

$$u = -\operatorname{sgn} \partial\varphi^\varepsilon(y) - \varepsilon\, \partial\varphi^\varepsilon(y) \tag{5.11}$$

is an optimal feedback control for problem (5.6). As a matter of fact any optimal control u^ε to problem (5.6) can be expressed as a function of optimal state y^ε by the formula (5.11), i.e.,

$$u^\varepsilon(t) \in -\operatorname{sgn} \partial\varphi^\varepsilon(y^\varepsilon(t)) - \varepsilon\, \partial\varphi^\varepsilon(y^\varepsilon(t)), \quad \text{a.e. } t > 0. \tag{5.12}$$

Here the multivalued mapping $\operatorname{sgn} : H \to H$ is defined as above, i.e., $\operatorname{sgn} y = y/|y|$ for $y \neq 0$ and $\operatorname{sgn} 0 = \{z \in H;\ |z| \leqq 1\}$.

By (4.16) and (5.10) we see that φ^ε is the solution to the stationary Hamilton–Jacobi equation

$$\frac{\varepsilon}{2}|\partial\varphi^\varepsilon(y)|^2 + |\partial\varphi^\varepsilon(y)| + (Fy, \partial\varphi^\varepsilon(y)) = g^\varepsilon(y), \tag{5.13}$$

i.e., there exist $\zeta_\varepsilon \subset \partial\varphi^\varepsilon$ and $\eta_\varepsilon \subset \partial\varphi^\varepsilon$ such that

$$(\zeta_\varepsilon(y), Fy + \operatorname{sgn}\eta_\varepsilon(y) + \varepsilon\eta_\varepsilon(y)) = g^\varepsilon(y) + \frac{\varepsilon}{2}|\eta_\varepsilon(y)|^2$$

for all y in a dense subset of $D(F)$.

Let us now assume that β satisfies condition (b). Then by Remark 4 (see (4.19)) we have

$$(p^\varepsilon(t), Fy^\varepsilon(t)) - |p^\varepsilon(t)| - \frac{\varepsilon}{2}|p^\varepsilon(t)|^2 + g^\varepsilon(y^\varepsilon(t)) = C, \quad \text{a.e. } t > 0. \tag{5.14}$$

Since $u^\varepsilon \in L^2(R^+; H) + L^\infty(R^+; H)$ it follows by (4.2) and the first equation in (5.7) that

$$|y^\varepsilon(t)| \leqq C, \qquad \int_0^t |Fy^\varepsilon(\tau)|^2\, d\tau \leqq Ct \quad \text{for } t \geqq 0$$

and by (5.8) we see that $p^\varepsilon \in L^2(R^+; H)$. Since $g^\varepsilon(y_\varepsilon) \in L^1(R^+)$ the latter estimate implies that there exists at least one sequence $\{t_n\} \to +\infty$ such that

$$(p^\varepsilon(t_n), F(y^\varepsilon(t_n)) - (p^\varepsilon(t_n)| - \frac{\varepsilon}{2}|p^\varepsilon(t_n)|^2 + g^\varepsilon(y^\varepsilon(t_n)) \to 0.$$

Hence $C = 0$ in (5.14) and so φ^ε satisfies (5.13) in the following stronger sense: *there exists a single valued section* $\zeta^\varepsilon \subset \partial\varphi^\varepsilon$ such that

$$\frac{\varepsilon}{2}|\zeta^\varepsilon(y)|^2 + |\zeta^\varepsilon(y)| + (Fy, \zeta^\varepsilon(y)) = g^\varepsilon(y) \quad \forall y \in E, \tag{5.15}$$

where E is a dense subset of $D(F)$.

The relevance of the function φ^ε in the time optimal control problem (5.1) becomes clear in Theorem 3 below.

THEOREM 3. *Let $y_0 \in H$ be given. Then*

$$\lim_{\varepsilon \to 0} \varphi^\varepsilon(y_0) = T(y_0) \tag{5.16}$$

and on some subsequence (again denoted ε) we have

$$u^\varepsilon \to u^* \quad \textit{weakly in } L^2_{\text{loc}}(R^+; H), \tag{5.17}$$

$$y^\varepsilon \to y^* \quad \textit{strongly in every } C([0, T]; H), \tag{5.18}$$

where u^* is a time optimal control and $y^*(t) = y(t, 0, y_0, u^*)$ is the corresponding state.

Proof. For the sake of simplicity we shall assume that $y_0 \in D(l)$ the proof in general case is completely analogous. Let $T^* = T(y_0)$ be the optimal time and let (y_1^*, u_1^*) be an optimal pair for problem (5.1). We have

$$\varphi^\varepsilon(y_0) \leqq \int_0^\infty (g^\varepsilon(y_1^*(t)) + h^\varepsilon(u_1^*(t)))\, dt$$

$$= \int_0^{T^*} (g^\varepsilon(y_1^*(t)) + h^\varepsilon(u_1^*(t)))\, dt = \int_0^{T^*} g^\varepsilon(y_1^*)\, dt.$$

Hence

$$\varphi^\varepsilon(y_0) \leqq T^* \quad \text{for all } \varepsilon > 0. \tag{5.19}$$

In particular it follows that $\{u^\varepsilon\}$ is bounded in every $L^2(0, T; H)$ and $\{y^\varepsilon\}$ is compact in every $C([0, T]; H)$. Hence there exist $(y^*, u^*) \in W^{1,2}_{\text{loc}}(R^+; H) \times L^2_{\text{loc}}(R^+; H)$ such

that $y^*(t) = y(t, 0, y_0, u^*)$ and

$$\begin{aligned} &u^\varepsilon \to u^* \quad \text{weakly in } L^2_{\text{loc}}(R^+; H), \\ &y^\varepsilon(t) \to y^*(t) \quad \text{uniformly on every } [0, T]. \end{aligned} \tag{5.20}$$

By (5.6) and (5.19) we see that

$$\limsup_{\varepsilon \to 0} \int_0^T g^\varepsilon(y^\varepsilon(t))\, dt \leqq T^* \quad \text{for every } T > 0. \tag{5.21}$$

This implies that there exist a sequence $\{\varepsilon_n\} \to 0$ and $T_0 > 0$ independent of n such that

$$|y^{\varepsilon_n}(t)| \leqq 2\varepsilon_n \quad \text{for } t \geqq T_0. \tag{5.22}$$

For, otherwise for every sequence $\{\varepsilon_n\}$ convergent to zero there would exist $t_n \to \infty$ such that $|y^{\varepsilon_n}(t_n)| > 2\varepsilon_n$ for all n. Let $\varepsilon_n = n^{-1/2}$. Then by an easy calculation involving (1.1)′ it would follow that

$$|y^{\varepsilon_n}(t_n)| \leqq |y^{\varepsilon_n}(t)| + \int_t^{t_n} |u^{\varepsilon_n}(s)|\, ds \quad \text{for } t \leqq t_n. \tag{5.23}$$

Since $(2\varepsilon_n)^{-1} \int_0^\infty ((|u^{\varepsilon_n}(t)| - 1)^+)^2\, dt \leqq C$, by (5.23) it follows that

$$|y^{\varepsilon_n}(t_n)| \leqq |y^{\varepsilon_n}(t)| + |t - t_n| + C(2\varepsilon_n |t - t_n|)^{1/2} \quad \text{for } t \leqq t_n.$$

Hence for n sufficiently large,

$$|y^{\varepsilon_n}(t)| \geqq \sqrt{2}\varepsilon_n \quad \text{for } t \in [t_n - \delta_n, t_n],$$

where $\delta_n = Cn^{-1}$. This would imply that $\lim_{n\to\infty} m\{t; |y^{\varepsilon_n}(t)| \geqq \sqrt{2}\varepsilon_n\} = +\infty$ (m is the Lebesgue measure) contrary to (5.21).

By (5.22) it follows that $y^*(t) = 0$ for $t \geqq T_0$. Let $\tilde{T} = \inf\{T; y^*(T) = 0\}$. We will show that $\tilde{T} = T^*$. To this end for any $\varepsilon > 0$ consider the set $E_\varepsilon = \{t \in [0, \tilde{T}]; |y^\varepsilon(t)| \geqq \sqrt{2}\varepsilon\}$. By (5.21) it follows that

$$\limsup_{\varepsilon \to 0} m(E_\varepsilon) \leqq T^*. \tag{5.24}$$

On the other hand, $\lim_{\varepsilon\to 0} m(E_\varepsilon) = \tilde{T}$, for otherwise there would exist $\delta > 0$ and $\varepsilon_n \to 0$ such that $m(E_{\varepsilon_n}) \leqq \tilde{T} - \delta$. In other words, there would exist a sequence of measurable subsets $A_n \subset [0, \tilde{T}]$ such that $m(A_n) \geqq \delta$ and $|y^{\varepsilon_n}(t)| \leqq \sqrt{2}\varepsilon_n$ for $t \in A_n$. In virtue of (5.20) this would imply that

$$|y^*(t)| \leqq \sqrt{2}\varepsilon_n + \nu_n \quad \text{for } t \in A_n,$$

where $\nu_n \to 0$. On the other hand, since $y^*(t) \neq 0$ for $t \in [0, \tilde{T}]$, $m\{t; |y^*(t)| \leqq \sqrt{2}\varepsilon_n + \nu_n\} \to 0$ for $n \to +\infty$. The contradiction we arrived at shows that indeed $\lim_{\varepsilon\to 0} m(E_\varepsilon) = \tilde{T}$. Along with (5.24) the latter implies that $\tilde{T} = T^*$ as claimed. Thus u^* is a time optimal control and the proof of Theorem 3 is complete.

Remark 5. Letting ε tend to zero in (5.12) we see that formally, T can be regarded as solution to the Bellman equation

$$(Fy, \partial T(y)) + |\partial T(y)| = 1. \tag{5.25}$$

Since T is locally Lipschitz on H we suspect that T is indeed a solution to (5.25) but we have failed to prove this. Anyway in virtue of Theorem 3 the feedback law (5.11) can be viewed as an *approximating feedback control* for problem (5.1). The implementation of this suboptimal feedback control requires a numerical procedure for the calculation of φ^ε either from (5.13) or directly from formula (5.6).

Now we shall study the following optimal control problem: Minimize

$$\frac{1}{2}\int_0^T |u(t)|^2\,dt+\alpha T \tag{5.26}$$

over all $(y, u)\in W^{1,2}_{\mathrm{loc}}(R^+; H)\times L^2_{\mathrm{loc}}(R^+; H)$ subject to

$$\begin{aligned} &y'(t)+Fy(t)=u(t), \quad \text{a.e. } t>0,\\ &y(0)=y_0, \quad y(T)=0. \end{aligned} \tag{5.27}$$

Here α is a positive constant.

We associate with (5.26) the approximating problem

$$\inf\int_0^\infty \left(\alpha(g^\varepsilon(y(t, 0, y_0, u)))+\frac{1}{2}|u(t)|^2\right)dt=\psi^\varepsilon(y_0). \tag{5.28}$$

Let $(y_\varepsilon, u_\varepsilon)\in W^{1,2}(R^+; H)\times L^2(R^+; H)$ be an optimal pair for problem (5.28). By Theorem 2 there exists $p_\varepsilon\in L^2_{\mathrm{loc}}(R^+; V)\cap L^2(R^+; H)\cap L^\infty(R^+; H)\cap AC_{\mathrm{loc}}(R^+; Y^*)$ which satisfies the system

$$\begin{aligned} &y_\varepsilon'+Fy_\varepsilon=p_\varepsilon, \quad \text{a.e. } t>0,\\ &p_\varepsilon'-Ap_\varepsilon-p_\varepsilon\,\partial\beta(y_\varepsilon)=\alpha G^\varepsilon(y_\varepsilon), \quad \text{a.e. } t>0, \end{aligned} \tag{5.29}$$

$$p_\varepsilon(t)+\partial\psi^\varepsilon(y_\varepsilon(t))\ni 0 \quad \text{for all } t\geqq 0, \tag{5.30}$$

$$u_\varepsilon=p_\varepsilon. \tag{5.31}$$

As noticed earlier, $\lim_{t\to\infty} y_\varepsilon(t)=0$ in H. By (4.16) we see that ψ^ε is the solution to the Hamilton–Jacobi equation

$$(Fy, \partial\psi^\varepsilon(y))+\tfrac{1}{2}|\partial\psi^\varepsilon(y)|^2=\alpha g^\varepsilon(y) \tag{5.32}$$

and

$$u=-\partial\psi^\varepsilon(y) \tag{5.33}$$

is an *optimal feedback control* for problem (5.28).

THEOREM 4. *For every* $y_0\in H$,

$$\lim_{\varepsilon\to 0}\psi^\varepsilon(y_0)=\psi_0(y_0) \tag{5.34}$$

and on some sequence $\varepsilon_n\to 0$

$$u_{\varepsilon_n}\to u_1^* \quad \textit{strongly in } L^2(R^+; H), \tag{5.35}$$

$$y_{\varepsilon_n}\to y_1^* \quad \textit{uniformly in } H \textit{ on } R^+, \tag{5.36}$$

where (y_1^*, u_1^*) *is an optimal pair of problem* (5.26).

Here $\psi_0(y_0)$ is the optimal value of problem (5.26). The proof which is essentially the same as that of Theorem 3 will be outlined only.

We have

$$\begin{aligned} \psi^\varepsilon(y_0)&=\int_0^\infty (\alpha g^\varepsilon(y_\varepsilon(t))+\tfrac{1}{2}|u_\varepsilon(t)|^2)\,dt\\ &\leqq\int_0^\infty (\alpha g^\varepsilon(\tilde{y}_1^*)+\tfrac{1}{2}|\tilde{u}_1^*|^2)\,dt, \end{aligned} \tag{5.37}$$

where $(\tilde{y}_1^*, \tilde{u}_1^*)$ is any optimal pair in problem (5.26). Selecting a subsequence if necessary, we may assume that for $\varepsilon \to 0$

$$\begin{aligned} &u_\varepsilon \to u_1^* \quad \text{weakly in } L^2(R^+; H), \\ &y_\varepsilon \to y_1^* \quad \text{uniformly in } H \text{ on every } [0, T]. \end{aligned} \tag{5.38}$$

By (5.37) it follows that

$$\limsup_{\varepsilon \to 0} \psi^\varepsilon(y_0) \leqq \psi_0(y_0). \tag{5.39}$$

As in the proof of Theorem 3 the latter implies that $y_1^*(t) = 0$ for $t \geqq T_0$. Let $\tilde{T} = \inf\{T;\ y_1^*(T) = 0$ and $E_\varepsilon = \{t \in [0, \tilde{T}];\ |y_\varepsilon(t)| \geqq \sqrt{2}\varepsilon\}$. Then $\lim_{\varepsilon \to 0} m(E_\varepsilon) = \tilde{T}$ and therefore by (5.39),

$$\alpha \tilde{T} + \frac{1}{2} \int_0^{\tilde{T}} |u_1^*(t)|^2 \, dt \leqq \psi_0(y_0).$$

Hence (y_1^*, u_1^*) is an optimal pair in problem (5.26). In (5.37) we take $\tilde{y}_1^* = y_1^*$, $\tilde{u}_1^* = u_1^*$ to get that

$$\limsup_{\varepsilon \to 0} \int_0^\infty |u_\varepsilon(t)|^2 \, dt \leqq \int_0^\infty |u_1^*(t)|^2 \, dt.$$

This yields (5.34), (5.35) and (5.36) as claimed.

Now we point out some immediate consequences of Theorem 4. First we notice that in virtue of (5.32) and (5.34) we may view ψ_0 as a generalized solution to the equation (the Bellman equation for the free-time optimal control problem (5.26))

$$(Fy, \partial\psi_0(y)) + \tfrac{1}{2} |\partial\psi_0(y)|^2 = \alpha. \tag{5.40}$$

We notice that if β *satisfies condition* (b) then by (5.29) and (5.31) we have (see (4.19))

$$-(Fy_\varepsilon(t), u_\varepsilon(t)) + \tfrac{1}{2}|u_\varepsilon(t)|^2 = \alpha g^\varepsilon(y_\varepsilon(t)), \quad \text{a.e. } t > 0. \tag{5.41}$$

Inasmuch as $g^\varepsilon(y_\varepsilon) \to 1$, a.e. $t \in [0, T_1^*]$, letting ε tend to zero in (5.41) we see that

$$-(Fy_1^*(t), u_1^*(t)) + \tfrac{1}{2}|u_1^*(t)|^2 = \alpha, \quad \text{a.e. } t \in]0, T_1^*[. \tag{5.42}$$

Here T_1^* is an optimal time in problem (5.26).

Now we shall use Theorem 4 to derive a maximum principle type result for problem (5.26).

COROLLARY 2. *Let u_1^* be the optimal control provided by Theorem* 4. *Then there exists $p \in L^2(0, T_1^*; V) \cap C_w([0, T_1^*]; H) \cap AC([0, T_1^*]; Y^*)$ which satisfies along with y_1^* and u_1^* the system*

$$\begin{aligned} &y_1^{*\prime} + Fy_1^* = u_1^*, \quad \textit{a.e. } t \in]0, T_1^*[, \\ &p' - Ap - p\ \partial\beta(y_1^*) \ni 0, \quad \textit{a.e. } t \in]0, T_1^*[, \end{aligned} \tag{5.43}$$

$$p = u_1^*, \quad \textit{a.e. } t \in]0, T_1^*[. \tag{5.44}$$

Proof. By (5.29), (5.36) and the definition of G^ε we see that for every $\delta > 0$ there exists $\varepsilon_0(\delta) > 0$ such that for all $0 < \varepsilon < \varepsilon_0(\delta)$

$$p_\varepsilon' - Ap_\varepsilon - p_\varepsilon\ \partial\beta(y_\varepsilon) \ni 0, \quad \text{a.e. } t \in]0, T_1^* - \delta[. \tag{5.45}$$

Since by (5.35), $\{p_\varepsilon\}$ is strongly convergent in $L^2(0, T_1^*; H)$ to $u_1^* = p$ and

$$\beta(y_\varepsilon) \to \beta(y_1^*) \quad \text{weakly in } L^2(0, T_1^*; H)$$

it follows by (2.22) and (5.45) that $\{p_\varepsilon\}$ is bounded in $L^2(0, T_1^* - \delta; V) \cap L^\infty(0, T_1^* - \delta; H)$ and $\{p'_\varepsilon\}$ is bounded in $L^1(0, T_1^* - \delta; Y^*)$. Then proceeding as in the proof of Proposition 2 we infer that p satisfies (5.43), (5.44).

Remark 6. Coming back to (5.7)–(5.9) let us assume that for every $\delta \in]0, T^*[$ there exists $t_0 \in [T^* - \delta, T^*]$ such that $\{|p^\varepsilon(t_0)|\}$ is bounded for $\varepsilon \to 0$ (this assumption is satisfied in several notable cases which will be discussed elsewhere). Since by (5.7) we see that for ε sufficiently small, $p^{\varepsilon'} - Ap^\varepsilon - p^\varepsilon \ni 0$, a.e. $t \in]0, T^* - \delta[$ we may conclude as in the proof of Corollary 2 that there exists $p_\delta \in L^2(0, T^* - \delta; V) \cap L^\infty(0, T^* - \delta; H) \cap AC([0, T^* - \delta]; Y^*)$ satisfying

$$\begin{aligned} &p'_\delta - Ap_\delta - p_\delta \ni 0, \quad \text{a.e. } t \in]0, T^* - \delta[, \\ &u^* = \operatorname{sgn} p_\delta, \qquad p_\delta \neq 0, \quad \text{a.e. } t \in]0, T^* - \delta[. \end{aligned} \tag{5.46}$$

6. Some remarks on feedback synthesis of variational inequalities. Consider the control problem with cost (1.2) and state system

$$\begin{aligned} &y_t + Ay + \gamma(y) \ni Bu, \quad \text{a.e. } t \in]0, T[, \\ &y(0) = y_0, \end{aligned} \tag{6.1}$$

where the operators A and B satisfy assumptions (i), (ii) and γ is a multivalued maximal monotone graph in $R \times R$ such that $\gamma(0) \ni 0$. This class of distributed parameter systems includes the control of several important variational inequalities of parabolic type. For instance the well-known "obstacle problem" can be written in the form (6.1) where A is the elliptic operator defined by (1.16) and γ is defined by

$$\gamma(r) = 0 \quad \text{for } r > 0, \qquad \gamma(0) = R^-, \qquad \gamma(r) = \varnothing \quad \text{for } r < 0.$$

Denote by $D(\gamma)$ the domain of γ, i.e., $D(\gamma) = \{r \in R; \gamma(r) \neq \varnothing\}$ and by D the set $\{y \in H; y(x) \in D(\gamma) \text{ a.e. } x \in \Omega\}$. Let $y(\cdot, t, y_0, u)$ be the solution to (6.1) on $[0, t]$ with initial value condition $y(t) = y_0$. We notice that relations (1.11)–(1.13) remain true for $y_0 \in \bar{D}$. Let $\varphi : [0, T] \times \bar{D} \to R$ be the optimal value function corresponding to control problem (6.1), (1.2), i.e., the function

$$\begin{aligned} \varphi(t, y_0) = \inf \Big\{ &\int_t^T (\Phi(y(s, t, y_0, u)) + h(u(s)))\, ds \\ &+ \psi(y(T, t, y_0, u)); u \in L^2(t, T; U) \Big\} \end{aligned} \tag{6.2}$$

where Φ, h and ψ satisfy assumption (iv).

Let $\varphi^\varepsilon : [0, T] \times H \to R$ be the optimal value function corresponding to control system

$$\begin{aligned} &y_t + Ay + \gamma_\varepsilon(y) = Bu, \quad \text{a.e. } t \in]0, T[, \\ &y(0) = y_0, \end{aligned} \tag{6.3}$$

with the functional cost (1.2). Here $\gamma_\varepsilon = \varepsilon^{-1}(1 - (1 + \varepsilon\gamma)^{-1})$ is the Yosida approximation of γ.

It should be noted that γ_ε is Lipschitz and therefore all the results of §§ 2 and 3 are applicable for the control system (6.3) with the cost functional (1.2).

Let $(y_\varepsilon, u_\varepsilon)$ be an optimal pair in the problem (6.3), (1.2). We have

PROPOSITION 5. *Under the above assumptions,*

$$\lim_{\varepsilon \to 0} \varphi^\varepsilon(t, y_0) = \varphi(t, y_0) \quad \textit{for all } (t, y_0) \in [0, T] \times \bar{D} \tag{6.4}$$

and on some subsequence $\varepsilon_n \to 0$

$$u_{\varepsilon_n} \to u^* \quad \textit{weakly in } L^2(0, T; U), \tag{6.5}$$

$$y_{\varepsilon_n} \to y^* \quad \textit{strongly in } L^2(0, T; H), \tag{6.6}$$

where (y^*, u^*) *is an optimal pair for the problem* (6.1), (1.2).

The proof is standard and it is left to the reader.

Along with Theorem 1 and its consequences, Proposition 5 amounts to saying that the feedback law

$$u = \partial h^*(-B^* \partial \varphi^\varepsilon(t, y)), \qquad t \in [0, T], \quad y \in H \tag{6.7}$$

is an approximating feedback control for problem (6.1), (6.2). In the same manner we may use Theorems 2, 3 and 4 to construct approximating feedback controls for the infinite time horizon problem and the optimal time problem associated with state system (6.1).

REFERENCES

[1] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Groningen, 1976.

[2] ———, *Convex control problems and Hamiltonian systems on infinite intervals*, this Journal, 16 (1978), pp. 687–702.

[3] ———, *Necessary conditions for distributed control problems governed by parabolic variational inequalities*, this Journal, 19 (1981), pp. 64–86.

[4] ———, *Boundary control problems with nonlinear state equation*, this Journal, 20 (1982), pp. 125–143.

[5] V. BARBU AND TH. PRECUPANU, *Convexity and Optimization in Banach Spaces*, Noordhoff and Sijthoff, Groningen, 1978.

[6] V. BARBU AND G. DA PRATO, *Existence and approximation for stationary Hamilton–Jacobi equations*, J. Nonlinear Anal., 6 (1981), pp. 1213–1224.

[7] ———, *Hamilton–Jacobi equations and synthesis of nonlinear control processes in Hilbert spaces*, J. Differential Equations, to appear.

[8] H. BREZIS, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[9] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[10] ———, *Generalized gradients of Lipschitz functionals*, Advances in Math., 40 (1981), pp. 52–67.

[11] G. DUVAUT AND J. L. LIONS, *Inequalities in Mechanics and Physics*, Springer-Verlag, New York, Heidelberg, Berlin, 1976.

[12] H. O. FATTORINI, *The time optimal control problem in Banach space*, Applied Math. Optim., Vol. 1 (1974), pp. 163–188.

[13] W. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, Berlin, Heidelberg, New York, 1975.

[14] J. L. LIONS, *Quelques methodes de resolution des problèmes aux limites non lineaires*, Dunod Gauthier-Villars, Paris, 1969.

[15] J. L. LIONS, *Optimal control of systems governed by partial differential equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1971.

[16] R. T. ROCKAFELLAR, *Directionally Lipschitzian functions and subdifferential calculus*, Proc. London Math. Soc., 39 (1979), pp. 331–355.

[17] D. TIBA, *Necessary conditions for distributed control problems with nonlinear state equation*, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0005 $01.25/0

ESTIMATION OF DELAYS AND OTHER PARAMETERS IN NONLINEAR FUNCTIONAL DIFFERENTIAL EQUATIONS*

H. T. BANKS† AND P. K. DANIEL LAMM‡

Abstract. We discuss a spline-based approximation scheme for nonlinear nonautonomous delay differential equations. Convergence results (using dissipative type estimates on the underlying nonlinear operators) are given in the context of parameter estimation problems which include estimation of multiple delays and initial data as well as the usual coefficient-type parameters. A brief summary of some of our related numerical findings is also given.

Key words. delay equations, parameter estimation, splines, numerical algorithms

1. Introduction. In [6] spline approximation ideas are developed in the context of numerical algorithms for the solution of functional differential equations (FDE). The theoretical framework is based on a functional analytic formulation (in an appropriately chosen Hilbert space Z) of Ritz–Galerkin type ideas where one approximates on finite-dimensional subspaces Z^N the underlying linear solution semigroup $T(t)$ (with infinitesimal generator A) for the FDE by linear semigroups $T^N(t)$ (with infinitesimal generators $A^N = P^N A P^N$, where P^N is the orthogonal projection of Z onto Z^N). These ideas were subsequently ([3], [4]) used in developing numerical schemes for parameter estimation and optimal control problems. The fundamental theoretical tool employed is the Trotter–Kato theorem (a functional analytic formulation of the Lax equivalence theorem: stability plus consistency yields convergence of approximation schemes) for linear semigroups.

In this paper we present approximation results that subsume those in [6], [3] and [4] in that we develop schemes to estimate parameters that include multiple delays, coefficients and initial data for nonlinear nonautonomous FDE. Our theoretical arguments avoid the Trotter–Kato linear semigroup formulation altogether. Rather, we combine dissipative type estimates with the use of Gronwall's inequality to develop a theory that not only allows for rather general nonlinearities but also accommodates with ease nonautonomous systems (both of which are features that the Trotter–Kato linear semigroup framework excludes). Of course, one could use an evolution operator analogue of the Trotter–Kato approximation theorem to obtain results for linear nonautonomous equations (see [7] for details), or a nonlinear Trotter–Kato type theorem for nonlinear autonomous FDE (see [13], [14]). Both of these separate approaches however are less direct than the one developed here when applied to parameter estimation problems.

While the approximation methods we develop can be used with great success to simply solve initial value problems for nonlinear nonautonomous FDE, the main focus of our treatment here is parameter identification or estimation. That our ideas can

* Received by the editors March 3, 1982, and in revised form September 3, 1982. Part of this research was completed while both authors were visitors at the Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA, operated under NASA contracts NAS1-15810 and NAS1-16394. This work was also supported in part by the Air Force Office of Scientific Research under contract AFOSR 76-3092D, in part by the National Science Foundation under grant NSF-MCS 7905774-02, and in part by the U.S. Army Research Office under contract ARO-DAAG29-79-C-0161.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

‡ Department of Mathematics, Southern Methodist University, Dallas, Texas 75275.

also be fruitfully employed in control problems is demonstrated in [9], [10] while application of the methods to estimation problems for certain partial differential equations can be found in [5].

The fundamental ideas (which were first presented for simple nonlinear autonomous, known delay, estimation problems in [1] and subsequently extended to treat nonautonomous, unknown delay, FDE problems in [10]) are really quite simple. However, the development of a theory for identification of the delays is a delicate matter since the "history space" for the delay system changes as one iteratively estimates the delays. This, unfortunately, results in a rather complicated presentation from the standpoint of technical notation regardless of the approach (e.g., see the treatment of the linear autonomous system case in [4]).

Our presentation is as follows: In § 2 we describe a parameter estimation problem for FDE's and give an equivalent Hilbert space formulation involving an abstract nonlinear evolution equation. Section 3 contains a discussion of approximate estimation problems based on spline subspaces; general convergence results are given. We conclude with a final section in which we present representative numerical findings obtained using the approximation scheme proposed in § 3.

Most of the notation (e.g., H^p for Sobolev spaces, L_p for Lebesgue spaces, etc.) is quite standard and is in accordance with popular usage. The symbol $|\cdot|$ will be used, in general, to denote the norm in various spaces in instances where no confusion will result. However in some situations it is absolutely essential to distinguish special weighted norms. These special norms will be defined as they are used in the discussions below. For convenience of the reader, we have summarized these definitions in a brief appendix for quick reference.

Finally we wish to mention the motivation behind our efforts to develop the methods presented below. In [4] and [10] one finds brief descriptions of nonlinear delay equation estimation problems arising in the study of enzymatically active column reactors. Although such problems actually prompted the theoretical investigations that we report here, a discussion of the application of our methods to these problems would be quite lengthy and thus will be the subject of a separate report.

2. Formulation of parameter estimation problems for nonlinear FDE. In the present section we describe the parameter estimation problem for a delay differential system and detail conditions under which solutions exist. Our approach then is to reformulate the FDE-governed identification problem as an abstract problem on an infinite-dimensional state space, concluding the section by establishing the equivalence between the FDE and the abstract state equation.

We consider the vector nonlinear delay equation

$$\begin{aligned} &\dot{x}(t) = f(\alpha, r_\nu, t, x(t), x_t, x(t-r_1), \cdots, x(t-r_\nu)) + g(t), \qquad a \leqq t \leqq b, \\ &(x(a), x_a) = (\eta, \phi), \end{aligned} \tag{2.1}$$

where x_t denotes the R^n-valued function $\Theta \to x_t(\Theta) = x(t+\Theta)$, $-r \leqq \Theta \leqq 0$, and g is a general $L_2^n(a, b)$ perturbation term. The equation depends on the parameters $\gamma = (\zeta, q)$, where $\zeta = (\eta, \phi)$ is the initial data in some set $S \subseteq W$, with

$$W \equiv \{(\psi(0), \psi) \in R^n \times L_2^n(-r, 0) \,|\, \psi \in H^1(-r, 0)\}.$$

The parameter $q = (\alpha, r_1, \cdots, r_\nu)$ is assumed to be in $Q = \mathscr{A} \times \mathscr{R}$ where α is a coefficient-type parameter in the set $\mathscr{A} \subseteq R^\mu$ and the discrete delays are chosen from

the set

$$\mathcal{R} = \{(r_1, \cdots, r_\nu) \in R^\nu \mid 0 = r_0 \leqq r_1 \leqq \cdots \leqq r_\nu \leqq r, r_\nu > 0\}$$

with $r > 0$ fixed and given throughout this paper.

To simplify notation we shall use $|\cdot|_{r_\nu}$ to denote the norm on $L_2^n(-r_\nu, 0)$ while we use $|\cdot|$ to denote the norm on $L_2^n(-r, 0)$ and $L_2^n(a, b)$. We make the following standing assumptions on f throughout the paper:

(H1) The mapping f satisfies a global Lipschitz condition on $R^n \times L_2^n(-r, 0) \times R^{n\nu}$ uniformly in $(\alpha, r_\nu) \in \mathcal{A} \times [0, r]$. That is, there exists $m_1 \in L_2(a, b)$, $m_1 > 0$, such that for all $(\xi, \psi, w_1, \cdots, w_\nu)$, $(\delta, \chi, y_1, \cdots, y_\nu) \in R^n \times L_2^n(-r, 0) \times R^{n\nu}$,

$$\begin{aligned} &|f(\alpha, r_\nu, t, \xi, \psi, w_1, \cdots, w_\nu) - f(\alpha, r_\nu, t, \delta, \chi, y_1, \cdots, y_\nu)| \\ &\quad \leqq m_1(t)\{|\xi - \delta| + |\psi - \chi|_{r_\nu} + \textstyle\sum_{i=1}^{\nu} |w_i - y_i|\} \end{aligned}$$

for all $(\alpha, r_\nu) \in \mathcal{A} \times [0, r]$ and a.a. $t \in [a, b]$.

(H2) For each $(\alpha, r_\nu) \in \mathcal{A} \times [0, r]$, $f(\alpha, r_\nu, \cdot, \cdots, \cdot)$: $[a, b] \times R^n \times L_2^n(-r, 0) \times R^{n\nu} \to R^n$ is differentiable, and $t \to f(\alpha, r_\nu, t, \psi(0), \psi, \psi(-r_1), \cdots, \psi(-r_\nu))$ is in $H^1(a, b)$ for every $\psi \in C^n(-r, 0) \equiv C([-r, 0]; R^n)$ and every $(\alpha, r_1, \cdots, r_\nu) \in \mathcal{A} \times \mathcal{R}$.

(H3) Given any $x \in C^n(a - r, b)$, the mapping

$$\sigma \to f_t(\alpha, r_\nu, \sigma, x(\sigma), x_\sigma, x(\sigma - r_1), \cdots, x(\sigma - r_\nu))$$

is in $L_2^n(a, b)$ for all $q \in Q$.

(H4) The function f is continuous on $\mathcal{A} \times [0, r] \times [a, b] \times R^n \times L_2^n(-r, 0) \times R^{n\nu}$.

Remark 2.1. It follows immediately from (H1) and (H2) that f satisfies an affine growth condition; that is, for a given $x \in L_2^n(a - r, b)$,

$$\begin{aligned} &|f(\alpha, r_\nu, t, x(t), x_t, x(t - r_1), \cdots, x(t - r_\nu))| \\ &\quad \leqq m_1(t)\left\{|x(t)| + |x_t|_{r_\nu} + \sum_{i=1}^{\nu} |x(t - r_i)|\right\} + m_2(\alpha, r_\nu, t), \end{aligned} \tag{2.2}$$

where $m_2(\alpha, r_\nu, t) = |f(\alpha, r_\nu, t, 0, \cdots, 0)|$ is in $L_2(a, b)$. Quite standard arguments may then be employed to demonstrate that, for each $q \in Q$,

$$t \to f(\alpha, r_\nu, t, x(t), x_t, x(t - r_1), \cdots, x(t - r_\nu))$$

is in $L_1^n(a, b)$ and that the mapping depends only on the equivalence class of x; therefore there will be no difficulty associated with point evaluations of x appearing in f since we shall write (2.1) as an equivalent integral equation.

Before we direct our attention to the estimation of the parameters appearing in (2.1) we shall first state results guaranteeing the existence, uniqueness and continuous dependence of solutions to the state equation for each choice of parameters $(\zeta, q) \in S \times Q$.

THEOREM 2.1. *Let $\gamma = (\zeta, q) = (\eta, \phi, \alpha, r_1, \cdots, r_\nu)$ be given in $S \times Q$. There exists a unique solution x to* (2.1) *on the interval $[a - r, b]$ which depends continuously on $\{\eta, \phi, g\}$ in the $R^n \times L_2^n(-r, 0) \times L_2^n(a, b)$ topology.*

In the proof, which may be found in [10, p. 6] and will not be detailed here, one employs general uniform contraction principles (see [12, p. 7]) and relies heavily on hypothesis (H1) and the growth condition (2.2).

We turn now to an examination of (2.1) when the parameters, including the delays $r_1, \cdots, r_\nu$, and initial data (η, ϕ) are to be estimated. We will restrict our attention to parameters in the admissible initial data-parameter set $\Gamma = S \times Q$ where we assume throughout that Γ has the following property:

(H5) Q is compact in $R^{\mu+\nu}$ and $S \subseteq W$ is compact in the $R^n \times L_2^n(-r, 0)$ topology.

The identification problem consists of finding $\bar{\gamma} \in \Gamma$ which provides the best least squares fit of the parameter-dependent solution (of the model equations (2.1)) to observations of the output at discrete sample times. The problem, which could also be reformulated as a maximum likelihood estimation problem, may be formally stated as follows:

Given g and observations $\{\hat{u}_i\}$, $\hat{u}_i \in R^s$, at times $\{t_i\}$, $i = 1, \cdots, M$, find $\bar{\gamma}$ in Γ which minimizes

$$J(\gamma) = \frac{1}{2} \sum_{i=1}^{M} |C(q)x(t_i; \gamma) - \hat{u}_i|^2 \tag{2.3}$$

over all $\gamma = (\zeta, q)$ in Γ. Here C is a given $s \times n$ matrix continuous in q, and $u(t; \gamma) = C(q)x(t; \gamma)$ represents the "observable part" of $x(t; \gamma)$, the solution to (2.1) corresponding to γ.

Remark 2.2. For $\bar{\gamma} = (\bar{\eta}, \bar{\phi}, \bar{\alpha}, \bar{r}_1, \cdots, \bar{r}_\nu)$ the optimal parameter, it may happen that $\bar{r}_\nu < r$ so that we actually only need $\bar{\phi}$ defined on $[-\bar{r}_\nu, 0]$ to integrate the state equations (2.1) (and, in fact, the $\bar{\phi}$ we obtain in practice will be defined on that interval only). We will view $\bar{\phi}$ as a function on all of $[-r, 0]$ by making an arbitrary, but definite, continuous extension from $-\bar{r}_\nu$ back to $-r$, so that $(\bar{\eta}, \bar{\phi})$ is an element of S as required.

Remark 2.3. The compactness assumption on S will not be difficult to satisfy in practice since a sufficient condition for compactness is that all elements (η, ϕ) in S are such that η belongs to a compact set in R^n and ϕ is bounded in $H^1(-r, 0)$. An example of one such admissible initial data set is the set of all polynomials on $[-r, 0]$ of order $\leqq k$ (k a nonnegative integer) with coefficients in a compact set.

2.1. An abstract reformulation of the estimation problem. We next reformulate (2.1) as an abstract evolution equation in an infinite-dimensional state variable. Our approach involving use of the state space $R^n \times L_2^n(-r, 0)$ is quite standard and well-established in the FDE literature (see, for example, [2] and the references therein); however, the dependence here of operators and state spaces on unknown parameters requires that we make such definitions in this and the following section with a certain amount of care.

We will let $Z = R^n \times L_2^n(-r, 0)$ with norm $|\cdot|$ induced by the inner product $\langle(\xi, \psi), (\delta, \chi)\rangle \equiv \xi^T\delta + \int_{-r}^{0} \psi^T\chi$. For $(q, t) \in Q \times [a, b]$, $(\xi, \psi) \in R^n \times C^n(-r, 0)$ define $F(q, t, \xi, \psi) = f(\alpha, r_\nu, t, \xi, \psi, \psi(-r_1), \cdots, \psi(-r_\nu))$ and $A(q, t): W \to Z$ by

$$A(q, t)(\psi(0), \psi) = (F(q, t, \psi(0), \psi), \mathcal{D}\psi), \tag{2.4}$$

where $\mathcal{D}\psi$ denotes the $L_2^n(-r, 0)$ function that is the derivative of ψ. In addition, let $G(t) = (g(t), 0) \in Z$, for $t \in [a, b]$.

The equivalence of the FDE (2.1) to an abstract evolution equation is detailed in Theorem 2.2; before proceeding, however, we need two results that also will be called upon frequently in § 3, so they are stated here as lemmas. Our first proposition is actually a restatement of the well-known result [8, p. 100] that $d/dt\, \frac{1}{2}|x(t)|^2 = \langle \dot{x}(t), x(t)\rangle$.

LEMMA 2.1. *If X is a Hilbert space and if $x: [a, b] \to X$ is given by $x(t) = x(a) + \int_a^t v(\sigma)\, d\sigma$, then*

$$|x(t)|^2 = |x(a)|^2 + 2\int_a^t \langle x(\sigma), v(\sigma)\rangle\, d\sigma.$$

The second result describes how to construct an equivalent topology for Z so that the nonlinear operator A satisfies a dissipative-type inequality. The Lemma, a nonlinear version of that found in [2, p. 186] and [6], greatly simplifies our calculations and is the foundation for our development without the use of semigroups.

LEMMA 2.2. *Let $q=(\alpha, r_1, \cdots, r_\nu)\in Q$ be given. For $y=(\xi,\psi)$, $z=(\delta,\chi)\in Z$ define a new inner product on Z by $\langle y, z\rangle_q \equiv \xi^T\delta+\int_{-r}^0 \psi(\Theta)^T\chi(\Theta)\tilde{\rho}(q)(\Theta)\,d\Theta$ where $\tilde{\rho}(q)$ is given on $[-r, 0]$ by*

$$\tilde{\rho}(q)(\Theta)=\begin{cases}1, & \Theta\in[-r, -r_\nu],\\ 2, & \Theta\in(-r_\nu, -r_{\nu-1}],\\ \vdots & \vdots\\ \nu+1, & \Theta\in(-r_1, 0].\end{cases} \tag{2.5}$$

Then

$$\langle A(q,t)y-A(q,t)z, y-z\rangle_q \leqq \omega(t)|y-z|_q^2$$

for all $q\in Q$, almost all $t\in[a,b]$ and all y, $z\in W$. The function $\omega>0$ is in $L_1(a,b)$ and is given by $\omega(t)=\frac{3}{2}m_1(t)+(\nu+1)/2+(\nu/2)m_1^2(t)$.

Proof. Let $y=(\psi(0),\psi)$, $z=(\chi(0),\chi)\in W$. Then

$$\langle A(q,t)y-A(q,t)z, y-z\rangle_q = [F(q,t,\psi(0),\psi)-F(q,t,\chi(0),\chi)]^T[\psi(0)-\chi(0)] + \int_{-r}^0 (\mathscr{D}\psi-\mathscr{D}\chi)^T(\psi-\chi)(\Theta)\tilde{\rho}(q)(\Theta)\,d\Theta,$$

where

$$\begin{aligned}&\int_{-r}^0 (\mathscr{D}\psi-\mathscr{D}\chi)^T(\psi-\chi)(\Theta)\tilde{\rho}(q)(\Theta)\,d\Theta\\ &=\int_{-r}^{-r_\nu}\mathscr{D}\left(\frac{|\psi(\Theta)-\chi(\Theta)|^2}{2}\right)d\Theta+\sum_{j=1}^{\nu}\int_{-r_j}^{-r_{j-1}}\mathscr{D}\left(\frac{|\psi(\Theta)-\chi(\Theta)|^2}{2}\right)(\nu+2-j)\,d\Theta\\ &=\frac{|\psi(-r_\nu)-\chi(-r_\nu)|^2}{2}-\frac{|\psi(-r)-\chi(-r)|^2}{2}+\frac{|\psi(0)-\chi(0)|^2}{2}(\nu+1)\\ &\quad-\sum_{j=1}^{\nu-1}\frac{|\psi(-r_j)-\chi(-r_j)|^2}{2}-\frac{|\psi(-r_\nu)-\chi(-r_\nu)|^2}{2}\cdot 2\\ &\leqq\frac{(\nu+1)}{2}|\psi(0)-\chi(0)|^2-\sum_{j=1}^{\nu}\frac{|\psi(-r_j)-\chi(-r_j)|^2}{2}.\end{aligned}$$

Therefore, for almost all $t\in[a,b]$,

$$\begin{aligned}&\langle A(q,t)y-A(q,t)z, y-z\rangle_q\\ &\leqq m_1(t)|\psi(0)-\chi(0)|^2+m_1(t)|\psi-\chi|_{r_\nu}|\psi(0)-\chi(0)|\\ &\quad+\sum_{j=1}^{\nu}|\psi(-r_j)-\chi(-r_j)|(m_1(t)|\psi(0)-\chi(0)|)\\ &\quad+\frac{\nu+1}{2}|\psi(0)-\chi(0)|^2-\frac{1}{2}\sum_{j=1}^{\nu}|\psi(-r_j)-\chi(-r_j)|^2\\ &\leqq\left(m_1(t)+\frac{m_1(t)}{2}+\frac{\nu m_1^2(t)}{2}+\frac{\nu+1}{2}\right)|\psi(0)-\chi(0)|^2+\frac{m_1(t)}{2}|\psi-\chi|_{r_\nu}^2,\end{aligned}$$

where we have used repeatedly the fact that $ab\leqq a^2/2+b^2/2$. It follows then that

$$\begin{aligned}\langle A(q,t)y-A(q,t)z, y-z\rangle_q &\leqq \omega(t)|\psi(0)-\chi(0)|^2+\omega(t)|\psi-\chi|_{r_\nu}^2\\ &\leqq \omega(t)|y-z|_q^2.\end{aligned}$$

□

It is clear that for all $q \in Q$, the norm $|\cdot|_q$ on Z induced by the $\tilde{\rho}(q)$ weighted inner product is equivalent to the usual Z norm since $1 \leqq \tilde{\rho}(q) \leqq \nu + 1$.

THEOREM 2.2. *For fixed $\gamma \in \Gamma$ let $y(t; \gamma, g) = (x(t; \gamma, g), x_t(\gamma, g))$, where x is the solution to (2.1) corresponding to $\gamma = (\eta, \phi, \alpha, r_1, \cdots, r_\nu)$ and $g \in L_2^n(a, b)$. Then $y(\gamma, g)$ is the unique solution on $[a, b]$ of*

$$z(t) = \zeta + \int_a^t \{A(q, \sigma) z(\sigma) + G(\sigma)\}\, d\sigma. \tag{2.6}$$

Furthermore, $y(t; \gamma, g) \in Z$ is continuous in $t \in [a, b]$, and uniformly continuous in $\{\zeta, g\} \in W \times L_2^n(a, b)$ (in the $Z \times L_2^n(a, b)$ topology) uniformly in $t \in [a, b]$.

Proof. We shall sketch the proof of the theorem, which demonstrates that (i) the integrand in (2.6) is well-defined and integrable, (ii) the equality in (2.6) holds for $z(t) = y(t; \gamma, g)$, and (iii) the solution y is unique and continuously dependent.

To prove (i) we first must show that $y(t; \gamma, g) \in \operatorname{dom}(A(q, t)) = W$ for each $t \in [a, b]$, or, since $x_t(0) = x(t)$, that $x_t \in H^1(-r, 0)$ for all t. Using the affine growth condition (2.2) on f and the continuity of x it is not difficult to show that $t \to f(\alpha, r_\nu, t, x(t), x_t, x(t - r_1), \cdots, x(t - r_\nu))$ is square-integrable on $[a, b]$ so that, using the fact that $\dot{x} = \dot{\phi}$ on $[a - r, a]$, we obtain $\dot{x} \in L_2^n(a - r, b)$. Standard estimates in [11, p. 254] may be invoked to demonstrate that

$$\frac{1}{\varepsilon}(x_{t+\varepsilon} - x_t) \to (\dot{x})_t \quad \text{in } L_2^n(-r, 0) \tag{2.7}$$

so that $\mathcal{D}(x_t) = (\dot{x})_t \in L_2^n(-r, 0)$, for each $t \in [a, b]$. Arguments similar to these may be used to show $A(t)y(t) + G(t)$ is integrable on $[a, b]$, concluding the proof of (i).

The argument that $y(t) = (x(t; \gamma, g), x_t(\gamma, g))$ satisfies equation (2.6) is trivial if this equation is examined componentwise: The R^n part of (2.6) is simply a restatement of (2.1) while the desired equality for the $L_2^n(-r, 0)$ component follows immediately from (2.7).

Finally, uniqueness and continuous dependence of solutions on $\{\zeta, g\}$ follow from arguments that will be repeated often throughout this paper and will be presented in detail here for the case of continuous dependence; uniqueness also follows from these arguments. Let z_1, z_2 denote solutions to (2.6) corresponding to $\{\zeta_1, g_1\}$, $\{\zeta_2, g_2\}$ respectively with $q \in Q$ fixed. Then, for $t \in [a, b]$,

$$z_1(t) - z_2(t) = \zeta_1 - \zeta_2 + \int_a^t \{A(\sigma) z_1(\sigma) - A(\sigma) z_2(\sigma) + (g_1(\sigma), 0) - (g_2(\sigma), 0)\}\, d\sigma$$

so that from Lemma 2.1,

$$\begin{aligned}
|z_1(t) - z_2(t)|_q^2 &\leqq |\zeta_1 - \zeta_2|_q^2 + 2\int_a^t \langle A(\sigma) z_1(\sigma) - A(\sigma) z_2(\sigma), z_1(\sigma) - z_2(\sigma)\rangle_q\, d\sigma \\
&\quad + 2\int_a^t |(g_1(\sigma) - g_2(\sigma), 0)|_q |z_1(\sigma) - z_2(\sigma)|_q\, d\sigma \\
&\leqq |\zeta_1 - \zeta_2|_q^2 + 2\int_a^t \omega(\sigma) |z_1(\sigma) - z_2(\sigma)|_q^2\, d\sigma \\
&\quad + \int_a^t |(g_1(\sigma) - g_2(\sigma), 0)|_q^2\, d\sigma + \int_a^t |z_1(\sigma) - z_2(\sigma)|_q^2\, d\sigma \\
&\leqq |\zeta_1 - \zeta_2|^2 (\nu + 1) + |g_1 - g_2|^2 + \int_a^b (2\omega(\sigma) + 1) |z_1(\sigma) - z_2(\sigma)|_q^2\, d\sigma.
\end{aligned}$$

Gronwall's inequality may be used to obtain

$$|z_1(t)-z_2(t)|_q^2 \leqq (|\zeta_1-\zeta_2|^2(\nu+1)+|g_1-g_2|^2)\cdot \exp\int_a^b (2\omega(\sigma)+1)\,d\sigma$$

from which continuous dependence (uniform in $t\in[a, b]$) follows at once. □

We have demonstrated the equivalence between an FDE in $x(t)\in R^n$ and an abstract evolution equation (AEE) in the infinite-dimensional state variable $z(t)$. We remark that the infinite dimensionality of (2.6) is inherited from (2.1) in that in the latter the history of x on $[t-r_\nu, t)$ is required before x may be determined at t. Thus the computational difficulties encountered with (2.6) are not simply an undesirable feature of this reformation of (2.1) but rather are a manifestation of the inherent infinite dimensionality of the underlying FDE.

In view of the established equivalence, the ID problem in (2.3) may be reformulated as an abstract ID problem, where we now wish to find $\bar\gamma\in\Gamma$ which minimizes

$$J(\gamma)=\frac{1}{2}\sum_{i=1}^{M}|C(q)\pi_0 z(t_i;\gamma)-\hat u_i|^2 \tag{2.8}$$

over all $\gamma\in\Gamma$ where $\pi_0: Z\to R^n$ is defined by $\pi_0(\xi,\psi)=\xi$.

In the next section we investigate the problem of approximating the infinite-dimensional identification problem (2.8) by a sequence of finite-dimensional state space identification problems (where the state variable satisfies an ordinary differential equation (ODE) on a finite-dimensional state space X^N). Fundamental to this undertaking is the establishment of the convergence of solutions of the approximating systems on X^N to solutions of the original equation on Z. Although our formulation is a classical one of the Ritz type (involving orthogonal projections of an infinite-dimensional system onto a sequence of finite-dimensional subspaces) our problem is complicated by the fact that the state space changes for each choice of parameters $q=(\alpha, r_1, \cdots, r_\nu)$. This concept is explained in detail in [4, p. 800] and involves the idea that the natural state space for $z(t)$ associated with the parameter choice $q=(\alpha, r_1, \cdots, r_\nu)$ is $X(q)=R^n\times L_2^n(-r_\nu, 0)$, where in general $X(q)\not\subseteq Z$. Since we would expect the (finite-dimensional) approximating spaces $X^N(q)$ associated with q to be subspaces of $X(q)$, we obtain a sequence of spaces $\{X^N(q)\}$ where $X^N(q)$ is different for each choice of q and is usually not contained in Z.

3. Approximate parameter estimation problem. Our focus in this section is the definition of finite-dimensional ODE-governed estimation problems which approximate the ID problem governed by the AEE (2.6), and their relationship to the original FDE-based ID problem. While we shall present the details for a scheme based on linear splines, arbitrary order spline approximations may be employed in a similar way with only slight modifications in the arguments detailed below (see the theory developed in [6] on which all of our development here is based).

For parameters $\gamma=(\zeta, q)\in\Gamma$ consider

$$z^N(t;\gamma)=P^N(q)\zeta+\int_a^t\{A^N(q,\sigma)z^N(\sigma;\gamma)+P^N(q)G(\sigma)\}\,d\sigma,\qquad t\in[a, b], \tag{3.1}$$

where A^N and P^N are defined via the following q-dependent operators and spaces. For a given $q=(\alpha, r_1, \cdots, r_\nu)$ we define the Hilbert space $X(q)$ as the set $R^n\times L_2^n(-r_\nu, 0)$ with inner product

$$\langle(\xi,\psi),(\delta,\chi)\rangle_{\rho,q}=\xi^T\delta+\int_{-r_\nu}^{0}\psi(\Theta)^T\chi(\Theta)\rho(q)(\Theta)\,d\Theta,$$

where $\rho(q)(\Theta)=\tilde{\rho}(q)(\Theta)-1$, with $\tilde{\rho}$ defined as in (2.5). We shall also use an equivalent topology on $X(q)$ given by the (unweighted) inner product $\langle(\xi,\psi),(\delta,\chi)\rangle_{X,q}=\xi^T\delta+\int_{-r_\nu}^0\psi(\Theta)^T\chi(\Theta)\,d\Theta$. The operator $\mathscr{E}^+(q):X(q)\to Z$ is the "continuous extension" operator defined by $\mathscr{E}^+(q)(\xi,\psi)=(\xi,\tilde{\psi})$ where $\tilde{\psi}=\psi$ on $[-r_\nu,0]$, $\tilde{\psi}(\Theta)=\psi(-r_\nu)$, $\Theta\in[-r,-r_\nu)$; $i(q):Z\to X(q)$ is defined by $i(q)(\xi,\psi)=(\xi,\hat{\psi})$ where $\hat{\psi}$ is the restriction of ψ to $[-r_\nu,0]$. The subspaces $X^N(q)$ of $X(q)$ are defined by $X^N(q)=\{(\psi(0),\psi)\,|\,\psi$ is a piecewise linear spline with knots at $t_j^N(q)=[-(j-(k-1)N)(r_k-r_{k-1})/N]-r_{k-1}$, $j=(k-1)N+1,\cdots,kN$, $k=1,\cdots,\nu$; $t_0^N=0\}$ and we denote by $\pi^N(q):X(q)\to X^N(q)$ the canonical orthogonal projection of $X(q)$ onto $X^N(q)$ along $X^N(q)^\perp$. Finally, $P^N(q):Z\to X^N(q)$ is defined by $P^N(q)=\pi^N(q)i(q)$ and $A^N(q,\sigma):X(q)\to X^N(q)$ is given by $A^N(q,\sigma)=\pi^N(q)\tilde{A}(q,\sigma)\pi^N(q)$ where here $\tilde{A}(q,\sigma)$ is defined as an operator in $X(q)$ given by $\tilde{A}(q,\sigma)(\psi_q(0),\psi_q)=(F(q,\sigma,\psi_q(0),\bar{\psi}_q),\mathscr{D}\psi_q)$ for $(\psi_q(0),\psi_q)\in W(q)\equiv\{(\psi(0),\psi)\,|\,\psi\in H^1(-r_\nu,0)\}$. Here and below, for any ψ_q defined on $[-r_\nu,0]$, $\bar{\psi}_q$ denotes the extension of ψ_q to all of $[-r,0]$ defined by $\bar{\psi}_q\equiv 0$ on $[-r,-r_\nu)$. (The arguments and proofs below are actually independent of how we extend ψ_q.)

Remark 3.1. A^N is well-defined since $X^N(q)$, the range of $\pi^N(q)$, is contained in the domain of $\tilde{A}(q,t)$. Note also that $A^N(q,t)$ actually may be considered as an operator from Z into $X^N(q)$ if it is defined by $A^N(q,\sigma)=P^N(q)\tilde{A}(q,\sigma)P^N(q)$. For present uses though, $P^N(q)\zeta$ and $z^N(\sigma;\gamma)$ are in $X^N(q)$ so that viewing $A^N(q,\sigma)$ as an operator from $X^N(q)$ to itself yields (3.1) as an equation on $X^N(q)$, a finite-dimensional space since each of its elements is completely determined by its value at each of $\nu N+1$ knots. Equation (3.1) is then equivalent to the ODE

$$\begin{aligned}&\dot{z}^N(t;\gamma)=A^N(q,t)z^N(t;\gamma)+P^N(q)G(t),\qquad t\in(a,b],\\&z^N(a;\gamma)=P^N(q)\zeta,\end{aligned}\tag{3.2}$$

which, as we shall show in the arguments that follow, approximates (2.1) in some sense.

When the parameter γ is unknown we may state an "approximate identification problem" $\mathscr{P}_N$ associated with (3.1) and (3.2):

Find $\bar{\gamma}^N=(\bar{\zeta}^N,\bar{q}^N)\in\Gamma$ so as to minimize

$$J^N(\gamma)=\frac{1}{2}\sum_{i=1}^M|C(q)\pi_0 z^N(t;\gamma)-\hat{u}_i|^2$$

over $\gamma\in\Gamma$, where g and observations $\hat{u}_i$ at times t_i, $i=1,\cdots,M$ are given and $z^N(t;\gamma)$ satisfies (3.1).

We now establish the existence of a unique solution to (3.1) for each choice of γ, and, further, the existence of a solution $\bar{\gamma}^N$ to the Nth ID problem, $\mathscr{P}_N$. First we must state an analogue of Lemma 2.2 which demonstrates a type of dissipativeness for A^N.

LEMMA 3.1. *Let $q=(\alpha,r_1,\cdots,r_\nu)\in Q$ be given. Then*

$$\langle A^N(q,t)y^N-A^N(q,t)z^N,\,y^N-z^N\rangle_{\rho,q}\leqq\omega(t)|y^N-z^N|_{\rho,q}^2$$

for all y^N, $z^N\in X^N(q)$ where ω, defined in Lemma 2.2, is independent of q and N.

Proof. Note first that for y, $z\in W(q)$, we may argue that

$$\langle\tilde{A}(q,t)y-\tilde{A}(q,t)z,\,y-z\rangle_{\rho,q}\leqq\omega(t)|y-z|_{\rho,q}^2$$

using estimates similar to those used to prove Lemma 2.2, where $\omega(t)$ is independent

of q, N. Then for $y^N, z^N \in X^N(q) \subseteq W(q)$,

$$\begin{aligned}
&\langle A^N(q,t)y^N - A^N(q,t)z^N, y^N - z^N\rangle_{\rho,q} \\
&\quad = \langle \pi^N(q)\tilde{A}(q,t)\pi^N(q)y^N - \pi^N(q)\tilde{A}(q,t)\pi^N(q)z^N, y^N - z^N\rangle_{\rho,q} \\
&\quad = \langle \tilde{A}(q,t)\pi^N(q)y^N - \tilde{A}(q,t)\pi^N(q)z^N, \pi^N(q)y^N - \pi^N(q)z^N\rangle_{\rho,q} \\
&\quad \leqq \omega(t)|\pi^N(q)y^N - \pi^N(q)z^N|^2_{\rho,q} \\
&\quad \leqq \omega(t)|y^N - z^N|^2_{\rho,q},
\end{aligned}$$

where we have used the properties of the (self-adjoint) orthogonal projection π^N and the dissipativeness of $\tilde{A}(q,t)$ on $W(q)$. □

Our next result demonstrates the existence of solutions to (3.1) as well as to the identification problem $\mathcal{P}_N$. In addition, the proof sheds light on the numerical procedure used to solve (3.1).

THEOREM 3.1. *Let $g \in L_2^n(a,b)$ and $\gamma = (\zeta, q) \in \Gamma$ be given. Then there exists a unique solution $z^N(t;\gamma,g) \in X^N(q)$ to (3.1) on $[a,b]$ with the property that the map $\{i(q)\zeta, g\} \to z^N(t;(\zeta,q),g)$ is uniformly continuous on $X(q)\times L_2^n(a,b)$, uniformly in N and t. Finally, there exists a solution $\bar{\gamma}^N$ to the Nth identification problem $\mathcal{P}_N$ for each $N = 1, 2, \cdots$.*

Remark 3.2. The continuity with respect to initial data given in this theorem is actually "uniform in $q \in Q$" in the following sense: Given $\varepsilon > 0$, there exists $\delta > 0$ independent of q and N such that for $\zeta_1, \zeta_2 \in S$ and $q \in Q$ with $|\zeta_1 - \zeta_2|_{X,q} < \delta$, we have $|z^N(t;(\zeta_1,q),g) - z^N(t;(\zeta_2,q),g)|_{X,q} < \varepsilon$. This type of "uniformity in q" follows from the arguments given below for Theorem 3.1 and will be used in establishing the convergence results of Theorem 3.3.

Proof. We first argue existence, uniqueness and continuous dependence of solutions to (3.1). We shall do this using arguments similar to those in [1] and [6] (where $z^N(t)$ is written in terms of basis vectors for $X^N(q)$). Let $q \in Q$ be fixed, $q = (\alpha, r_1, \cdots, r_\nu)$, and let e_j^N denote the scalar first-order spline function on $[-r_\nu, 0]$ characterized by

$$e_j^N(t_i^N) = \delta_{ij}, \qquad i,j = 0, 1, \cdots, \nu N$$

where δ_{ij} is the Kronecker symbol and $t_i^N = t_i^N(q)$ are the knots defined for functions in $X^N(q)$, $i = 0, \cdots, \nu N$. Define

$$\hat{e}_j^N = (e_j^N(0), e_j^N), \quad j = 0, \cdots, \nu N, \qquad \beta^N = (e_0^N, \cdots, e_{\nu N}^N) \otimes I,$$

where I is the $n \times n$ identity matrix and $\otimes$ denotes the Kronecker product so that β^N is an $(n \times n(\nu N+1))$-matrix-valued function on $[-r_\nu, 0]$. We represent by $\hat{\beta}^N$ the matrix-valued pair, $\hat{\beta}^N = (\beta^N(0), \beta^N)$, and whenever w is an $n(\nu N+1)$ vector, we adopt the notation $\hat{\beta}^N w \equiv (\beta^N(0)w, \beta^N w)$.

From [6], $X^N(q) = \text{span}\{\hat{\beta}_j^N\}$, $j = 1, \cdots, n(\nu N+1)$ where the basis vectors are given by $\hat{\beta}_j^N = (\beta_j^N(0), \beta_j^N)$, β_j^N the jth column of β^N. (Note that $\hat{\beta}_j^N$ is *not* the jth column of $\hat{\beta}^N$.) It follows then that since $z^N(t) \in X^N(q)$, there exists $w^N(t) \in R^{n(\nu N+1)}$ such that

$$z^N(t) = \hat{\beta}^N w^N(t) = \sum_{j=0}^{\nu N} w_j^N(t)\hat{e}_j^N = \left(w_0^N(t), \sum_{j=0}^{\nu N} w_j^N(t)e_j^N\right)$$

for $w_j^N(t) \in R^n$, $j = 0, \cdots, \nu N$. Furthermore, since $P^N G(t)$ and $P^N\zeta$ are vectors in $X^N(q)$, there exist $G^N(t), \zeta^N \in R^{n(\nu N+1)}$ such that

$$P^N G(t) = \hat{\beta}^N G^N(t) \quad \text{and} \quad P^N\zeta = \hat{\beta}^N\zeta^N$$

so that (3.2) may now be written in terms of $\hat{\beta}^N$ as

$$\begin{aligned} &\hat{\beta}^N \dot{w}^N(t) = A^N(q,t)\hat{\beta}^N w^N(t) + \hat{\beta}^N G^N(t), \qquad t \in (a,b], \\ &\hat{\beta}^N w^N(a) = \hat{\beta}^N \zeta^N. \end{aligned} \tag{3.3}$$

Let $\mathscr{A}^N(q,t)$ denote the representation of $A^N(q,t)$ (restricted to $X^N(q)$) with respect to the basis of $X^N(q)$. Here $\mathscr{A}^N(q,t)$ is nonlinear as opposed to the matrix (linear) version of the operator arising in [6]. As in [1] and [6], usual Galerkin calculations establish that the coefficients $w^N(t)$ in (3.3) satisfy

$$\begin{aligned} &\dot{w}^N(t) = \mathscr{A}^N(q,t)w^N(t) + G^N(t), \qquad t \in (a,b], \\ &w^N(a) = \zeta^N. \end{aligned} \tag{3.4}$$

We next establish a representation of $\mathscr{A}^N(q,t)$ which will enable us to consider the existence and uniqueness of solutions to (3.4) as well as the realization of numerical solution techniques for the system. Note first that

$$\begin{aligned} A^N(q,t)z^N(t) &= \pi^N(q)\tilde{A}(q,t)\pi^N(q)\Big(w_0^N(t), \sum_{j=0}^{\nu N} w_j^N(t)e_j^N\Big) \\ &= \pi^N(q)\Big(f(\alpha, r_\nu, t, w_0^N(t), \overline{\sum_{j=0}^{\nu N} w_j^N(t)e_j^N}, \\ &\qquad \sum_{j=0}^{\nu N} w_j^N(t)e_j^N(-r_1), \cdots, \sum_{j=0}^{\nu N} w_j^N(t)e_j^N(-r_\nu)), \sum_{j=0}^{\nu N} w_j^N(t)\mathscr{D}e_j^N\Big) \\ &= \pi^N(q)\Big(\tilde{f}(\alpha, r_\nu, t, w^N(t)), \sum_{j=0}^{\nu N} w_j^N(t)\mathscr{D}e_j^N\Big) \end{aligned}$$

where $\tilde{f}: \mathscr{A} \times [0,r] \times R^{1+n(\nu N+1)} \to R^n$ is defined by

$$\tilde{f}(\alpha, r_\nu, t, (v_0, \cdots, v_{\nu N})^T) = f\Big(\alpha, r_\nu, t, v_0, \overline{\sum_{j=0}^{\nu N} v_j e_j^N}, \sum_{j=0}^{\nu N} v_j e_j^N(-r_1), \cdots, \sum_{j=0}^{\nu N} v_j e_j^N(-r_\nu)\Big)$$

for $v_j \in R^n$, $j = 0, \cdots, \nu N$, and can be shown to be globally Lipschitz in $(v_0, \cdots, v_{\nu N})^T \in R^{n(\nu N+1)}$ since f satisfies such a condition. Thus, $\mathscr{A}^N(q,t)w^N(t) = \alpha^N(t)$, where $\alpha^N(t) \in R^{n(\nu N+1)}$ is such that

$$\hat{\beta}^N \alpha^N(t) = A^N(q,t)\hat{\beta}^N w^N(t) = \pi^N(q)\Big(\tilde{f}(\alpha, r_\nu, t, w^N(t)), \sum_{j=0}^{\nu N} w_j^N(t)\mathscr{D}e_j^N\Big).$$

It follows from [6, p. 508] that whenever $\pi^N(q)(\xi,\psi) = \hat{\beta}^N \delta^N$, $(\xi,\psi) \in X(q)$, $\delta^N \in R^{n(\nu N+1)}$, we have

$$\delta^N = (Q^N)^{-1} h^N(\xi,\psi),$$

where the nonsingular matrix Q^N is given by $Q^N = (\beta^N(0))^T\beta^N(0) + \int_{-r_\nu}^0 \beta^N(\Theta)^T\beta^N(\Theta)\rho(q)(\Theta)\,d\Theta$ and $h^N(\xi,\psi) = (\beta^N(0))^T\xi + \int_{-r_\nu}^0 \beta^N(\Theta)^T\psi(\Theta)\rho(q)(\Theta)\,d\Theta$. We may apply these results to obtain

$$\begin{aligned} \alpha^N(t) &= (Q^N)^{-1}h^N\Big(\tilde{f}(\alpha, r_\nu, t, w^N(t)), \sum_{j=0}^{\nu N} w_j^N(t)\mathscr{D}e_j^N\Big) \\ &= (Q^N)^{-1}\begin{pmatrix} \tilde{f}(\alpha, r_\nu, t, w^N(t)) \\ 0 \\ \vdots \\ 0 \end{pmatrix} + (Q^N)^{-1}H_{12}^N w^N(t), \end{aligned}$$

where H_{12}^N is given in [6] and [10] by

$$H_{12}^N = \begin{pmatrix} \langle \dot{e}_0^N, e_0^N \rangle & \cdots & \langle \dot{e}_{\nu N}^N, e_0^N \rangle \\ \vdots & & \vdots \\ \langle \dot{e}_0^N, e_{\nu N}^N \rangle & \cdots & \langle \dot{e}_{\nu N}^N, e_{\nu N}^N \rangle \end{pmatrix} \otimes I.$$

In this matrix $\langle \cdot, \cdot \rangle$ denotes the $\rho(q)$-weighted $L_2^1(-r_\nu, 0)$ inner product.

Similarly, $G^N(t)$ in (3.4) is given by $G^N(t) = (Q^N)^{-1} h^N(G(t))$, $h^N((g(t), 0)) = (g(t), 0, \cdots, 0)^T \in R^{n(\nu N+1)}$, so that (3.4) may be rewritten as

$$\begin{aligned} \dot{w}^N(t) &= (Q^N)^{-1}(\tilde{f}(\alpha, r_\nu, t, w^N(t)) + g(t), 0, \cdots, 0)^T \\ &\quad + (Q^N)^{-1} H_{12}^N w^N(t), \qquad t \in (a, b], \\ w^N(a) &= \zeta^N, \end{aligned} \tag{3.5}$$

an ODE in $w^N(t) = w^N(t; \gamma, g) \in R^{n(\nu N+1)}$. Since $\tilde{f}$ satisfies a global Lipschitz condition in $w^N(t)$, the form of (3.5) allows one to employ standard ODE theory to obtain the existence of a unique solution $w^N(t)$ on $[a, b]$. We can therefore conclude that

$$z^N(t) = \hat{\beta}^N w^N(t)$$

is the unique solution to (3.1) (and (3.2)) on $[a, b]$ for $\zeta \in Z$ given.

The proof of the continuous dependence on ζ and g as stated in the theorem is identical to the corresponding proof in Theorem 2.2 where dissipativeness for $A^N(q, t)$ is now used to show that whenever $|\zeta_1 - \zeta_2|_{X,q} < \delta$ and $|g_1 - g_2| < \delta$, $\delta = \delta(\varepsilon, \omega, a, b)$ independent of t, q and N, we have for the corresponding solutions, $z_1^N(t; (\zeta_1, q), g_1)$ and $z_2^N(t; (\zeta_2, q), g_2)$,

$$|z_1^N(t) - z_2^N(t)|_{X,q} \leqq |z_1^N(t) - z_2^N(t)|_{\rho,q} < \varepsilon$$

for all $t \in [a, b]$.

Finally, to establish existence of a solution $\bar{\gamma}^N$ to $\mathcal{P}_N$, one argues continuity (for fixed N) of the map $\gamma = (\zeta, q) \to \pi_0 z^N(t; \gamma) = w_0^N(t, \gamma)$ and thus infers continuity of $\gamma \to J^N(\gamma)$ on the compact set Γ. But it is not difficult to see that the right side of (3.5) depends continuously on q as do the basis elements $\hat{e}_j^N(q)$. Continuous dependence results (with respect to parameters and initial data) from the theory of ordinary differential equations can then be invoked to obtain the desired conclusions. □

In view of the last result, we are assured of a solution $\bar{\gamma}^N$ to the Nth estimation problem $\mathcal{P}_N$ (which is a standard least squares problem governed by an ODE). Since an application of conventional optimization techniques requires a solution to (3.1) for each choice of γ, straightforward computational schemes may be devised to solve (3.5), the associated ODE in the "Fourier" coefficients $w^N(t)$. Although it may be relatively easy to solve the finite-dimensional problem $\mathcal{P}_N$, the solution $\bar{\gamma}^N$ we find is meaningful only if $\bar{\gamma}^N$ approximates the solution $\bar{\gamma}$ to the original ID problem. Fundamental to the establishment of this fact (i.e., the convergence of $\bar{\gamma}^N$ to $\bar{\gamma}$ in some sense) is the demonstration that the sequence of state variables $\{z^N(t; \gamma^N, g)\}$ converges to $z(t; \tilde{\gamma}, g)$ given *any* sequence $\{\gamma^N\}$ with $\gamma^N \to \tilde{\gamma} = (\tilde{\zeta}, \tilde{q})$ in Γ. We shall first consider this problem for limits $\tilde{\gamma}$ and perturbing functions g such that $\{\tilde{\zeta}, g\}$ lies in a smooth but dense subset of $W \times L_2^n(a, b)$ (which simplifies our calculations). We then extend the convergence results for all limits $\tilde{\zeta}$ and perturbations g such that $\{\tilde{\zeta}, g\} \in W \times L_2^n(a, b)$.

3.1. Convergence of state variables. We shall assume that a sequence of parameters $\{\gamma^N\}$ in Γ has been given, $\gamma^N = (\zeta^N, q^N) = (\eta^N, \phi^N, \alpha^N, r_1^N, \cdots, r_\nu^N)$, and that $\gamma^N \to \tilde{\gamma} = (\tilde{\zeta}, \tilde{q}) = (\tilde{\eta}, \tilde{\phi}, \tilde{\alpha}, \tilde{r}_1, \cdots, \tilde{r}_\nu)$, in the sense that (i) $q^N \to \tilde{q}$ in $R^{\mu+\nu}$ and (ii)

$|i(q^N)(\tilde{\zeta}-\zeta^N)|_{\rho,q^N}\to 0$ as $N\to\infty$. We make the following standing assumptions on $\tilde{\gamma}$ and S:

(H6) There exists some $\delta_r>0$ such that $|\tilde{r}_k-\tilde{r}_{k-1}|\geqq\delta_r$, $k=1,2,\cdots,\nu$.

(H7) If $\zeta\in S$ then $\mathscr{C}^+(q^N)i(q^N)\zeta\in S$ for all N.

Remark 3.3. We note that the set S involving all polynomials of order $\leqq k$ on $[-r,0]$ mentioned in Remark 2.3 does not, strictly speaking, satisfy (H7). However, the reader can easily see from the arguments below that a modification in defining the extension operator $\mathscr{C}^+$ (rather than extending from $[-r_\nu,0]$ to $[-r,0]$ by constant values, extend any polynomial on $[-r_\nu,0]$ to $[-r,0]$ by simply extending the domain of definition of the polynomial) would allow the set S of Remark 2.3 to satisfy (H7) and not require any change in the convergence arguments to follow.

In what follows we will simplify notation by abbreviating $X^N\equiv X^N(q^N)$, $A^N(t)\equiv A^N(q^N,t)$, $A(t)\equiv A(\tilde{q},t)$, $P^N\equiv P^N(q^N)$, $i^N\equiv i(q^N)$, $\pi^N\equiv\pi^N(q^N)$, $\mathscr{C}_N^+\equiv\mathscr{C}^+(q^N)$, and $|\cdot|_N\equiv|\cdot|_{\rho,q^N}$. We shall also use $|\cdot|_N$ to denote the $L_2^n(-r_\nu^N,0)$ norm weighted with $\rho(q^N)$. We remind the reader that $|\cdot|$ denotes either the Z or $L_2^n(-r,0)$ norm while $|\cdot|_{r_\nu^N}$ denotes the unweighted $L_2^n(-r_\nu^N,0)$ norm. When no confusion results we shall also write $z(t)$ instead of $z(t;\tilde{\gamma},g)$ and $z^N(t)$ for $z^N(t;\gamma^N,g)$ the solutions to (2.6) and (3.1) associated with $\tilde{\gamma}$ and γ^N, respectively.

For q given in Q, define $\mathscr{J}(q)=\{\{\zeta,g\}\in W\times L_2^n(a,b)|\zeta=(\psi(0),\psi),\ \psi\in H^2(-r,0),\ g\in H^1(a,b),\ \dot{\psi}(0)=F(q,a,\psi(0),\psi)+g(a)\}$ and define $\mathscr{S}=\{(\psi(0),\psi)\in Z\,|\,\psi\in H^2(-r,0)\}$.

LEMMA 3.2. *For any $q\in Q$, $\mathscr{J}(q)$ is dense in $W\times L_2^n(a,b)$ (in the $Z\times L_2^n(a,b)$ topology). Furthermore, if $\{\zeta,g\}\in\mathscr{J}(q)$, then the solution $z(t;(\zeta,q),g)=(x(t;(\zeta,q),g),x_t((\zeta,q),g))$ to (2.6) corresponding to ζ,q,g satisfies $z(t)\in\mathscr{S}$ for all $t\in[a,b]$.*

Proof. Let $q\in Q$ and $\zeta=(\psi(0),\psi)$ be fixed in $\mathscr{S}$ and define $\mathscr{J}(q,\zeta)=\{g\in L_2^n(a,b)\,|\,g\in H^1(a,b),\ g(a)=\dot{\psi}(0)-F(q,a,\psi(0),\psi)\}$. Then for $g\in L_2^n(a,b)$ given and $\varepsilon>0$, standard arguments may be used to construct a $\hat{g}$ that is piecewise $-C^{(1)}$ satisfying $\hat{g}(a)=\dot{\psi}(0)-F(q,a,\psi(0),\psi)$ with $|g-\hat{g}|<\varepsilon$. That is, $\mathscr{J}(q,\zeta)$ is dense in $L_2^n(a,b)$. Furthermore, for $\zeta\in\mathscr{S}$, the pair $\{\zeta,g\}$ belongs to $\mathscr{J}(q)$ whenever $g\in\mathscr{J}(q,\zeta)$, so that

$$\bigcup_{\zeta\in\mathscr{S}}[\{\zeta\}\times\mathscr{J}(q,\zeta)]\subseteq\mathscr{J}(q)\subseteq W\times L_2^n(a,b),$$

where the first set is dense in the last since $\mathscr{S}$ is dense in W. It follows that $\mathscr{J}(q)$ is dense in $W\times L_2^n(a,b)$.

Required for the proof of the second part of the theorem is a verification that $\ddot{x}\in L_2(a-r,b)$ (since $\mathscr{D}^2(x_t)=(\ddot{x})_t$ for $t\in[a,b]$). If $\{\zeta,g\}\in\mathscr{J}(q)$, $\zeta=(\eta,\phi)$, it follows that $x\in C^{(1)}[a-r,b]$ since: (1) $\dot{x}(t)=\dot{\phi}(t-a)$ for $t\in[a-r,a)$; (2) for $t\in(a,b)$, $\dot{x}(t)=f(\alpha,r_\nu,t,x(t),x_t,x(t-r_1),\cdots,x(t-r_\nu))+g(t)$, which is continuous by assumption (H2) and the definition of $\mathscr{J}(q)$; and (3) $\dot{x}(a^-)=\dot{\phi}(0)=F(q,a,\phi(0),\phi)+g(a)=\dot{x}(a^+)$. Further, the differentiability of f and g yields

$$\begin{aligned}\ddot{x}(t)=&f_\sigma(\alpha,r_\nu,t,x(t),x_t,\cdots)+f_\xi(\alpha,r_\nu,t,x(t),x_t,\cdots)\dot{x}(t)\\&+f_\psi[\alpha,r_\nu,t,x(t),x_t,\cdots;\dot{x}_t]\\&+\sum_{i=1}^{\nu}f_{y_i}(\alpha,r_\nu,t,x(t),x_t,\cdots)\dot{x}(t-r_i)+\dot{g}(t)\end{aligned}$$

for $t\in(a,b)$, where f_δ denotes the Fréchet derivative of $f(\alpha,r_\nu,\sigma,\xi,\psi,y_1,\cdots,y_\nu)$ with respect to δ, $\delta=\sigma,\xi,\cdots,y_\nu$. The global Lipschitz condition on f ensures that these

derivatives (excluding f_σ) are bounded, so that, for almost all $t \in (a, b)$,

$$|f_\xi(\alpha, r_\nu, t, x(t), x_t, \cdots)\dot{x}(t)| \leqq m_1(t)|\dot{x}(t)|,$$
$$|f_\psi[\alpha, r_\nu, t, x(t), x_t, \cdots; \dot{x}_t]| \leqq m_1(t)|\dot{x}_t|,$$

and similarly for f_{y_i}. Therefore,

$$|\ddot{x}(t)| \leqq |f_\sigma(\alpha, r_\nu, t, x(t), x_t, \cdots)| + cm_1(t) + |\dot{g}(t)|$$

almost everywhere on (a, b), where $m_1, \dot{g} \in L_2^n(a, b)$, and c is a constant. Using (H3) we thus obtain that $\ddot{x} \in L_2^n(a, b)$ and hence it follows that $\ddot{x} \in L_2^n(a-r, b)$ since $\ddot{x}(t) = \ddot{\phi}(t-a)$, $t \in (a-r, a)$, $\ddot{\phi} \in L_2^n(-r, 0)$. □

Essential to our convergence proofs are certain standard estimates from the theory of spline approximations, in particular the Schmidt inequality and the results from [16, Thm. 2.5]. These inequalities are stated in the next lemma.

LEMMA 3.3. *Let $z = (\psi(0), \psi)$ be given in $\mathcal{S}$, and denote by $(\psi^N(0), \psi^N)$ the element $P^N z$ of X^N. Then the following estimates may be obtained for N sufficiently large:*

$$|P^N z - z|_N \leqq \frac{k_1}{N^2}|\mathcal{D}^2\psi|, \tag{3.6}$$

$$|\mathcal{D}\psi^N - \mathcal{D}\psi|_N \leqq \frac{k_2}{N}|\mathcal{D}^2\psi|, \tag{3.7}$$

$$|\psi^N(\Theta) - \psi(\Theta)| \leqq \left(\frac{k_1}{N^2} + \frac{r^{1/2}k_2}{N}\right)|\mathcal{D}^2\psi|, \qquad \Theta \in [-r_\nu^N, 0], \tag{3.8}$$

where k_1 and k_2 are positive constants independent of N and q^N.

Proof. We have

$$|P^N z - z|_N = |\pi^N i^N z - i^N z|_N \leqq |z_I^N - i^N z|_N = |\psi_I^N - \psi|_N,$$

where $z_I^N = (\psi_I^N(0), \psi_I^N)$, ψ_I^N the interpolating spline for $\psi \in H^2(-r, 0)$ with knots $\{t_j^N\}$. From [4, (6.10)] we use

$$|\psi_I^N - \psi|_N \leqq \frac{r^2\nu^{1/2}}{\pi^2 N^2}|\mathcal{D}^2\psi|,$$

thus obtaining (3.6). The calculations for the estimates in (3.7), (3.8) are found in [4, pp. 814–15]. □

These estimates may now be employed to show convergence of $z^N(t; \gamma^N, g)$ to $z(t; \tilde{\gamma}, g)$ (in the proper sense) when $z(t) \in \mathcal{S}$; i.e., when $\{\tilde{\zeta}, g\} \in \mathcal{I}(\tilde{q})$.

THEOREM 3.2. *Let $\{\gamma^N\}$ be arbitrary in Γ with $\gamma^N \to \tilde{\gamma}$, $\gamma^N = (\zeta^N, q^N)$, $\tilde{\gamma} = (\tilde{\zeta}, \tilde{q}) \in \Gamma$, where $\{\tilde{\zeta}, g\} \in \mathcal{I}(\tilde{q})$, and let $z^N(t; \gamma^N, g)$, $z(t; \tilde{\gamma}, g)$ denote the solutions to* (3.1) *and* (2.6) *associated with γ^N and $\tilde{\gamma}$ respectively. Then*

$$|z^N(t; \gamma^N, g) - P^N z(t; \tilde{\gamma}, g)|_N \to 0$$

as $N \to \infty$ uniformly in $t \in [a, b]$.

Proof. We have

$$\begin{aligned}\Delta^N(t) &\equiv z^N(t) - P^N z(t)\\ &= P^N\zeta^N + \int_a^t \{A^N(\sigma)z^N(\sigma) + P^N G(\sigma)\}\,d\sigma\\ &\quad - P^N\tilde{\zeta} - \int_a^t \{P^N A(\sigma)z(\sigma) + P^N G(\sigma)\}\,d\sigma\\ &= \Delta^N(a) + \int_a^t \{A^N(\sigma)z^N(\sigma) - P^N A(\sigma)z(\sigma)\}\,d\sigma\end{aligned}$$

so that from Lemma 2.1 we obtain

$$\begin{aligned}|\Delta^N(t)|_N^2 &= |\Delta^N(a)|_N^2 + 2\int_a^t \langle A^N(\sigma)z^N(\sigma) - A^N(\sigma)P^N z(\sigma), \Delta^N(\sigma)\rangle_N\,d\sigma\\ &\quad + 2\int_a^t \langle A^N(\sigma)P^N z(\sigma) - P^N A(\sigma)z(\sigma), \Delta^N(\sigma)\rangle_N\,d\sigma\\ &\leqq |\Delta^N(a)|_N^2 + 2\int_a^t \omega(\sigma)|\Delta^N(\sigma)|_N^2\,d\sigma\\ &\quad + 2\int_a^t |A^N(\sigma)P^N z(\sigma) - P^N A(\sigma)z(\sigma)|_N |\Delta^N(\sigma)|_N\,d\sigma\\ &\leqq |\Delta^N(a)|_N^2 + \int_a^t (2\omega(\sigma)+1)|\Delta^N(\sigma)|_N^2\,d\sigma\\ &\quad + \int_a^b |A^N(\sigma)P^N z(\sigma) - P^N A(\sigma)z(\sigma)|_N^2\,d\sigma.\end{aligned}$$

Gronwall's inequality may be employed (since the L_1 function $2\omega+1$ is positive and Δ^N is continuous in t) to obtain

$$|\Delta^N(t)|_N^2 \leqq (\varepsilon_1(N) + \varepsilon_2(N)) \exp \int_a^b (2\omega(\sigma)+1)\,d\sigma,$$

where

$$\begin{aligned}\varepsilon_1(N) &= |\Delta^N(a)|_N^2,\\ \varepsilon_2(N) &= \int_a^b |A^N(\sigma)P^N z(\sigma) - P^N A(\sigma)z(\sigma)|_N^2\,d\sigma.\end{aligned}$$

It remains to show that $\varepsilon_i(N) \to 0$ as $N \to \infty$; that the convergence is uniform in t is readily seen. First,

$$\varepsilon_1(N) = |P^N\zeta^N - P^N\tilde{\zeta}|_N^2 \leqq |i^N\zeta^N - i^N\tilde{\zeta}|_N^2$$

converges to 0 as $N \to \infty$ from the definition of convergence of γ^N to $\tilde{\gamma}$. We will also obtain $\varepsilon_2(N) \to 0$ once we demonstrate the dominated convergence of

$$|A^N(\sigma)P^N z(\sigma) - P^N A(\sigma)z(\sigma)|_N^2 \to 0.$$

Let $z(\sigma)=(y_\sigma(0), y_\sigma)$, $P^N z(\sigma)=(y_\sigma^N(0), y_\sigma^N)$, $y_\sigma^N \in L_2^n(-r_\nu^N, 0)$. Then

$$\begin{aligned}
&|A^N(\sigma)P^N z(\sigma)-P^N A(\sigma)z(\sigma)|_N^2 \\
&\quad=|\pi^N \tilde{A}(q^N,\sigma)\pi^N P^N z(\sigma)-\pi^N i^N A(\tilde{q},\sigma)z(\sigma)|_N^2 \\
&\quad=|\pi^N(F(q^N,\sigma,y_\sigma^N(0),\bar{y}_\sigma^N),\mathscr{D}y_\sigma^N)-\pi^N(F(\tilde{q},\sigma,y_\sigma(0),y_\sigma),\mathscr{D}y_\sigma)|_N^2 \\
&\quad\leqq|F(q^N,\sigma,y_\sigma^N(0),\bar{y}_\sigma^N)-F(\tilde{q},\sigma,y_\sigma(0),y_\sigma)|^2+|\mathscr{D}y_\sigma^N-\mathscr{D}y_\sigma|_N^2 \\
&\quad\equiv T_1^N(\sigma)+T_2^N(\sigma),
\end{aligned}$$

where $\bar{y}_\sigma^N$ is the usual extension of y_σ^N to all of $[-r, 0]$. From (3.7), $T_2^N(\sigma)\to 0$ as $N\to\infty$. Furthermore,

$$\begin{aligned}
(T_1^N(\sigma))^{1/2}&\leqq|F(q^N,\sigma,y_\sigma^N(0),\bar{y}_\sigma^N)-F(q^N,\sigma,y_\sigma(0),y_\sigma)| \\
&\quad+|F(q^N,\sigma,y_\sigma(0),y_\sigma)-F(\tilde{q},\sigma,y_\sigma(0),y_\sigma)| \\
&\equiv\tau_1^N(\sigma)+\tau_2^N(\sigma),
\end{aligned}$$

where $\tau_2^N(\sigma)\to 0$ as $N\to\infty$ since quite standard arguments may be used to show that the map

$$\begin{aligned}
&q=(\alpha, r_1, \cdots, r_\nu) \\
&\quad\to F(q,\sigma,\psi(0),\psi)=f(\alpha,r_\nu,\sigma,\psi(0),\psi,\psi(-r_1),\cdots,\psi(-r_\nu))
\end{aligned}$$

is continuous whenever ψ is continuous. In addition, $\tau_1^N(\sigma)$ is $O(1/N)$ (for almost all σ) from (3.6), (3.8) since, for almost all σ,

$$\begin{aligned}
&|F(q^N,\sigma,y_\sigma^N(0),\bar{y}_\sigma^N)-F(q^N,\sigma,y_\sigma(0),y_\sigma)| \\
&\qquad\leqq m_1(\sigma)\{|y_\sigma^N(0)-y_\sigma(0)|+|y_\sigma^N-y_\sigma|_N+\sum_{i=1}^{\nu}|y_\sigma^N(-r_i^N)-y_\sigma(-r_i^N)|\}
\end{aligned}$$

and $|y_\sigma^N-y_\sigma|_N\leqq|P^N z(\sigma)-z(\sigma)|_N$. Therefore, for almost all $\sigma\in[a, b]$, $T_1^N(\sigma)\to 0$ as $N\to\infty$, and the convergence (a.e.) to zero of the integrand of $\varepsilon_2(N)$ is assured. Dominated convergence follows from similar arguments:

$$|A^N(\sigma)P^N z(\sigma)-P^N A(\sigma)z(\sigma)|_N^2\leqq(\tau_1^N(\sigma)+\tau_2^N(\sigma))^2+T_2^N(\sigma)$$

as before where, from (3.7), $T_2^N(\sigma)\leqq k_2^2|\mathscr{D}^2 y_\sigma|^2\leqq k_2^2 M_0<\infty$,

$$M_0=\sup_{\sigma\in[a,b]}|\mathscr{D}^2 y_\sigma|^2=\sup_{\sigma\in[a,b]}\int_{\sigma-r}^{\sigma}|\ddot{y}(\Theta)|^2\,d\Theta\leqq\int_{a-r}^{b}|\ddot{y}(\Theta)|^2\,d\Theta=|\ddot{y}|_{L_2^n(a-r,b)}^2<\infty$$

(we have made use of Lemma 3.2 to assert that $z(\sigma)\in\mathscr{S}$ for all σ; i.e., $y\in H^2(a-r, b)$). The Lipschitz condition on f and estimates (3.6), (3.8) may be used to show

$$\tau_1^N(\sigma)\leqq cm_1(\sigma)M_0$$

for a constant $c>0$ and almost all σ. Finally,

$$\begin{aligned}
\tau_2^N(\sigma)&=|F(q^N,\sigma,y(\sigma),y_\sigma)-F(\tilde{q},\sigma,y(\sigma),y_\sigma)| \\
&\leqq 2\sup_{(q,\sigma)\in Q\times[a,b]}|F(q,\sigma,y(\sigma),y_\sigma)|,
\end{aligned}$$

where y was determined by a fixed $\tilde{q}\in Q$ and is thus independent of q and is continuous. Again the continuity of $F(q,\sigma,y(\sigma),y_\sigma)$ in (q,σ) may be easily established, (q,σ) in the compact set $Q\times[a, b]$, so that there is some (q^*,σ^*) in $Q\times[a, b]$ such that

$$\tau_2^N(\sigma)\leqq 2|F(q^*,\sigma^*,y(\sigma^*),y_{\sigma^*})|\equiv M_1<\infty.$$

It follows then that, for almost all $\sigma \in [a, b]$,

$$|A^N(\sigma)P^N z(\sigma) - P^N A(\sigma) z(\sigma)|_N^2 \leq (cm_1(\sigma)M_0 + M_1)^2 + k_2^2 M_0 \equiv h(\sigma),$$

where $h \in L_1(a, b)$ (since $m_1 \in L_2(a, b)$). Thus the theorem is proved. □

We now turn to the main state variable convergence result for arbitrary $\{\tilde{\zeta}, g\} \in W \times L_2^n(a, b)$; it contains the key arguments needed to prove Theorem 3.4 below, which describes how solutions $\bar{\gamma}^N$ (to $\mathcal{P}_N$) converge to $\bar{\gamma}$, a solution to the original parameter estimation problem.

THEOREM 3.3. *Suppose* $\gamma^N \to \tilde{\gamma}$ *where* $\gamma^N = (\zeta^N, q^N)$ *and* $\tilde{\gamma} = (\tilde{\zeta}, \tilde{q})$ *are arbitrary in* Γ. *Then for any* $g \in L_2^n(a, b)$,

$$\pi_0 z^N(t; \gamma^N, g) \to \pi_0 z(t; \tilde{\gamma}, g)$$

as $N \to \infty$ *uniformly in* $t \in [a, b]$.

Proof.

$$\begin{aligned} |\pi_0 z^N(t; \gamma^N, g) - \pi_0 z(t; \tilde{\gamma}, g)| &\leqq |\pi_0 z^N(t; \gamma^N, g) - \pi_0 P^N z(t; \tilde{\gamma}, g)| \\ &\quad + |\pi_0 P^N z(t; \tilde{\gamma}, g) - \pi_0 z(t; \tilde{\gamma}, g)| \\ &\equiv T_1^N(t) + T_2^N(t), \end{aligned}$$

where $T_2^N(t) \to 0$ as $N \to \infty$ uniformly in $t \in [a, b]$ from the convergence $\pi_0 P^N z \to \pi_0 z$, $z \in Z$, demonstrated in [4, p. 814]. (Uniformity here is due to the fact that $z \in \{z(t; \tilde{\gamma}, g) \mid t \in [a, b]\}$, a compact set in Z.) Further, since $\mathcal{I}(\tilde{q})$ is dense in $W \times L_2^n(a, b)$, a pair $\{\hat{\zeta}, \hat{g}\}$ may be chosen in $\mathcal{I}(\tilde{q})$ arbitrarily close to $\{\tilde{\zeta}, g\}$ so that, given that

$$\begin{aligned} T_1^N(t) &\leqq |z^N(t; (\zeta^N, q^N), g) - P^N z(t; (\tilde{\zeta}, \tilde{q}), g)|_N \\ &\leqq |z^N(t; (\zeta^N, q^N), g) - z^N(t; (\mathscr{C}_N^+ i^N \hat{\zeta}, q^N), \hat{g})|_N \\ &\quad + |z^N(t; (\mathscr{C}_N^+ i^N \hat{\zeta}, q^N), \hat{g}) - P^N z(t; (\hat{\zeta}, \tilde{q}), \hat{g})|_N \\ &\quad + |P^N z(t; (\hat{\zeta}, \tilde{q}), \hat{g}) - P^N z(t; (\tilde{\zeta}, \tilde{q}), g)|_N, \end{aligned}$$

the first and third terms may be made as small as desired from the continuous dependence of z^N, z on $\{i^N \zeta, g\} \in \boldsymbol{X}(q^N) \times L_2^n(a, b)$ and $\{\zeta, g\} \in Z \times L_2^n(a, b)$ respectively, uniform in N and t (we may use this result for the first term since $|\zeta^N - \mathscr{C}_N^+ i^N \hat{\zeta}|_{X,q} = |i^N \zeta^N - i^N \hat{\zeta}|_{X,q} \leqq |i^N \zeta^N - i^N \tilde{\zeta}|_N + |i^N \tilde{\zeta} - i^N \hat{\zeta}|_N$ is arbitrarily small from the convergence of ζ^N to $\tilde{\zeta}$). Finally, the middle term goes to 0 uniformly in $t \in [a, b]$ as $N \to \infty$ since $\{\hat{\zeta}, \hat{g}\} \in \mathcal{I}(\tilde{q})$ and the parameters involved, $(\mathscr{C}_N^+ i^N \hat{\zeta}, q^N)$, converge to $(\hat{\zeta}, \tilde{q})$ in the sense required ($|\tilde{q} - q^N| \to 0$ and $|i^N \hat{\zeta} - i^N (\mathscr{C}_N^+ i^N \hat{\zeta})|_N = 0$), so that we are guaranteed the uniform convergence of $\pi_0 z^N(t; \gamma^N, g)$ to $\pi_0 z(t; \tilde{\gamma}, g)$. □

3.2. Convergence of parameters. Our attention to this point has been focused on the convergence of solutions z^N (to (3.1)) to the solution z (to (2.6)), once the convergence of any sequence of parameters has been established. In reality, though, we have yet to determine that any sequence of solutions $\{\bar{\gamma}^N\}$ to $\mathcal{P}_N$ is in fact convergent; even then, we must prove that the limiting value $\bar{\gamma}$ is indeed a solution to the original parameter identification problem. The result we now state addresses this question and indicates when an approximate ID problem $\mathcal{P}_N$ may be used to compute numerical solutions for the original problem.

THEOREM 3.4. *Let* $\{\bar{\gamma}^N\}$, $\bar{\gamma}^N \in \Gamma$, *be a sequence of solutions to the approximate parameter estimation problems* $\mathcal{P}_N$. *Then there exists* $\bar{\gamma} \in \Gamma$ *and a subsequence* $\{\bar{\gamma}^{N_k}\}$

such that $\bar{\gamma}^{N_k} \to \bar{\gamma}$ *and, if* $\bar{\gamma}$ *and* S *satisfy hypotheses* (H6) *and* (H7), $\bar{\gamma}$ *is a solution to the original parameter identification problem.*

Proof. From (H7), the sequence $\{\mathscr{C}_N^+ i^N \bar{\zeta}^N\}$ belongs to S and S is compact in the Z topology, so that a subsequence satisfies $|\mathscr{C}_{N_k}^+ i^{N_k} \bar{\zeta}^{N_k} - \bar{\zeta}| \to 0$ for some $\bar{\zeta} \in S$. The compactness of Q guarantees the convergence of a subsequence of $\{\bar{q}^{N_k}\}$, $\bar{q}^{N_{k_j}} \to \bar{q}$ for some $\bar{q} \in Q$. Relabelling as $\bar{\gamma}^{N_k}$, we have a sequence $\bar{\gamma}^{N_k} = (\bar{\zeta}^{N_k}, \bar{q}^{N_k})$ in Γ that converges to $\bar{\gamma} = (\bar{\zeta}, \bar{q})$ in the required sense because $\bar{q}^{N_k} \to \bar{q}$ and

$$|i^{N_k} \bar{\zeta}^{N_k} - i^{N_k} \bar{\zeta}|_{N_k} \leqq \nu^{1/2} |i^{N_k} \mathscr{C}_{N_k}^+ i^{N_k} \bar{\zeta}^{N_k} - i^{N_k} \bar{\zeta}|_{X,q}$$
$$\leqq \nu^{1/2} |\mathscr{C}_{N_k}^+ i^{N_k} \bar{\zeta}^{N_k} - \bar{\zeta}| \to 0.$$

It remains to show that $\bar{\gamma}$ is a solution to the original ID problem. We have (see (2.8) and J^N in the definition of $\mathscr{P}_N$)

$$J(\bar{\gamma}) = \frac{1}{2} \sum_{i=1}^{M} |C(\bar{q}) \pi_0 z(t_i; \bar{\gamma}, g) - \hat{u}_i|^2 = \lim_{N_k \to \infty} J^{N_k}(\bar{\gamma}^{N_k}) \leqq \lim_{N_k \to \infty} J^{N_k}(\gamma),$$

where the continuity of C and the convergence of $\pi_0 z^{N_k}(t; \bar{\gamma}^{N_k}, g)$ to $\pi_0 z(t; \bar{\gamma}, g)$ is used to obtain the second expression and the final inequality holds for any $\gamma \in \Gamma$ since $\bar{\gamma}^{N_k}$ is a solution to $\mathscr{P}_{N_k}$. The convergence of $\pi_0 z^{N_k}(t; \gamma, g)$ to $\pi_0 z(t; \gamma, g)$ for any $\gamma \in \Gamma$ also follows from Theorem 3.3, so it follows that $J^{N_k}(\gamma) \to J(\gamma)$ as $N_k \to \infty$, or that

$$J(\bar{\gamma}) \leqq J(\gamma)$$

for any $\gamma \in \Gamma$. Thus $\bar{\gamma}$ is a solution to the original identification problem. □

4. Numerical results. In this concluding section we present a sample of numerical findings obtained using the spline approximation estimation schemes discussed above. The test examples we investigated were chosen with certain types of applications and/or difficulties in mind. Example 4.1 deals with a nonlinear pendulum (small oscillations are *not* assumed) with damping through a linear feedback on the velocity, i.e., $U(\dot{x}) = k\dot{x}$. We assume the existence of actuator delays in effecting the feedback laws. (Delayed damping and delayed restoring forces are quite common in mechanical systems—see [15, Chapt. 21].) A possible application is associated with the design of a damped "pendulum" to "track" a given course or program $\hat{x}(t)$. Example 4.2 involves a nonlinear nonautonomous multiple delay equation in which the nonlinearity is of the Michaelis–Menten, Briggs–Haldane velocity approximation type. Such nonlinearities occur in biological applications in which enzyme mediated reactions must be modeled. Our third example concerns a linear multiple delay system with unknown coefficients such as might arise in multi-compartment transport models, while Example 4.4 contains a nonlinearity that is only locally Lipschitz and thus it does not satisfy the hypotheses detailed above. It is interesting (although not at all surprising) to observe that the methods under investigation also perform admirably when applied to examples of this type. Indeed, we believe that a convergence theory for problems with only locally Lipschitz systems could be developed, but the technical details would be even more unpleasant than those in the theory presented above.

The computations reported below were performed on the IBM 370/158 at Brown University. The goal of our numerical efforts was to test convergence properties of the estimation algorithm on selected examples. This was done in the following manner. "True" values of the parameters to be estimated were chosen and an independent method was used to integrate the systems with these values. These "exact" solutions or these solutions with random noise added were used as observed "data" (a number

of "sample" data points were chosen) and the spline-based methods were employed with a least squares criterion. For a given N, an IMSL package (ZXSSQ) for the Levenberg–Marquardt method was used to iteratively find the corresponding parameters. Our experience with this package has been very positive; convergence often was obtained after relatively few iterations (e.g., less than 10). In most cases, we employed the default values for the several input parameters required in the IMSL version of the Levenberg–Marquardt algorithm.

Example 4.1 (nonlinear pendulum with delayed damping).

We considered the system

$$\ddot{x}(t)+k\dot{x}(t-r)+\left(\frac{g}{l}\right)\sin x(t)=0, \qquad 0\leqq t\leqq 7,$$

$$x(\Theta)=1, \qquad \Theta\leqq 0,$$

$$\dot{x}(\Theta)=0, \qquad \Theta\leqq 0.$$

"Data" consisting of 28 sample points at times in $[0, 7]$ were generated for "true" values $\bar{r}=2$, $\bar{k}=4$, and $\overline{g/l}=9.81$. Several different estimation problems were investigated.

(a) We sought to estimate r with $k=\bar{k}$, $g/l=\overline{g/l}$ given (start-up value: $r^0=2.5$). We denoted by $\bar{r}^N$ the "converged" values for r corresponding to a fixed value N of the approximation index. (See Table 1.)

TABLE 1.

N	$\bar{r}^N$
2	2.429
4	2.412
8	1.908
16	2.003
32	2.002

(b) We estimated r, g/l with $k=\bar{k}$ given (start-up values: $r^0=2.2$, $(g/l)^0=8.6$). For $N=16$, we obtained $\bar{r}^{16}=2.002$, $\overline{g/l}^{16}=9.84$.

(c) We estimated r, k with $g/l=\overline{g/l}$ given (start-up values: $r^0=2.5$, $k^0=8.0$). For $N=16$, we obtained $\bar{r}^{16}=1.999$, $\bar{k}^{16}=3.977$.

Example 4.2. The nonlinear nonautonomous multiple delay equation for consideration is

$$\dot{x}(t)=-tx(t)+2x(t-r_1)+\frac{3x(t-r_2)}{K+x(t-r_2)}, \qquad 0\leqq t\leqq 4,$$

$$x(\Theta)=\begin{cases}-m\Theta, & -2\leqq\Theta\leqq 0,\\ 20+m\Theta, & -4\leqq\Theta\leqq -2.\end{cases}$$

"Data" were generated for 16 sampling times in $[0, 4]$ using true values $\bar{r}_1=1$, $\bar{r}_2=2$, $\bar{K}=10$, $\bar{m}=5$. The following problems were studied and results obtained.

(a) We estimated r_1, r_2 with $K=\bar{K}$, $m=\bar{m}$ (start-up values: $r_1^0=.5$, $r_2^0=2.5$). (See Table 2.)

TABLE 2.

N	$\bar{r}_1^N$	$\bar{r}_2^N$
2	1.055	1.600
4	1.013	1.896
8	1.007	1.943
16	.9995	2.003
32	.9998	2.003

(b) We estimated K for $r_1 = \bar{r}_1$, $r_2 = \bar{r}_2$, $m = \bar{m}$ (start-up value: $K^0 = .05$). (See Table 3.)

TABLE 3.

N	$\bar{K}^N$
2	8.345
4	9.706
8	9.816
16	10.027
32	9.9998

(c) We estimated m for $r_1 = \bar{r}_1$, $r_2 = \bar{r}_2$, $K = \bar{K}$ (start-up value: $m^0 = -4.0$). (See Table 4.)

TABLE 4.

N	$\bar{m}^N$
2	5.114
4	5.028
8	5.014
16	4.998
32	4.999

(d) We repeated the calculations of (c) except that we corrupted the data with random noise (Gaussian with zero mean and standard deviation $\sigma = .1$). (See Table 5.)

TABLE 5.

N	$\bar{m}^N$
2	5.059
4	4.973
8	4.956
16	4.940
32	4.940

Example 4.3. We consider next the linear multiple delay example

$$\dot{x}(t) = -\tfrac{1}{2}x(t) + \beta x(t - r_1) + x(t - r_2), \qquad 0 \leqq t \leqq 3,$$

$$x(\Theta) = \alpha\Theta^2 - 3\Theta, \qquad -4 \leqq \Theta \leqq 0.$$

"True" values of $\bar{\beta} = 3$, $\bar{r}_1 = 1$, $\bar{r}_2 = 2$, $\bar{\alpha} = -.75$ were used to produce 24 data points on the interval [0, 3].

(a) We estimated α for $\beta = \bar{\beta}$, $r_1 = \bar{r}_1$, $r_2 = \bar{r}_2$, (start-up value: $\alpha^0 = 5.0$). (See Table 6.)

TABLE 6.

N	$\bar{\alpha}^N$
2	−.661
4	−.724
8	−.742
16	−.748
32	−.749

(b) We estimated r_1, r_2, β with $\alpha = \bar{\alpha}$ (start-up values: $r_1^0 = 1.3$, $r_2^0 = 1.7$, $\beta^0 = 3.5$). (See Table 7.)

TABLE 7.

N	$\bar{r}_1^N$	$\bar{r}_2^N$	$\bar{\beta}^N$
2	1.1233	1.600	3.1642
4	1.0028	1.957	3.0323
8	.9993	2.009	3.0064
16	.9996	2.005	3.0007
32	.9998	2.002	3.0000

(c) We repeated the calculations of (b) with data that had been corrupted by noise. (See Table 8.)

TABLE 8.

N	$\bar{r}_1^N$	$\bar{r}_2^N$	$\bar{\beta}^N$
2	1.096	1.600	3.152
4	.9998	1.970	3.023
8	.9940	2.024	2.994
16	.9934	2.025	2.987
32	.9941	2.023	2.987

Example 4.4. As our final example, we present a multiple delay equation with nonlinearity satisfying only a local Lipschitz condition.

$$\dot{x}(t) = -1.5x(t) - 1.25x(t - r_1) + cx(t - r_2)\sin x(t - r_2), \qquad 0 \leqq t \leqq 5,$$

$$x(\Theta) = 10\Theta + 1, \qquad \Theta \leqq 0.$$

True values were $\bar{c} = 1$, $\bar{r}_1 = 1$, $\bar{r}_2 = 2$, and data were generated corresponding to 20 sampling times in [0, 5]. We estimated r_1, r_2, c with start-up values of $r_1^0 = 1.4$, $r_2^0 = 2.2$, $c^0 = .2$. (See Table 9.)

TABLE 9.

N	$\bar{r}_1^N$	$\bar{r}_2^N$	$\bar{c}^N$
2	1.0814	1.9863	1.0606
4	1.0537	1.9900	.9757
8	.9998	1.9906	.9745
16	.9992	1.9993	.9981
32	.9996	1.9995	.9986

5. Appendix. Notation.

$|\cdot|$ = standard norm on R^n, $L_2^n(-r, 0)$, or more generally $L_2^n(a, b)$, or on $Z = R^n \times L_2^n(-r, 0)$,
$|\cdot|_{r_\nu}$ = standard norm on $L_2^n(-r_\nu, 0)$,
$|\cdot|_q$ = $\tilde{\rho}(q)$ weighted norm on Z,
$|\cdot|_{X,q}$ = standard norm on $X(q) = R^n \times L_2^n(-r_\nu, 0)$,
$|\cdot|_{\rho,q}$ = $\rho(q)$ weighted norm on $X(q)$,
$|\cdot|_N$ = $\rho(q^N)$ weighted norm on either $L_2^n(-r_\nu^N, 0)$ or $X(q^N)$.

REFERENCES

[1] H. T. BANKS, *Identification of nonlinear delay systems using spline methods*, Proc. International Conf. on Nonlinear Phenomena in the Math. Sciences, Univ. Texas, Arlington, June 16–20, 1980, Academic Press, New York, 1982, pp. 47–55.

[2] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, this Journal, 16 (1978), pp. 169–208.

[3] H. T. BANKS, J. A. BURNS AND E. M. CLIFF, *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.

[4] ———, *Parameter estimation and identification for systems with delays*, this Journal, 19 (1981), pp. 791–828.

[5] H. T. BANKS AND P. L. DANIEL, *Parameter estimation of nonlinear nonautonomous distributed systems*, Proc. 20th IEEE Conference on Decision and Control, San Diego, Dec. 16–18, 1981, pp. 228–232.

[6] H. T. BANKS AND F. KAPPEL, *Spline approximations for functional differential equations*, J. Differential Equations, 34 (1979), pp. 496–522.

[7] H. T. BANKS AND I. G. ROSEN, *Spline approximations for linear nonautonomous delay systems*, ICASE Rep. No. 81-33, Oct. 1981, NASA Langley Research Center, Hampton, VA, J. Math. Anal. Appl., to appear.

[8] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, 1976.

[9] P. L. DANIEL, *Spline approximations for nonlinear hereditary control systems*, ICASE Rep. No. 82-10, NASA Langley Research Center, Hampton, VA, April 1982, J. Opt. Theory Appl., to appear.

[10] ———, *Spline-based approximation methods for the identification and control of nonlinear functional differential equations*, Ph.D. thesis, Brown Univ., Providence, RI, June, 1981.

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

[12] J. K. HALE, *Ordinary Differential Equations*, Interscience, New York, 1969.

[13] F. KAPPEL, *An approximation scheme for delay equations*, Proc. International Conf. on Nonlinear Phenomena in the Math Sciences, Univ. Texas, Arlington, June 16–20, 1980, Academic Press, 1982, pp. 585–595.

[14] ———, *Spline approximation for autonomous nonlinear functional differential equations*, J. Nonlinear Analysis, to appear.

[15] N. MINORSKY, *Nonlinear Oscillations*, Van Nostrand, New York, 1962.

[16] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0006 $01.25/0

FINITE DIMENSIONAL DETERMINISTIC NONLINEAR FILTERS VIA RICCATI TRANSFORMATION AND VOLTERRA SERIES*

MARKKU T. NIHTILÄ†

Abstract. Filtering in two classes of nonlinear differential and difference systems is studied. The system structures admit the representation of the optimal recursive filter in the form of a finite dimensional differential or difference system. The filtering problem is posed as a set of fixed interval optimization problems. The deterministic least-squares problem statement results in the filter which is an ordinary nonstochastic differential or difference system driven by the observed signal. No stochastic concepts are used. The system classes considered are described by two linear subsystems with a polynomial link map between them. In one class the link map affects the input of the latter subsystem. In the other class the coefficient matrix of the latter subsystem is a polynomial in the state of the preceding subsystem satisfying a Lie algebraic nilpotency condition. The former class is a special case of the systems described by finite Volterra series for which finite dimensionality of the optimal filter is also proved. Some examples and comparisons on the basis of examples with the corresponding stochastic systems admitting finite dimensional optimal conditional mean filters studied recently are performed. The comparisons show that except in some special cases including a linear one the deterministic least-squares filter and the stochastic conditional mean filter are not equivalent.

Key words. nonlinear least-squares filtering, finite Volterra series, Riccati transformation

1. Introduction. A deterministic approach to nonlinear filtering, both in time-discrete and time-continuous cases, is applied here. The filtering problem is formulated as a set of fixed interval optimization problems (FIOP), the final values of the optimal trajectories of which give the filtering. The formulation is the same as originally was used by Detchmendy and Sridhar [1], Bellman et al. [2], and later on by Mortensen [3]. Via Pontryagin's principle the optimization problem is converted into a two-point boundary-value problem (TPBVP). The solving of the TPBVP in a recursive finite dimensional form for some classes of systems as the final time t (or k) goes from 0 to T (or from 1 to k_f) is the main result of this paper.

The system classes to be considered cover the classes studied via stochastic formulation by Marcus et al. [4], [5]. The key mathematical machinery to be used is in part adopted from Marcus and Willsky [4] and from Brockett [6].

Filtering in the following nonlinear systems is studied.

$$\text{(1)}\quad 1:\ \dot{x} = A(t)x + B(t)w,$$

$$\text{(2)}\quad \dot{z} = [F(t) + Q(x)]z + P(t, x),$$

$$\text{(3)}\quad g(t, y(t), x(t)) = v(t),$$

$$\text{(4)}\quad 2:\ x_{i+1} = A_i x_i + B_i w_i,$$

$$\text{(5)}\quad z_{i+1} = F_i z_i + P(i, x_i),$$

$$\text{(6)}\quad g(i, y_i, x_i) = v_i,$$

where the state-linear output mapping g is defined by

$$\text{(7)}\qquad g(\sigma, y, x) = G_0(\sigma, y) + G_1(\sigma, y)x.$$

In (2) F is constant and P is identically zero, or Q equals to zero. The state components x and z are of the dimension n and n_1, respectively. The observed signal y has values

* Received by the editors March 23, 1982, and in revised form August 24, 1982.

† Department of Electrical Engineering, Control Engineering Laboratory, Helsinki University of Technology, Otakaari 5A, SF-02150 Espoo 15, Finland.

in $\mathbf{R}^m$. w and v stand for the unknown system and generalized observation errors of the dimension l and m_1, respectively. Furthermore, w and v are functions of the time only, not stochastic processes. Consequently, the systems studied are ordinary differential and difference systems. P and Q are polynomials in x; P arbitrary time-varying and Q a specified time-invariant one, with values in n_1-vectors and $n_1 \times n_1$ matrices, respectively. Furthermore, it is assumed that $\{A, B\}$ is completely controllable and $\{A, G_1(\cdot, y(\cdot))\}$ completely observable along the given observation y. G_0 and G_1 are appropriate smooth functions of σ and y with values in m_1-vectors and $m_1 \times n$ matrices, correspondingly.

The observation equation (3), or (6), expresses the dependence of the observation y on the state x and on the generalized observation error v. The dimension m_1 of the generalized error need not be the same as the dimension m of the observation. However, (3), or (6), must possess a feasible (not necessarily unique) solution for the given triple $(t, x(t), v(t))$, or (i, x_i, v_i).

The partial problem (1) and (3), and (4) and (6), with a state-linear output mapping (7), was studied in [7], [8].

It is seen by setting

$$g(\sigma, y, x) = y - H(\sigma)x,$$

that the state-linear output mapping also includes the standard linear case.

Assuming certain properties for the polynomial mappings P and Q the finite dimensionality of the filters is proved by applying a Riccati transformation technique or a successive differentiation method in the relevant multidimensional Volterra series. Comparisons on the basis of some examples with the corresponding stochastic finite dimensional conditional mean filters developed by Marcus et al. [4], [5] are performed. A similar polynomial structure of the optimal discrete filter with respect to the innovation was also found via the deterministic formulation. The comparisons show that the deterministic filters and the conditional mean filters are not for the given systems formally equivalent[1] although special cases exist. The formal equivalence means here that formally the same system equations generate formally the same filters. In the linear case it has been proved by Fleming and Rishel [9] that the conditional mean filtering problem has a formally equivalent deterministic counterpart. In nonlinear problems this equivalence breaks down.

The organization of the paper is as follows. Section 2 presents the problem statement in a deterministic framework for both continuous and discrete time systems. Some motivation of the nonstatistical least-squares approach is also discussed. Section 3 gives the main results in the form of three theorems. Some additional lemmas and proofs of the theorems are included. The proof of the second theorem is based on the finite generalized Volterra series defined and studied in the appendix which also includes a side result concerning finite dimensional computability of the optimal filtering for systems described by finite generalized Volterra series. Some Lie algebraic concepts are also introduced in the appendix. Section 4 is devoted to applications. Concluding remarks complete the paper.

[1] The word *formally* is here used due to the fact that there exists a very essential difference between deterministic and classical stochastic approaches. The deterministic formulation produces a deterministic filter which is an ordinary differential or difference system the input of which is the observed signal. Stochastic filters in Marcus et al. [4], [5] are stochastic differential or difference equations driven by the stochastic observation process.

2. Problem statement and motivation.

2.1. Deterministic formulation. In the following both continuous and discrete filtering problems are posed as sets of fixed interval optimization problems.

Filtering problem 2.1. For every final time $t \in [0, T]$ $(T < \infty)$ and for a given observation y on the time interval $[0, t]$ find the final value of the optimal trajectory of the fixed interval optimization problem: Minimize the performance index

$$\mathbf{J}(t, x, z, d) = \tfrac{1}{2}\|x(0) - \xi\|_J^2 + \tfrac{1}{2}\|z(0) - \zeta\|_{\tilde{J}}^2 + \tfrac{1}{2}\int_0^t [\|g(\sigma, y(\sigma), x(\sigma))\|_{R(\sigma)}^2 + \|d(\sigma)\|_{S(\sigma)}^2]\, d\sigma \tag{8}$$

with respect to (x, z, d) subject to the constraint

$$\dot{x}(\sigma) = A(\sigma)x(\sigma) + B(\sigma)d(\sigma), \tag{9}$$

$$\dot{z}(\sigma) = [F(\sigma) + Q(x(\sigma))]z(\sigma) + P(\sigma, x(\sigma)), \tag{10}$$

where g is given by (7).

Filtering problem 2.2. For every final time $k \in [0, k_f]$ $(\triangleq\{0, 1, \cdots, k_f\})$ and for a given observation sequence y_i, $i = 1, 2, \cdots, k$, find the final value x_k of the optimal trajectory of the fixed interval optimization problem: Minimize

$$\mathbf{J}(k, x, z, d) = \tfrac{1}{2}\|x_0 - \xi\|_J^2 + \tfrac{1}{2}\|z_0 - \zeta\|_{\tilde{J}}^2 + \tfrac{1}{2}\sum_{i=1}^{k} [\|g(i, y_i, x_i)\|_{R_i}^2 + \|d_{i-1}\|_{S_{i-1}}^2] \tag{11}$$

with respect to $(x, z, d)_i$, $i = 0, 1, \cdots, k$, subject to

$$x_{i+1} = A_i x_i + B_i d_i, \tag{12}$$

$$z_{i+1} = F_i z_i + P(i, x_i), \tag{13}$$

where g is given by (7).

J, $\tilde{J}$ and values of S and R are positive definite weighting matrices of appropriate dimensions. ξ and ζ are the given (approximate) initial state vectors. $\|\cdot\|_J$ denotes a norm of $\mathbf{R}^n$ defined by $\|\xi\|_J^2 = \xi^T J \xi$.

2.2. Motivation. The problem statements can be motivated as follows. In the deterministic filtering the objective is to get the response of the given state system which is disturbed (by the pseudo-control d) *as slightly as possible* driven through the generalized observation equation *as close as possible* to the observed signal on the whole time interval. The two goals are in disagreement if we have not a common measure for them. So, our selection is to minimize the weighted sum of the squared pseudo-control d and the generalized observation error, or residual $g(\sigma, y, x)$ integrated (or summed up) over the observation interval. Almost whatever reasonable measure instead of the chosen quadratic measure could be used. However the chosen measure leads into relatively simple algorithms; in the purely linear case into the finite dimensional formal Kalman–Bucy filter, and expecially in the given system classes also into finite dimensional filters. It has to be emphasized that the selection of the quadratic measure corresponds in the stochastic formulation to Gaussian assumptions of the process and observation noise. Furthermore, the weighting matrices in the cost functional can be interpreted as the inverses of the noise covariances in the stochastic formulation. From a practical viewpoint the weighting matrices are tuning parameters in the filters.

2.3. Denote the optimal solution of the FIOP for every fixed t (or k) by $(x(\cdot|t), z(\cdot|t))$ and $(x(\cdot|k), z(\cdot|k))$. The total time derivative of the solution of the continuous problem is given by

$$\dot{x}(t|t) = x_\sigma(t|t) + x_t(t|t), \tag{14}$$

$$\dot{z}(t|t) = z_\sigma(t|t) + z_t(t|t). \tag{15}$$

The subindices σ and t denote partial derivatives with respect to the first and second argument position, respectively.

Remark 2.1. It is seen that both problems are split into a state-linear filtering problem and a feedforward problem. In the continuous case (for instance for $Q(x) = 0$) we have to solve the feedforward problem

$$z_\sigma(\sigma|t) = F(\sigma)z(\sigma|t) + P(\sigma, x(\sigma|t)), \tag{16}$$

$$z(0|t) = \zeta, \tag{17}$$

to get $z(t|t)$. In principle it could be obtained by integrating the equations of the linear two-point boundary-value problem (see e.g. [7]), which corresponds to the state-linear subproblem, from the known final values $(x(t|t), 0)$ backwards to the initial time. This would give the function $x(\cdot|t)$ which then could be used to obtain $z(\cdot|t)$ (i.e. $z(t|t)$). However, this procedure is neither time-recursive nor practical because the procedure should be performed separately for every final time $t \in [0, T]$ due to the fact that in general $x(\sigma|t_1) \neq x(\sigma|t_2)$ for $t_1 \neq t_2$.

DEFINITION 2.1. The filtering $s(t) \triangleq z(t|t)$ (or $s_k \triangleq z(k|k)$) is finite dimensionally computable if it can be obtained from a finite dimensional differential (or difference) system with a fixed initial state driven by the observed signal $y(t)$ (or y_k).

Remark 2.2. The inclusion of an additional disturbance term $G(\sigma)\tilde{d}(\sigma)$, or $G_i\tilde{d}_i$, in the feedforward system (10), or (13), and the inclusion of the corresponding quadratic term $\frac{1}{2}\|\tilde{d}(\sigma)\|^2_{\tilde{S}(\sigma)}$, or $\frac{1}{2}\|\tilde{d}_{i-1}\|^2_{\tilde{S}_{i-1}}$, in the performance index do not extend the problem classes studied via the deterministic formulation. This is due to the fact that the optimal $\tilde{d}$ in the both cases would be the zero function.

3. Main results.

THEOREM 3.1. *The filtering $s(t)$ is finite dimensionally computable if $Q = 0$ and P is an arbitrary polynomial.*

THEOREM 3.2. *The filtering $s(t)$ is finite dimensionally computable if $P = 0$ and Q is a polynomial given by*

$$Q(x) = \sum_{k=1}^{\nu} \sum_{\substack{i=1 \\ \sum \nu_j = k}}^{\binom{n+k-1}{k}} F_{j(k,i)} x_1^{\nu_1} \cdots x_n^{\nu_n} \tag{18}$$

where ν is the degree of the polynomial Q, and if the Lie algebra generated by the n_1 by n_1 matrices $\{F_j\}$ is nilpotent.[2]

Without loss of generality the constant term F_0 has been omitted from the polynomial. This is based on the result by Marcus and Willsky [4, Lemma 3.2]. For subsequent purposes we express the polynomial Q in the form

$$Q(x) = \sum_{j=1}^{N} F_j u_j, \tag{19}$$

where $u \in \mathbf{R}^N$ is formed by homogeneous terms in x in the obvious way.

[2] The index $j(k, i)$ is given by $j(k, i) = \sum_{l=1}^{k-1} \binom{n+l-1}{l} + i$, $i \leqq \binom{n+k-1}{k}$.

THEOREM 3.3. *The filtering s_k is finite dimensionally computable if P is an arbitrary polynomial.*

The proof of Theorems 3.1 and 3.3 is straightforward. It is based on the application of a polynomial type Riccati transformation in (10) and (13), respectively, and on the following lemma.

LEMMA 3.1. *The solution of the state-linear subproblem in the continuous case, i.e. the state-linear filter, is obtained by applying the linear Riccati transformation*

$$x(\sigma|t) = r(\sigma) - K(\sigma)p(\sigma|t) \tag{20}$$

in the corresponding TPBVP. *K is an $n \times n$ matrix valued function, and p is the co-state of the* TPBVP.

The proof of Lemma 3.1 is presented in [8]. It turns out that $r(t)$ is independent of p, and the pair (r, K) satisfies the deterministic equations of the state-linear filter driven by the observed signal $y(t)$, i.e. $r(t)$ is finite dimensionally computable. The proof of Theorem 3.3 is analogous to that of Theorem 3.1. So, it is omitted.

Proof of Theorem 3.1. Let the degree of the polynomial P be ν. Define a νth degree Riccati transformation

$$z(\sigma|t) = s(\sigma) - M(\sigma, p(\sigma|t)) \tag{21}$$

where M is a νth degree polynomial on $\mathbf{R}^n$ with differentiable tensor coefficients (see Appendix)

$$M(\sigma, p) = \sum_{i=1}^{\nu} \frac{1}{i!} M^i(\sigma) p^i. \tag{22}$$

By differentiating (21) with respect to σ and by substituting the result and (20) into (16) a new νth degree polynomial $\tilde{M}(\sigma, p)$ is obtained. By setting all the coefficient tensors $\tilde{M}^i$, $i = 1, 2, \cdots, \nu$, and the zeroth order term equal to zero we obtain $\nu + 1$ coupled ordinary inhomogeneous differential systems for $s(\cdot)$ and $M^i(\cdot)$, $i = 1, 2, \cdots, \nu$, with the time parameter σ. By setting $\sigma = t$ we obtain a finite dimensional system for the filtering $s(t)$. The initial values are obtained at $t = 0$ from (17):

$$s(0) = \zeta, \tag{23}$$

$$M^i(0) = 0^i \text{ (zero tensors).} \tag{24}$$

Q.E.D.

For the proof of the second theorem some additional lemmas are needed. In the proof we show that the second term in (15) is finite dimensionally computable. Multi-dimensional and generalized Volterra terms needed are studied in the appendix.

LEMMA 3.2. *The partial derivative $x_t(\sigma|t)$ is given by*

$$x_t(\sigma|t) = -K(\sigma)\Psi(\sigma, t)G_1(t, y(t))^T R g(t, y(t), r(t)) \tag{25}$$

where the $n \times n$ matrix valued Ψ satisfies

$$\Psi_\sigma(\sigma, t) = -[A^T - W(\sigma)K(\sigma)]\Psi(\sigma, t), \tag{26}$$

$$\Psi(t, t) = I, \tag{27}$$

where W is defined by

$$W(\sigma) = G_1(\sigma, y(\sigma))^T R G_1(\sigma, y(\sigma)), \tag{28}$$

and K is the solution of the matrix Riccati differential equation

$$\dot{K} = AK + KA^T - KWK + BS^{-1}B^T, \tag{29}$$

$$K(0) = J^{-1}, \tag{30}$$

obtained as a result of Lemma 3.1.

Proof. From (20) we obtain

$$x_t(\sigma|t) = -K(\sigma)p_t(\sigma|t). \tag{31}$$

Differentiate then the differential equation of the co-state p of the linear TPBVP, which corresponds to the state-linear subproblem, with respect to the final time t; change the order of differentiation, i.e. set $p_{t\sigma} = p_{\sigma t}$; and use (31) to obtain

$$\frac{\partial}{\partial \sigma} p_t(\sigma|t) = -[A^T - W(\sigma)K(\sigma)]p_t(\sigma|t), \tag{32}$$

which is interpreted as an ordinary differential equation for $p_t(\sigma|t)$. The final condition is obtained via differentiation from the same TPBVP:

$$p_t(t|t) = G_1(t, y(t))^T R g(t, y(t), x(t|t)). \tag{33}$$

Consequently, (32) and (33) yield

$$p_t(\sigma|t) = \Psi(\sigma, t)G_1(t, y(t))^T R g(t, y(t), x(t|t)) \tag{34}$$

resulting in (with (31)) (25). Q.E.D.

LEMMA 3.3. *The partial derivative $z_t(\sigma|t)$ is given by*

$$z_t(\sigma|t) = \Lambda(\sigma|t)z(\sigma|t), \tag{35}$$

where[3]

$$\Lambda(\sigma|t) = \Phi(\sigma|t, 0)\int_0^\sigma \Phi(\alpha|t, 0)^{-1}[Q_x(x(\alpha|t))x_t(\alpha|t)]\Phi(\alpha|t, 0)\, d\alpha\ \Phi(\sigma|t, 0)^{-1}, \tag{36}$$

where Φ is the state transition matrix of the system

$$z_\sigma(\sigma|t) = Q(x(\sigma|t))z(\sigma|t). \tag{37}$$

Proof. Differentiate the equations of the state transition matrix Φ

$$\Phi_\sigma(\sigma|t, 0) = Q(x(\sigma|t))\Phi(\sigma|t, 0), \tag{38}$$

$$\Phi(0|t, 0) = I \tag{39}$$

with respect to the final time t, change the order of differentiation, i.e. set $\Phi_{\sigma t} = \Phi_{t\sigma}$ and solve the resulting linear differential equation of $\Phi_t(\cdot|t, 0)$. Application of the obvious relations

$$z_t(\sigma|t) = \Phi_t(\sigma|t, 0)\zeta, \tag{40}$$

$$\zeta = \Phi(\sigma|t, 0)^{-1}z(\sigma|t), \tag{41}$$

and of the transition property of Φ gives the desired equation (36). Q.E.D.

LEMMA 3.4. *Λ can be represented in the form*

$$\Lambda(\sigma|t) = -S(\sigma|t)\Psi(\sigma, t)G_1(t, y(t))^T R g(t, y(t), x(t|t)) \tag{42}$$

where S has a finite generalized Volterra series (see Appendix) with values on $T_1(\mathbf{R}^{n_1 \times n_1}, \mathbf{R}^n)$.

[3] The derivative Q_x has to be interpreted as an element of $T_1(\mathbf{R}^{n_1 \times n_1}, \mathbf{R}^n)$ (see the appendix). It is defined componentwise by $(Q_x)_{ijk} = \partial Q_{ij}/\partial x_k$, $i, j = 1, \cdots, n_1$, $k = 1, \cdots, n$.

Proof. Due to the nilpotency of the Lie algebra generated by $\{F_k\}$ in the representation of Q (19) there exists an invertible n_1 by n_1 matrix P with the following property [4]. The similarity transformation $\Phi \mapsto \Gamma = P\Phi P^{-1}$ decomposes Γ into l_i-dimensional subsystems satisfying

$$\Gamma^i_\sigma(\sigma|t, 0) = Q^i(x(\sigma|t))\Gamma^i(\sigma|t, 0), \tag{43}$$

$$\Gamma^i(0|t, 0) = I, \tag{44}$$

where Q^i are polynomials in x of the degree ν at most. Furthermore, in the representation of Q^i corresponding to (19) the coefficient matrices F^i_k satisfy

$$F^i_k = \alpha^i_k I + B^i_k, \tag{45}$$

where B^i_k are strictly upper triangular, i.e., have zero elements on and below the main diagonal. For notational simplicity we assume that the polynomial Q is directly of the form (45). It is easily seen [4] that

$$\Phi(\sigma|t, 0) = \left[\exp\left(\sum_{k=1}^{N} \alpha_k \int_0^\sigma u_k(\tau|t)\, d\tau\right)\right] V(\sigma|t, 0) \tag{46}$$

where V satisfies

$$V_\sigma(\sigma|t, 0) = \left[\sum_{k=1}^{N} u_k(\sigma|t)B_k\right] V(\sigma|t, 0), \tag{47}$$

$$V(0|t, 0) = I. \tag{48}$$

Due to the strict upper triangularity of B_k V can be represented by a finite Volterra series [4], [6] with the input N-vector $u(\cdot|t)$.

In the definition (36) of Λ, Φ can be replaced by V due to (46). In the resulting expression $V(\sigma|t, 0)^{-1}$ can be represented by a finite Volterra series due to the finiteness of the series of V itself and due to the fact that V is upper triangular with units in the main diagonal. The second property is easily deduced from the strict upper triangularity of the matrices B_k.

Consider now the representations of V and V^{-1} as generalized finite Volterra series with the input $x(\cdot|t)$. Then the product in (36) is also of the same type. This is seen by suitably using representations of the Volterra series with triangular or symmetric kernels. Consequently, $\Lambda(\sigma|t)$ admits a generalized finite Volterra series.

The extraction of the innovation term $G_1(t, y)^T R g(t, y, x)$ from (36) is obtained by defining first two multiplications

$$(A\mathcal{B})_{ijk} = \sum_{l=1}^{n_1} A_{il}\mathcal{B}_{ljk},$$

$$(\mathcal{B} \cdot A)_{ijk} = \sum_{l=1}^{n_1} \mathcal{B}_{ilk}\mathcal{A}_{lj}, \qquad A \in \mathbf{R}^{n_1 \times n_1}, \quad \mathcal{B} \in \mathbf{R}^{n_1 \times n_1 \times n},$$

the latter of which gives the obvious identity

$$(\mathcal{B}\nu)A = (\mathcal{B} \cdot A)\nu, \qquad \nu \in \mathbf{R}^n.$$

On the basis of the identity and the expression (25) of $x_t(\sigma|t)$ we obtain (42) where

$$S(\sigma|t) = V\left(\int_0^\sigma V^{-1}(Q_x K \Psi) \cdot V\, d\alpha\right) \cdot V^{-1} \tag{49}$$

as an element of $T_1(\mathbf{R}^{n_1\times n_1}, \mathbf{R}^n)$ (identified with $\mathbf{R}^{n_1\times n_1\times n}$) has a finite generalized Volterra series. Q.E.D.

Now it is simple to deduce Theorem 3.2.

Proof of Theorem 3.2. The filtering defined by $s(t) = z(t|t)$ satisfies

$$\dot{s}(t) = [Q(r(t)) + \Lambda(t|t)]s(t), \tag{50}$$

$$s(0) = \zeta \tag{51}$$

due to (15) and (35). Λ has a finite generalized Volterra series on the basis of Lemma 3.4. Theorem A.1 in the appendix says that then $\Lambda(t|t)$ is finite dimensionally computable. Consequently, $s(t)$ has the same property. Q.E.D.

4. Examples. Explicit nonlinear filters in the following classes are first derived.

$$\dot{x} = Ax + Bw, \qquad x_{k+1} = Ax_k + Bw_k,$$

$$\dot{z} = Fz + \tfrac{1}{2}\sum_{i=1}^{n} x_i A_i x, \qquad z_{k+1} = Fz_k + \tfrac{1}{2}\sum_{i=1}^{n} x_{i,k} A_i x_k,$$

$$g(t, y, x) = y - Hx, \qquad g(k, y_k, x_k) = y_k - Hx_k,$$

where the dimension of x, z and y are n, n_1 and m, respectively. By introducing a triply indexed matrix $\mathscr{A}$ ($\in \mathbf{R}^{n_1\times n\times n}$) by

$$\mathscr{A}_{ijk} = (A_j)_{ik}, \qquad i = 1, \cdots, n_1, \quad j, k = 1, \cdots, n,$$

which is symmetric with respect to the two last indices, we can write the quadratic term in the feedforward systems in a compact form $\frac{1}{2}(\mathscr{A}x)x$ where the first multiplication has to be interpreted in the sense of (A.8). Componentwise, it is expressed by

$$(\mathscr{A}x)_{ij} = \sum_{k=1}^{n} \mathscr{A}_{ijk} x_k.$$

4.1. Continuous problem. As stated in the proof of Theorem 3.1 we apply a second degree Riccati transformation

$$z(\sigma|t) = s(\sigma) - L(\sigma)p(\sigma|t) - \tfrac{1}{2}[\mathscr{M}(\sigma)p(\sigma|t)]p(\sigma|t).$$

For typographical reasons we denote by L the n_1 by n matrix and by $\mathscr{M}$ a triply indexed n_1 by n by n generalized matrix. By differentiating with respect to σ by using the equations of the feedforward system and of the co-state of the corresponding linear TPBVP we obtain a polynomial equation of the form

$$v + \tilde{L}p + \tfrac{1}{2}(\tilde{M}p)p = 0,$$

which must be valid for all p. By setting v and the coefficients $\tilde{L}$ and $\tilde{M}$ equal to zero we obtain the algorithm (for $\sigma = t$) for $s(t)$

$$\dot{s} = Fs + \tfrac{1}{2}(\mathscr{A}r)r + LH^T R(y - Hr),$$

$$\dot{L} = FL + L(A^T - H^T RHK) + (\mathscr{A}r)K - \mathscr{M}H^T R(y - Hr),$$

$$\dot{\mathscr{M}} = F\mathscr{M} + \mathscr{M}(A^T - H^T RHK) + \mathscr{M} \cdot (A^T - H^T RHK) - (\mathscr{A}K) \cdot K,$$

$$s(0) = \zeta, \qquad L(0) = 0, \qquad \mathscr{M}(0) = 0,$$

where the multiplication has been defined in the proof of Lemma 3.4.

4.2. Discrete problem. It is shown in [7] that the discrete linear TPBVP

$$x(i+1|k) = Ax(i|k) - BS^{-1}B^T p(i+1|k),$$
$$p(i|k) = A^T p(i+1|k) - H^T R[y_i - Hx(i|k)],$$
$$x(0|k) = \xi - J^{-1}p(0|k),$$
$$p(k+1|k) = 0,$$

corresponding to the linear subproblem of the filtering problem 2.2 for $g = y - Hx$ can be solved by applying the discrete Riccati transformation

$$r_i = x(i|k) + K_i A_i^T p(i+1|k)$$

in the TPBVP. The obvious result is the formal discrete Kalman–Bucy filter for the pair (r, K). As is well known the gain K satisfies the recursion

$$K_{k+1} = (I + \bar{K}_{k+1}H^T RH)^{-1}\bar{K}_{k+1},$$
$$\bar{K}_{k+1} = AK_k A^T + BS^{-1}B^T,$$
$$K_0 = J^{-1}.$$

Application of the second degree Riccati transformation

$$z(i|k) = s_i - L_i A^T p(i+1|k) - \tfrac{1}{2}[\mathcal{M}_i A^T p(i+1|k)][A^T p(i+1|k)]$$

in the feedforward system and by using the equations of the TPBVP, we obtain a polynomial equation

$$v_i + \tilde{L}_i p(i+1|k) + \tfrac{1}{2}[\tilde{\mathcal{M}}_i p(i+1|k)]p(i+1|k) = 0$$

for all $i = 1, 2, \cdots, k$. By setting v_i and the coefficients $\tilde{L}_i$ and $\tilde{\mathcal{M}}_i$ equal to zero we obtain for $\sigma = t$ the discrete filter

$$s_{k+1} = Fs_k + \tfrac{1}{2}(\mathcal{A}r_k)r_k + L_{k+1}\nu_{k+1} + \tfrac{1}{2}(\mathcal{M}_{k+1}\nu_{k+1})\nu_{k+1},$$
$$\nu_{k+1} = H^T R(y_{k+1} - HAr_k),$$
$$L_{k+1} = FL_k\tilde{A}^T + (\mathcal{A}r_k)K_k\tilde{A}^T - \mathcal{M}_{k+1}\nu_{k+1},$$
$$\tilde{A} = (I + \bar{K}_{k+1}H^T RH)^{-1}A,$$
$$\mathcal{M}_{k+1} = F(\mathcal{M}_k\tilde{A}^T)\cdot\tilde{A}^T - (\mathcal{A}K_k\tilde{A}^T)\cdot(K_k\tilde{A}^T),$$
$$s_0 = \zeta, \qquad L_0 = 0, \qquad \mathcal{M}_0 = 0.$$

As compared with the continuous filter an essential difference is that the filtering equation itself is driven by a second degree polynomial in the innovation ν_{k+1}.

An immediate conclusion which can be easily proved is the following. If the degree of the polynomial link map is l then the filtering equation is driven by the lth degree polynomial in the innovation. Furthermore, every coefficient tensor M_k^j in the corresponding Riccati transformation

$$z(i|k) = s_i - \sum_{j=1}^{l} \frac{1}{j!} M_i^j (A^T p(i+1|k))^j$$

is obtained from a (tensor) difference equation driven by a polynomial of degree $l-j$ in the innovation. The highest degree coefficient M_k^l is independent of the innovation.

This phenomenon was, however, expected on the basis of some approximate similar filters proposed in [10] for nonlinear difference systems. The stochastic formulation in the case of a two-dimensional system performed by Marcus et al. [5] also resulted in the optimal conditional mean filter driven by a second degree innovation polynomial.

4.3. Comparisons. Marcus et al. [4], [5] proved for the aforementioned system classes the existence of finite dimensional optimal conditional mean filters. As an explicit example they considered the stochastic counterpart of the following system.

$$\dot{x}_1 = -\alpha x_1 + w_1, \qquad \dot{x}_2 = -\beta x_2 + w_2, \qquad \dot{z} = -\gamma z + x_1 x_2, \qquad y = Hx + v,$$

where H is the identity matrix, $J = 0$, and R and S are diagonal (in (8)). By applying the general equations of § 4.1 we obtain the deterministic least-squares filter

$$\begin{aligned}
\dot{s} &= -\gamma s + r_1 r_2 + M_1 R(y_1 - r_1) + M_2 R(y_2 - r_2),\\
\dot{M}_1 &= -(\alpha + \gamma + RK_{11})M_1 + K_{11} r_2 - M_3 R(y_2 - r_2),\\
\dot{M}_2 &= -(\beta + \gamma + RK_{22})M_2 + K_{22} r_1 - M_3 R(y_1 - r_1),\\
\dot{M}_3 &= -(\alpha + \beta + \gamma + RK_{11} + RK_{22})M_3 - K_{11}K_{22}
\end{aligned}$$

where K_{11} and K_{22} are the diagonal elements of the Riccati matrix of the linear subproblem. The nonlinear filter was obtained by using the Riccati transformation

$$z = s - M_1 p_1 - M_2 p_2 - M_3 p_1 p_2.$$

A detailed inspection shows that stochastic and deterministic filters for this system are formally equivalent. Although, in our case, we have one differential equation less than in the stochastic case. It is seen in the next example that the formal equivalence breaks down due to the dependence of the components in the product affecting the feedforward system.

The example

$$\dot{x} = -\alpha x + w, \qquad \dot{z} = -\gamma z + x^2, \qquad y = x + v,$$

gives for s the filtering equation

$$\dot{s} = -\gamma s + r^2 + M_1 R(y - r)$$

where M_1 is obtained from a two-dimensional system. The stochastic counterpart is now [5]

$$d\hat{z} = (-\gamma \hat{z} + \hat{x}^2 + K)\, dt + M_1 R(d\tilde{y} - \hat{x}\, dt)$$

where $\hat{}$ stands for the conditional expectation, K is the covariance of $\hat{x}$ and $d\tilde{y}$ is defined by $d\tilde{y} = x dt + dv$ (v is a Wiener process with the incremental covariance R^{-1}). The obvious difference as compared with the first example is that instead of $\hat{x}^2$ in the stochastic filter we have here $\hat{x}^2 + K$, the conditional expectation of x^2. The stochastic filter of the first example can be considered to include the similar term $E\{x_1 x_2\}$ which is, however, equal to the product of expectations of x_1 and x_2 resulting in the formal equivalence of the stochastic and deterministic filter.

4.4. Nilpotent class. Optimal filtering for a system of the class of Theorem 3.2 is then studied. Let the linear subsystem be scalar ($\dot{x} = -Ax + w$, $g = y - x$) and the

feedforward system given by

$$\dot{z} = (F_0 + u_1 F_1 + u_2 F_2) z,$$

$$F_0 = -\begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}, \qquad F_1 = \begin{bmatrix} 1 & -1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}, \qquad F_2 = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$u_1 = x, \qquad u_2 = \tfrac{1}{2}x^2.$$

The transformation $\eta = \exp(-F_0\sigma)z$ rejects the constant matrix F_0. Furthermore, we arrive at the differential equation (see Marcus et al. [4])

$$\dot{\eta} = \left(\sum_{i=1}^{4} H_i w_i\right)\eta,$$

where $w = D(\sigma)u$, D is defined by

$$D(\sigma) = \begin{bmatrix} -\alpha(\sigma) & 2\alpha(\sigma) \\ -\gamma(\sigma) & \gamma(\sigma) \\ \beta(\sigma) & 0 \\ 1 & 1 \end{bmatrix}, \quad \alpha = \exp(a-b)\sigma, \quad \beta = \exp(a-c)\sigma, \quad \gamma = \exp(b-c)\sigma,$$

and $\{H_i\}$ span the Lie algebra generated by $\{F_1, F_2\}$. They are given by

$$H_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad H_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad H_3 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad H_4 = I$$

and satisfy $[H_i, H_j] = 0$, except $[H_1, H_2] = H_3$.

Next we consider the finite generalized Volterra series of the matrix S given in Lemma 3.4. The state transition matrix Φ of the feedforward system satisfies

$$\Phi(\sigma|t, 0) = \exp\left(\int_0^\sigma w_4(\tau|t)\, d\tau\right) V(\sigma|t, 0)$$

where

$$V_\sigma(\sigma|t, 0) = \left(\sum_{i=1}^{3} H_i w_i(\sigma|t)\right) V(\sigma|t, 0).$$

Via a successive approximation technique it is seen that V is given by

$$V(\sigma|t, 0) = I + \int_0^\sigma \sum_{i=1}^{3} H_i w_i(\tau|t)\, d\tau + \int_0^\sigma \int_0^{\tau_1} H_3 w_1(\tau_1|t) w_2(\tau_2|t)\, d\tau_2\, d\tau_1,$$

and that it is of the form

$$V = \begin{bmatrix} 1 & v_1 & v_3 \\ 0 & 1 & v_2 \\ 0 & 0 & 1 \end{bmatrix}.$$

On the basis of (49) it is seen that S is also upper triangular with equal but nonconstant diagonal terms whose elements all have a finite generalized Volterra series. Going back into the estimate $z(t|t) = \exp(F_0 t)\eta(t|t)$ we obtain for the filtering $s(t)$ the

differential equation

$$\dot{s} = (F_0 + rF_1 + \tfrac{1}{2}r^2F_2)s + LsR(y - r),$$

where L is defined by $L = \exp(F_0 t)S \exp(-F_0 t)$ and is also upper triangular. The detailed equations for the four different elements of L are obtained via differentiation of the Volterra series of the elements of S and using the above-mentioned transformation formula.

5. Concluding remarks. It is shown that the least-squares formulation of the filtering in a deterministic framework resulted in the same results as the stochastic conditional mean filtering: finite dimensionality of the optimal recursive filters for some system classes. Explicit filter structures were developed for a class of continuous and discrete systems. The Riccati transformation technique employed proved to be a useful tool in the TPBVP's. A side result is presented in the appendix: existence of finite dimensional filters for systems driven by the state of a linear system and described by a finite generalized Volterra series. It can also be proved using the realization results of Crouch [11] and the proof of Theorem 3.1.

Appendix A. Multilinear mappings and generalized Volterra terms. The following definitions and notation are in part from Greub [12], Krasnoselskii et al. [13], and Brockett [6]. U and V are finite dimensional vector spaces over $\mathbf{R}$.

DEFINITION A.1. ϕ is a multilinear mapping of order k (k-linear) from U into V if it is linear in each of the k arguments $x_1, \cdots, x_k \in U$.

DEFINITION A.2. *ϕ is symmetric if*

$$\phi(\pi(x_1), \pi(x_2), \cdots, \pi(x_k)) = \phi(x_1, x_2, \cdots, x_k) \tag{A.1}$$

where π is any permutation on $\{x_1, x_2, \cdots, x_k\}$.

DEFINITION A.3. M is a homogeneous operator of order k from U into V whose value is denoted by Mx^k if

$$Mx^k = \phi(x, x, \cdots, x). \tag{A.2}$$

DEFINITION A.4. A covariant tensor M^k of order k (k-tensor) is a k-linear mapping from U into V. The linear subspace of k-tensors is denoted by $T_k(U, V)$. The subspace of symmetric (in the sense of (A.1)) k-tensors is denoted by $Y_k(U, V)$.

DEFINITION A.5. A mapping Q with values in V is a polynomial of order ν in $x \in U$ if

$$Q(x) = M^0 + \sum_{k=1}^{\nu} M^k x^k \tag{A.3}$$

where $M^k \in T_k(U, V)$, and $M^0 \in V$.

Consider the multidimensional Volterra term which depends on the time parameter t

$$N_i^k(\sigma|t) = \int_0^\sigma \cdots \int_0^\sigma W_i^k(\sigma, \tau_1, \tau_2, \cdots, \tau_i) \cdot (u_1(\tau_1|t), u_2(\tau_2|t), \cdots, u_i(\tau_i|t))\, d\tau_1\, d\tau_2 \cdots d\tau_i \tag{A.4}$$

where $u_j(\sigma|t) \in \mathbf{R}^N$ and $N_i^k(\sigma|t) \in T_k(\mathbf{R}^n, \mathbf{R}^{n_1})$. The kernel is assumed symmetric in the time arguments, i.e. for any permutation π on $\{\tau_1, \cdots, \tau_i\}$

$$W_i^k(\sigma, \pi(\tau_1), \cdots, \pi(\tau_i)) = W_i^k(\sigma, \tau_1, \cdots, \tau_i). \tag{A.5}$$

The value of the kernel is an element of $Y_i(\mathbf{R}^N, T_k(\mathbf{R}^n, \mathbf{R}^{n_1}))$ for each $\sigma, \tau_1, \cdots, \tau_i$. The value of the kernel as a k-tensor valued i-tensor can be interpreted as a generalized matrix with $k+i+1$ running indices, i.e. it is formed by the scalars (see also Crouch [11])

$$W_i^k(\sigma, \cdots)_{j,l_1,\cdots,l_k,m_1,\cdots,m_i}, \qquad j=1,2,\cdots,n_1, \quad l_\nu=1,2,\cdots,n, \quad m_\nu=1,2,\cdots,N.$$

The kernel is said to be *separable* if every scalar component of $W_i^k(\sigma, \cdots)$ can be written in the form

$$W_i^k(\sigma, \tau_1, \cdots, \tau_i)_{\text{comp}} = \sum_{l=1}^{\mu} \gamma_0^l(\sigma)\gamma_1^l(\tau_1)\cdots\gamma_i^l(\tau_i).$$

The components are assumed to be piecewise continuous and locally bounded [6].

Volterra terms can also be defined by using *triangular* kernels by

$$N_i^k(\sigma|t) = \int_0^{\sigma}\int_0^{\tau_1}\cdots\int_0^{\tau_{i-1}} \tilde{W}_i^k(\sigma, \tau_1, \cdots, \tau_i)(\cdots)\, d\tau_1\cdots d\tau_i \tag{A.6}$$

where $\tilde{W}_i^k(\sigma, \tau_1, \cdots, \tau_i) = 0$ if $\tau_{k+j} > \tau_j$, $k, j = 1, 2, \cdots$. Brockett [6] has stated that there exists a one-to-one correspondence between triangular and symmetric kernels. So, depending on the situation to be considered we use triangular or symmetric kernels.

The differentiation of time varying tensors with respect to the scalar variables t or σ is interpreted componentwise.

Define a multiplication of k-tensor $N_i^k(\sigma|t)$ and an $n \times n$ matrix C, and an n-vector v, by

$$(N_i^k(\sigma|t)C)(x_1, \cdots, x_k) = N_i^k(\sigma|t)(x_1, \cdots, x_{k-1}, Cx_k), \tag{A.7}$$

$$(N_i^k(\sigma|t)v)(x_1, \cdots, x_{k-1}) = N_i^k(\sigma|t)(x_1, \cdots, x_{k-1}, v). \tag{A.8}$$

LEMMA A.1. *Let $u_j(\cdot|t) = x(\cdot|t)$, $j = 1, \cdots, N$, the optimal trajectory of the problem* (8)–(10), *in the Volterra term* (A.4) *with a symmetric separable kernel W_i^k. Then*

$$\frac{\partial}{\partial t} N_i^k(\sigma|t) = -N_{i-1}^{k+1}(\sigma|t)\Psi(\sigma, t)G_1(t, y(t))^T R g(t, y(t), x(t|t)) \tag{A.9}$$

where N_{i-1}^{k+1} is a $(k+1)$-tensor valued Volterra term of order $i-1$ with a symmetric separable kernel W_{i-1}^{k+1} defined by

$$W_{i-1}^{k+1}(\sigma, \tau_1, \cdots, \tau_{i-1}) = i\int_0^{\sigma} W_i^k(\sigma, \tau_1, \cdots, \tau_i)K(\tau_i)\Psi(\tau_i, \sigma)\, d\tau_i. \tag{A.10}$$

K and Ψ are given in the main text in Lemma 3.2.

Proof. By differentiating the Volterra term (A.4) with respect to t we obtain the sum of i terms. In the jth term of the sum, instead of $x(\tau_j|t)$, there is $x_t(\tau_j|t)$ which is given by Lemma 3.2. Because of the symmetry of the kernel W_i^k and of the tensor $W_i^k(\sigma, \tau_1, \cdots, \tau_i)$ all the i terms are equal. By using multiplication rules (A.7–A.8) we have a $(i-1)$-tensor (in x) in the integrand. On the basis of the transition property $\Psi(\tau_i, t) = \Psi(\tau_i, \sigma)\Psi(\sigma, t)$ the term $\Psi(\sigma, t)G_1(t, y(t))^T R g(t, y(t), x(t|t))$ can be extracted from the integrand. The remaining term is consequently a $(k+1)$-tensor. By integrating with respect to τ_i the kernel (A.10) is obtained. The separability of the kernel W_{i-1}^{k+1} is due to the separability of W_i^k and to the transition property of Ψ. The symmetry of W_{i-1}^{k+1} is obvious due to the symmetry of W_i^k. Q.E.D.

DEFINITION A.6. Let $u_j(\tau|t) = P^{\nu_j}(\tau)x(\tau|t)^{\nu_j}$, $j = 1, 2, \cdots, N$, *where* $P^{\nu_j}(\tau) \in Y_{\nu_j}(\mathbf{R}^n, \mathbf{R}^N)$, and $x(\cdot|t)$ is as in Lemma A.1. Then the i-fold integral (A.4) is called a generalized Volterra term of order $l = \sum_{j=1}^{N} \nu_j$.

LEMMA A.2. *Let* (A.4) *be a generalized Volterra term as defined above. Then* $W_i^k(\sigma, \cdots)$ *is a tensor of order* l *in* x. *Relabel it by* $V_l^k(\sigma, \cdots)$, *and the corresponding* N_i^k *by* M_l^k. *It satisfies*

$$\frac{\partial}{\partial t} M_l^k(\sigma|t) = -M_{l-1}^{k+1}(\sigma|t)\Psi(\sigma, t)G_1(t, y(t))^T R g(t, y(t), x(t|t)) \tag{A.11}$$

where $M_{l-1}^{k+1}(\sigma|t)$ *is a* $(k+1)$*-tensor valued generalized Volterra term of order* $l-1$ *with a separable kernel, say* V_{l-1}^{k+1}, *and* Ψ *is given in Lemma* 3.2.

The proof is analogous to that of Lemma A.1, so it is omitted. The only exception is that the kernel V_{l-1}^{k+1} cannot be represented in the explicit integrated form as was the case for W_{i-1}^{k+1}.

Remark A.1. It is obvious that the generalized Volterra term of order l with i-fold integration can be expressed as a usual Volterra term of order l with l-fold integration. But then the kernel would include an $(l-i)$-fold product of δ-functions (see Brockett [6]).

On the basis of Lemmas A.1 and A.2 and by using the results of Brockett [6] and Crouch [11] the following theorem is obtained.

THEOREM A.1. *Let an input–output map be given in the form of a finite generalized Volterra series with symmetric separable kernels*

$$w(\sigma|t) = W_0(\sigma) + \sum_{i=1}^{N_w} \int_0^\sigma \cdots \int_0^\sigma W_i^0(\sigma, \tau_1, \cdots, \tau_i) \cdot (u(\tau_1|t), \cdots, u(\tau_i|t))\, d\tau_1 \cdots d\tau_i, \tag{A.12}$$

where $w(\sigma|t) \in \mathbf{R}^{n_1}$, *and* $u(\tau_i|t) \in \mathbf{R}^N$ *are polynomials in* $x(\tau_i|t)$ *given in Definition* A.6. *Then* $\mu(t) \triangleq w(t|t)$ *is finite dimensionally computable, i.e. it is obtained from a finite dimensional differential system with a fixed initial value driven by* $y(t)$ *and polynomials of* $x(t|t)$.

Proof. In the formula $\dot{\mu}(t) = w_\sigma(t|t) + w_t(t|t)$ we prove that the both partials are finite dimensionally computable. By keeping t as a fixed parameter the series considered as a normal Volterra series with input $u(\cdot|t)$ is realizable by a finite dimensional differential system. The result has been proved by Brockett [6], and later on by Crouch [11] who also presented a detailed structure of the corresponding differential system. By setting in it $\sigma = t$ we obtain a system for $w_\sigma(t|t)$. Successive application of Lemma A.2 in each of the individual terms of (A.12) gives as the final step the equation

$$\frac{\partial}{\partial t} M_1^{k+l_i-1}(\sigma|t) = -M_0^{k+l_i}(\sigma|t)\Psi(\sigma, t)G_1(t, y(t))^T R g(t, y(t), r(t)). \tag{A.13}$$

Furthermore, we obtain a finite set of tensors $M_j^{k+1_i-j}(\sigma|t)$ which all are finite dimensionally computable at $\sigma = t$ with obvious zero initial values. This is seen by realizing every tensor for the fixed time t by a finite differential system and by using Lemma A.2 for $\sigma = t$.

At last, it is seen that $M_0^{k+l_i}(\sigma|t)$ is a zeroth order Volterra term, independent of the final time t, and it satisfies

$$\begin{aligned} &M_0^{k+l_i}(t|t) = W_1^{k+l_i-1}(t, t)K(t) + M_0^{k+l_i}(t|t)[A^T - W(t)K(t)], \\ &M_0^{k+l_i}(0|0) = 0. \end{aligned} \tag{A.15}$$

Consequently, every individual Volterra term in (A.12) for $\sigma = t$ is obtained from a finite set of differential equations driven by $y(t)$ and polynomials of $r(t)$ (i.e. of $x(t|t)$). Q.E.D.

Remark A.2. It has to be noted that if the Volterra series were infinite we could not consider it in general as an input–output map between x and w, but only between u and w [14].

Appendix B. On Lie algebras [15].

DEFINITION B.1. A vector space L where a bilinear product $L \times L \to L$ is defined satisfying the anticommutativity and the Jacobi conditions

$$[A, B] = -[B, A], \qquad [[A, B], C] + [[B, C], A] + [[C, A], B] = 0,$$

is a Lie algebra.

DEFINITION B.2. L is nilpotent if the lower central series of ideals

$$L^0 = L, \qquad L^{i+1} = [L, L^i] = \{[A, B], B \in L^i\},$$

is $\{0\}$ for some i.

It is seen that the n_1 by n_1 matrices considered in the main text form a Lie algebra if the product is defined by using a standard matrix product by

$$[A, B] = AB - BA.$$

REFERENCES

[1] D. DETCHMENDY AND R. SRIDHAR, *Sequential estimation of states and parameters in noisy nonlinear dynamical systems*, J. Basic. Eng., 88 (1966), pp. 362–368.

[2] R. BELLMAN, H. KAGIWADA, R. KALABA AND R. SRIDHAR, *Invariant imbedding and nonlinear filtering theory*, J. Astronaut. Sci., 13 (1966), pp. 110–115.

[3] R. E. MORTENSEN, *Maximum likelihood recursive filtering*, J. Optim. Theory Appl., 2 (1968), pp. 386–394.

[4] S. I. MARCUS AND A. S. WILLSKY, *Algebraic structure and finite dimensional estimation*, SIAM J. Math. Anal., 9 (1978), pp. 312–327.

[5] S. I. MARCUS, S. K. MITTER AND D. OCONE, *Finite dimensional nonlinear estimation for a class of systems in continuous and discrete time*, Proc. International Conference on Analysis and Optimization of Stochastic systems, Oxford, England, September 1978.

[6] R. W. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167–176.

[7] M. T. NIHTILÄ, *Optimal state-linear filtering through implicit output equation*, Preprints 8th World Congress of IFAC, Kyoto, Japan, Aug. 1981, vol. V, pp. 79–83.

[8] ———, *Optimal finite dimensional solution for a class of nonlinear observation problems*, J. Optim. Theory Appl., 38 (1982), pp. 231–240.

[9] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[10] M. T. NIHTILÄ, *Non-statistical filtering in nonlinear time-discrete systems*, Third IMA Conference on Control Theory, J. E. Marshall et al., eds., Academic Press, London, 1981.

[11] P. E. CROUCH, *Dynamical realizations of finite Volterra series*, this Journal, 19 (1981), pp. 177–202.

[12] W. H. GREUB, *Linear Algebra*, Springer-Verlag, New York, 1963.

[13] M. A. KRASNOSELSKII, G. M. VAINIKKO, P. P. ZABREIKO, YA. B. RUTITSKII AND U. YA. STETSENKO, *Approximate Solution of Operator Equations*, Wolters-Noordhoff, Groningen, 1972.

[14] C. LESIAK AND A. J. KRENER, *The existence and uniqueness of Volterra series for nonlinear systems*, IEEE Trans. Automatic Control, AC-23 (1978), pp. 1090–1095.

[15] M. HAUSNER AND J. T. SCHWARTZ, *Lie Groups and Lie Algebras*, Gordon and Breach, New York, 1968.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0007 $01.25/0

NONLINEAR SYSTEM IDENTIFICATION BASED ON A FOCK SPACE FRAMEWORK*

L. V. ZYLA† AND RUI J. P. DEFIGUEIREDO‡

Abstract. A method is presented for the identification of a nonlinear dynamical system whose input-output map is assumed to be a Volterra series mapping inputs u belonging to $L^2(I)$ to outputs y in the Sobolev space $H_n^2(I)$, where I is a compact interval of the real line and n is a nonnegative integer. The input-output relation is denoted by $y(t) = V(u)(t) \triangleq V_t(u)$, $t \in I$. It is assumed also that a set of test input-output pairs $\{u_j \in L^2(I), y_j \in H_n^2(I): j = 1, \cdots, m\}$ are provided.

Our system identification procedure is based on the construction of a reproducing kernel Hilbert space, namely a symmetric Fock space F_ρ, for the nonlinear functional V_t. The corresponding nonlinear operator V then belongs to a Bochner space B_n^2. We obtain the best estimate $\hat{V}$ of V, based on the input-output data, in the form of a generalized inverse in B_n^2. Both the noncausal and causal versions of $\hat{V}$ are derived. The concept of ε-causality, which is weaker than that of causality, is also introduced and motivated, and an ε-causal solution to the system identification problem is derived. Finally, the modifications needed to be introduced in the above solutions when the data is noisy are indicated.

Key words. nonlinear systems, system identification, generalized inverses, approximation, Fock spaces, Volterra series, operator theory

1. Introduction.

We consider the problem of identifying a nonlinear dynamical system whose input-output map V is described by a Volterra series

$$y(t) = V(u)(t) = V_t(y) = \sum_{k=0}^{\infty} \frac{1}{k!} \int_I \cdots \int_I h_k(t; t_1, \cdots, t_k) u(t_1) \cdots u(t_k)\, dt_1 \cdots dt_k, \tag{1}$$

where the input u belongs to real $L^2(I)$, I being an interval of the real line, and, for some nonnegative integer n, the output y is a member of the Sobolev space $H_n^2(I)$ of real-valued functions y on I such that $y^{(i)} = d^i y/dt^i$ is absolutely continuous on I, $i = 0, 1, \cdots, n-1$, and $y^{(n)} \in L^2(I)$. To guarantee the smoothness property of the output just stated, it will be sufficient that each kernel $h_k : R^{k+1} \to R^1$ satisfy the following condition:

(a)

$$h_k^{(i)} \triangleq \partial^i h_k / \partial t^i \text{ when viewed as a function from } I \text{ to } L^2(I^k) \text{ is absolutely continuous}^1 \text{ over } t \in I \text{ for } i = 0, \cdots, n-1, \text{ and } h_k^{(n)} \in L^2(I^{k+1}). \tag{2}$$

The following two additional restrictions on the kernels are assumed so that, as shown in § 2, the nonlinear Volterra functional V_t can be made to belong to a reproducing kernel Hilbert space F_ρ, called a symmetric Fock space (or, simply, Fock space), and the corresponding Volterra operator V to a Bochner space B_n^2, where ρ

* Received by the editors June 18, 1979, and in final revised form December 27, 1982. This research was supported in part by the National Science Foundation under grant ENG 74-17955, by a contract from the Rome Air Development Center, Air Force Systems Command, Griffiss Air Force Base, New York, and by the Office of Naval Research under contract N00014-79-C-0442 (Mathematical Statistics and Probability Program).

† McDonnell Douglas Technical Services Company, Houston, Texas 77058. Formerly with the Department of Mathematical Sciences, Rice University, Houston, Texas 77251.

‡ Department of Mathematical Sciences and Department of Electrical Engineering, Rice University, Houston, Texas 77251.

[1] Absolute continuity of $h_k^{(i)} : I = [a, b] \to L^2(I^k)$ is defined by the statement: For every $\varepsilon > 0$, there is a $\delta > 0$, such that if $a \leqq z_1 < y_1 \leqq z_2 < y_2 \cdots \leqq z_N < y_N \leqq b$, then $\sum_{j=1}^{N} \| h_k^{(i)}(y_j; \cdots) - h_k^{(i)}(z_j; \cdots) \|_{L^2(I^k)} < \varepsilon$ whenever $\sum_{j=1}^{N} |y_j - z_j| < \delta$.

is a positive constant:

(b)

(3) $h_k(t; t_1, \cdots, t_k)$ is symmetric with respect to the argument variables $t_1, \cdots, t_k$;

(c)

$$\sum_{k=0}^{\infty} \frac{\rho^k}{k!} \|h_k^{(i)}(\cdot, \cdots)\|^2_{L^2(I^{k+1})} < \infty, \qquad i = 0, \cdots, n. \tag{4}$$

Finally, we assume that:

(d) Specific information on the system under consideration is provided in the form of a set of test input-output pairs $\{u_j \in L^2(I),\ y_j \in H_n^2(I): j = 1, \cdots, m\}$.

Under the above conditions, we obtain the best estimate $\hat{V}_t$ of V_t as a generalized inverse in the space F_ρ. The knowledge of $\hat{V}_t$ for every $t \in I$ then determines the best estimate $\hat{V}$ of the operator V.

In §§ 3 and 4, we present respectively the noncausal and causal versions of the above solution (generalized inverse). In particular, we discuss in § 4 some of the difficulties which may arise in obtaining the expression for the causal solution in a neighborhood of $t = 0$, and introduce the weaker concept of "ε-causality," which is easy to satisfy.

Finally, in § 5, we indicate how the above results are to be modified when the data is contaminated by noise.

The present approach to the system identification problem is similar to the one proposed by deFigueiredo and Caprihan [1], [2] for the identification of linear systems, with the basic difference that in the linear case, the space, to which the operator to be identified belonged, was assumed to be the space of "trace class" operators.

We would also like to point out that the idea of using the generalized inverse approach to the solution of the system identification problem has been extensively discussed in the literature and notably by Balakrishnan [3], Hsieh [4], Root [5], and more recently by Beutler and Root [6], Franklin [7], Mosca [8] and Porter [9]–[11]. The new feature of the present results in the use of the infinite-dimensional Fock space framework, and the reproducing kernel pertaining to this space, to derive a *parameter-free* solution to the nonlinear system identification problem under consideration. The solution obtained is an approximate infinite-dimensional Volterra series which constitutes the best approximation to the entire (untruncated) Volterra series representing the system to be identified, subject to the input-output data constraints.

2. Fock spaces and Bochner spaces. Let ρ and I be defined as before and assume $t \in I$ to be a fixed parameter. By a (symmetric) Fock space F_ρ we mean the set of Volterra functional series maps $V_t: L^2(I) \to R^1$ described by (1) and satisfying the conditions stated in the preceding section, each member of F_ρ being uniquely characterized by its sequence of kernels, and the inner product between any two members V_t, with associated kernels $\{h_0(t), h_1(t; t_1), \cdots, h_k(t; t_1, \cdots, t_k), \cdots\}$, and W_t, with associated kernels $\{\tilde{h}_0(t), \tilde{h}_1(t, t_1), \cdots, \tilde{h}_k(t; t_1, \cdots, t_k), \cdots\}$, being defined by

$$\langle V_t, W_t \rangle_{F_\rho} = \sum_{k=0}^{\infty} \frac{\rho^k}{k!} \langle h_k(t; \cdots), \tilde{h}_k(t; \cdots) \rangle_{L^2(I^k)}. \tag{5}$$

Under the conditions stated, F_ρ can be shown [12]–[15] to be a Hilbert space with the inner product (5).

Let us introduce the functional $K: L^2(I)\times L^2(I)\to R^1$ expressed by

$$K(u,v)=\exp\left\{\frac{1}{\rho}\langle u,v\rangle_{L^2(I)}\right\}. \tag{6}$$

Then

$$K(u,\cdot)=\exp\left\{\frac{1}{\rho}\langle u,\cdot\rangle_{L^2(I)}\right\}=\sum_{k=0}^{\infty}\frac{1}{k!\rho^k}u\otimes u\otimes\cdots\otimes u \tag{7}$$

(where the symbol $\otimes$ denotes tensor product between elements of $L^2(I)$) is clearly a member of F_ρ. Furthermore, it follows from (5), (7) and (1) that

$$\langle V_t, K(u,\cdot)\rangle_{F_\rho}=\left\langle V_t,\exp\left\{\frac{1}{\rho}\langle u,\cdot\rangle_{L^2(I)}\right\}\right\rangle_{F_\rho}=V_t(u). \tag{8}$$

Thus (7) and (8) imply:

PROPOSITION 1 [15]. *Under the conditions stated, F_ρ is a reproducing kernel Hilbert space* (RKHS) *with the reproducing $K(u,v)$ defined by* (6).

We next construct the Hilbert space B_n^2 for the operator V. Formally, we define B_n^2 as the space of operators W from I to F_ρ (the value of W at t being denoted by W_t) such that the strong derivatives (in the F_ρ norm) $W_t^{(i)}$ of W at t, for $i=0, 1,\cdots, n-1$, are absolutely continuous on I, and $W_t^{(n)}$ satisfies

$$\int_I \|W_t^{(n)}\|_{F_\rho}^2\,dt<\infty. \tag{9}$$

It is clear that the conditions (2) through (4) on the kernels of the Volterra series (1) guarantee that the Volterra operator V, introduced earlier, belongs to B_n^2.

B_n^2 can be made a Hilbert space [15], [16] under a variety of inner products defined in much the same way as in the Sobolev space $H_n^2(I)$. A typical such inner product for two elements V and W of B_n^2 is

$$\langle V,W\rangle_{B_n^2}=\sum_{i=0}^{n-1}a_i\int_I\langle V_t^{(i)},W_t^{(i)}\rangle_{F_\rho}\,dt, \tag{10}$$

where a_i, $i=0, 1,\cdots, n-1$, are positive constants.

3. Noncausal solution in the noiseless case. Under the conditions stated, the noncausal solution to the noiseless nonlinear system identification problem may now be obtained as the solution to the following minimum norm problem:

$$\inf\|V\|_{B_n^2},\qquad V\in B_n^2, \tag{11a}$$

$$V(u_j)=y_j,\qquad j=1,\cdots,m \tag{11b}$$

(where $\{u_j\in L^2(I),\ y_j\in H_n^2(I): j=1,\cdots,m\}$ are test input-output pairs).

Since, according to (13),

$$\|V\|_{B_n^2}^2=\sum_{i=0}^{n}a_i\int_I\|V_t^{(i)}\|_{F_\rho}^2\,dt, \tag{12}$$

the minimum of (12) is achieved by the solution of the $(n+1)$ minimization problems

$$\min\|V_t^{(i)}\|_{F_\rho}\quad\forall t\in I,\quad V_t^{(i)}\in F_\rho, \tag{13a}$$

$$V_t^{(i)}(u_j)=y_j^{(i)}(t),\qquad i=0,\cdots,n,\quad j=1,\cdots,m. \tag{13b}$$

Now, for every $t \in I$, each u_j may be viewed as a continuous linear functional on F_ρ. In fact if, for all $W_t \in F_\rho$, we write

$$W_t(u_j) = \tilde{u}_j(W_t), \tag{14}$$

then the functional $\tilde{u}_j$ is linear, since $\tilde{u}_j(aW_t + b\tilde{W}_t) = aW_t(u_j) + b\tilde{W}_t(u_j) = a\tilde{u}_j(W_t) + b\tilde{u}_j(\tilde{W}_t)$ for all $a, b \in R^1$ and $W_t, \tilde{W}_t \in F_\rho$, and also $\tilde{u}_j$ is bounded because

$$\|\tilde{u}_j(W_t)\| = \|W_t(u_j)\| \leqq \exp\left\{\frac{1}{2\rho}\|u_j\|_{L^2(I)}\right\}\|W_t\|_{F_\rho}. \tag{15}$$

The following is also true.

LEMMA 1. *If u_j, $j = 1, \cdots, n$, are distinct elements of $L^2(I)$, then the $m \times m$ matrix*

$$G = \left\{\exp\left\{\frac{1}{\rho}\langle u_i, u_j\rangle_{L^2(I)}\right\}\right\}_{i,j=1,\cdots,m} \tag{16}$$

is nonsingular.

Proof. If u_j, $j = 1, \cdots, m$, are distinct, then $\exp\{(1/\rho)\langle u_j, \cdot\rangle_{L^2(I)}\}$, $j = 1, \cdots, m$, are linearly independent elements of F_ρ (see Guichardet [17]). Since, according to (5) and (7),

$$\left\langle \exp\left\{\frac{1}{\rho}\langle u_i, \cdot\rangle_{L^2(I)}\right\}, \exp\left\{\frac{1}{\rho}\langle u_j, \cdot\rangle_{L^2(I)}\right\}\right\rangle_{F_\rho} = \exp\left\{\frac{1}{\rho}\langle u_i, u_j\rangle_{L^2(I)}\right\}, \tag{17}$$

it follows that G defined by (16) is the Gram matrix for the linearly independent elements $\exp\{(1/\rho)\langle u_j, \cdot\rangle_{L^2(I)}\}$, $j = 1, \cdots, m$, of F_ρ, and hence G is nonsingular. □

From now on we will assume that u_j, $j = 1, \cdots, m$, are all distinct. According to (8), we have

$$V_t^{(i)}(u_j) = \left\langle V_t^{(i)}, \exp\left\{\frac{1}{\rho}\langle u_j, \cdot\rangle_{L^2(I)}\right\}\right\rangle_{F_\rho}. \tag{18}$$

Hence, by the projection theorem, the minimum of (13) is achieved by projecting $V_t^{(i)}$ into the span of the representers $\exp\{(1/\rho)\langle u_j, \cdot\rangle_{L^2(I)}\}$, $j = 1, \cdots, m$, of $\tilde{u}_j$, $j = 1, \cdots, m$, in F_ρ. In other words, the optimal estimate $\hat{V}_t^{(i)}$ of $V_t^{(i)}$ is of the form:

$$\hat{V}_t^{(i)} = \sum_{j=1}^{m} c_{ij}(t) \exp\left\{\frac{1}{\rho}\langle u_j, \cdot\rangle_{L^2(I)}\right\}, \qquad i = 0, \cdots, n, \tag{19}$$

where the constants $c_{ij}(t)$ are determined by the constraints (13b). A simple calculation toward satisfaction of all these constraint requirements shows that $c_{ij}(t) = c_j^{(i)}(t)$, where $c_j(t)$ are the coefficients in the following expression for $\hat{V}_t$:

$$\hat{V}_t = \sum_{j=1}^{m} c_j(t) \exp\left\{\frac{1}{\rho}\langle u_j, \cdot\rangle_{L^2(I)}\right\}. \tag{20}$$

The coefficients are obtained from:

$$\mathbf{c}(t) = G^{-1}\mathbf{y}(t), \tag{21}$$

$$\mathbf{c}(t) = \text{col}\,(c_1(t), \cdots, c_m(t)), \tag{22}$$

$$\mathbf{y}(t) = \text{col}\,(y_1(t), \cdots, y_m(t)), \tag{23}$$

$$G = \left\{\exp\left\{\frac{1}{\rho}\langle u_i, u_j\rangle_{L^2(I)}\right\}\right\}_{i,j=1,\cdots,m}. \tag{24}$$

We summarize the above results in the form of the following:

THEOREM 1. *Given that the input-output map V to be identified belongs to B_n^2 and assuming that we are provided with a set of test input-output pairs $\{u_j \in L^2(I), y_j \in H_n^2(I): j = 1, \cdots, m\}$, where $u_j, j = 1, \cdots, m$, are distinct elements of $L^2(I)$, the problem* (13) *has a unique solution expressed by*

$$\hat{V}_t(u) = \sum_{j=1}^{m} c_j \exp\left\{\frac{1}{\rho}\langle u_j, u\rangle_{L^2(I)}\right\}, \tag{25}$$

where the functions $c_j \in H_n^2(I)$ are obtained by (21) *through* (24).

Remark 1. It can be shown [15] that (25) is also the solution of the min-max problem:

$$\min_{W \in \chi} \{\sup_{W \in C} \|V - W\|_{B_n^2}\}, \tag{26a}$$

where

$$\chi = \{V \in B_n^2: V(u_j) = y_j, j = 1, \cdots, m\} \tag{26b}$$

and C is the uncertainty class

$$C = \{W \in B_n^2: \|W\|_{B_n^2}^2 \leqq \gamma^2\}, \tag{26c}$$

where γ is an appropriate positive constant, sufficiently large for C to have nonzero intersection with χ. In a specific application, γ^2 may be equated to the largest eigenvalue of the covariance associated with the probability of V over B_n^2.

Remark 2. It is of interest to obtain an expression for the error

$$\xi = \|\hat{V}_t - V_t\|_{F_\rho}^2. \tag{27}$$

According to the projection theorem, we have

$$\xi = (\|V_t\|_{F_\rho}^2 - \mathbf{y}^T(t) G^{-1} \mathbf{y}(t)). \tag{28}$$

If $u_j, j = 1, \cdots, m$, are orthonormal, then the diagonal elements of G are e (Napierian base) and its off-diagonal elements equal unity. Then (28) can be expressed as

$$\xi = \|V_t\|_{F_\rho}^2 - \alpha \sum_{j=1}^{m} |y_j(t)|^2 - \beta \sum_{i,j=1}^{m} y_i(t) y_j(t), \tag{29}$$

where

$$\alpha = \frac{e + m - 2}{e^2 + (m-2)e - m - 1}, \qquad \beta = \frac{-1}{e^2 + (m-2)e - (m-1)}. \tag{30}$$

An estimate of ξ may be obtained by replacing $\|V_t\|_{F_\rho}^2$ in the above formula by γ^2, where γ is the constant introduced in (26c).

4. Causal solution in the noiseless case. We now present and discuss the causal solution to problem (11) for the case in which $n = 0$, that is, in the space B_0^2. For this purpose it is necessary to add to the minimization problem the additional constraint that $V \in B_0^2$ satisfy the causality condition:

$$P_t V P_t(u) = P_t V(u), \qquad t \in I, \quad u \in L^2(I), \tag{31}$$

where $P_t: L^2(I) \to L^2(I)$ is defined by

$$P_t u(s) = \begin{cases} u(s) & \text{if } s \leqq t, \\ 0 & \text{if } s > t. \end{cases} \tag{32}$$

Let $\tilde{u}_j$ denote the continuous linear functional on F_ρ defined by

$$V(u_j)(t) = \tilde{u}_j(V_t), \qquad u_j \in L^2(I), \tag{33}$$

where now V is causal. Since the representer for $\tilde{u}_j$ in F_ρ is

$$\exp\left\{\frac{1}{\rho}\langle P_t u_j, \cdot\rangle_{L^2(I)}\right\}, \tag{34}$$

it follows from an argument similar to the one in the preceding section that the following expression applies to the causal solution of (11):

$$\hat{V}(u)(t) = \sum_{j=1}^{m} \tilde{c}_j(t) \exp\left\{\frac{1}{\rho}\langle P_t u_j, P_t u\rangle_{L^2(I)}\right\}, \tag{35}$$

where now $\tilde{\mathbf{c}}(t) = \operatorname{col}(\tilde{c}_1(t), \cdots, \tilde{c}_m(t))$ is the solution of

$$\tilde{G}(t)\tilde{\mathbf{c}}(t) = \mathbf{y}(t) \tag{36}$$

where $\tilde{G}(t)$ is the causal Gramm matrix

$$\tilde{G}(t) = \left\{\exp\left(\frac{1}{\rho}\langle P_t u_i, P_t u_j\rangle_{L^2(I)}\right)\right\}_{i,j=1,\cdots,m} \tag{37}$$

and $\mathbf{y}(t)$ is defined by (23).

4.1 Strictly causal solution. For simplicity in presentation and without loss in generality, let $I = [0, 1]$.

We assume:

$$\{P_t u_1, \cdots, P_t u_m\} \text{ are distinct elements of } L^2(I) \text{ for every } t \in (0, 1]. \tag{38}$$

It is clear that, under this condition, $\tilde{G}(t)$ is invertible for every $t \in (0, 1]$, and hence (35) makes sense for all such t. A difficulty arises, however, at $t = 0$, since $\tilde{G}(0)$ is a singular matrix with all its elements equal to unity. To guarantee that (35) be well defined at $t = 0+$, additional conditions on the system to be identified need to be imposed, by our selecting the test inputs appropriately and observing that the corresponding test outputs satisfy suitable conditions.

Specifically, according to (36), $\{(u_j, y_j): j = 1, \cdots, m\}$ must be such that the conditions (39) and (40) below are satisfied:

(i) for $j = 1, \cdots, m$,

$$y_j(0+) = \lim_{\substack{t\to 0\\ t>0}} y_j(t)$$

exists and

$$y_1(0+) = y_2(0+) = \cdots = y_m(0+) \tag{39}$$

(each being equal to $\tilde{c}_1(0+) + \tilde{c}_2(0+) + \cdots + \tilde{c}_m(0+)$);

(ii)

$$\lim_{\substack{t\to 0\\ t>0}} \tilde{G}^{-1}(t)\mathbf{y}(t) = \tilde{G}^{-1}(0+)\mathbf{y}(0+) \tag{40}$$

exists as a finite vector.

Remark 3. It is of interest to derive sufficient conditions on $\{(u_j, y_j): j = 1, \cdots, m\}$ so that (40) is satisfied.

For $m = 1$, (39) implies (40). The situation however is nontrivial for $m > 1$.

For $m = 2$, (40) is assured by requiring that

$$u_1(0+),\ u_2(0+),\ y_1'(0+),\ y_2'(0+) \tag{41}$$

exist as finite real numbers and $u_1(0+) \neq u_2(0+)$ (where, as before, for a given function f on I we denote by $f(0+)$ the limit approached by f at 0 from the right). This is gleaned from the fact that (for $m = 2$)

$$c_1(t) = \frac{\tilde{G}_{22}(t)y_1(t) - \tilde{G}_{12}(t)y_2(t)}{\tilde{G}_{11}(t)\tilde{G}_{22}(t) - \tilde{G}_{12}^2(t)}, \tag{42}$$

$$c_2(t) = \frac{\tilde{G}_{11}(t)y_2(t) - \tilde{G}_{12}(t)y_1(t)}{\tilde{G}_{11}(t)\tilde{G}_{22}(t) - \tilde{G}_{12}^2(t)}. \tag{43}$$

Now, according to (37),

$$\tilde{G}_{ij}(0+) = 1, \qquad i, j = 1, \cdots, m, \tag{44}$$

$$\tilde{G}_{ij}'(0+) = \frac{1}{\rho} u_i(0+)u_j(0+). \tag{45}$$

Hence, by l'Hôpital's rule,

$$\begin{aligned} \tilde{c}_1(0+) &= \lim_{\substack{t \to 0 \\ t>0}} \frac{\tilde{G}_{22}'(t)y_1(t) + \tilde{G}_{22}(t)y_1'(t) - \tilde{G}_{12}'y_2(t) - \tilde{G}_{12}(t)y_2'(t)}{\tilde{G}_{11}'(t)\tilde{G}_{22}(t) + \tilde{G}_{11}(t)\tilde{G}_{22}'(t) - 2\tilde{G}_{12}(t)\tilde{G}_{12}'(t)} \\ &= \frac{(1/\rho)u_2(0+)(u_2(0+) - u_1(0+))y_1(0+) + y_1'(0+) - y_2'(0+)}{(1/\rho)(u_1(0+) - u_2(0+))^2} \end{aligned} \tag{46}$$

and similarly with $\tilde{c}_2(t)$, which shows that (41) is a sufficient condition for (40).

For $m > 2$, using an approach similar to the one above, we have, by Cramer's rule,

$$\tilde{c}_j(t) = \frac{|\tilde{G}(j;t)|}{|\tilde{G}(t)|}, \tag{47}$$

where $|(\cdots)|$ denotes the determinant of $(\cdots)$ and $\tilde{G}(j;t)$ is the matrix obtained from $\tilde{G}(t)$ by replacing the jth column by $\mathbf{y}(t)$.

Letting $\tilde{G}_j(t)$ denote the jth column of $\tilde{G}(t)$, we have by Leibniz's rule

$$\begin{aligned} |\tilde{G}'(t)| &= |\tilde{G}_1'(t), \tilde{G}_2(t), \cdots, \tilde{G}_m(t)| + |\tilde{G}_1(t), \tilde{G}_2'(t), \tilde{G}_3(t), \cdots, \tilde{G}_m(t)| \\ &\quad + \cdots + |\tilde{G}_1(t), \tilde{G}_2(t), \cdots, \tilde{G}_{m-1}(t), \tilde{G}_m'(t)|. \end{aligned} \tag{48}$$

Since

$$\tilde{G}_j(0+) = \operatorname{col}(1, \cdots, 1), \qquad j = 1, \cdots, m, \tag{49}$$

each of the determinants in the right side of (48) vanishes because the respective matrix has $m-1$ columns consisting of ones. By continuing the differentiation of $\tilde{G}(t)$ k times, we deduce that only if $k \geqq m-1$, does $|\tilde{G}^{(k)}(0+)|$ consist of a sum of determinants of matrices each of which need not possess a repeated column of ones. Hence we conclude that:

(50) For $m > 2$, (40) holds by our requiring that (for $k \geqq m-1$ and for $j = 1, \cdots, m$) $u_j^{(i)}(0+)$, $i = 0, 1, \cdots, \kappa - 1$, and $y_j^{(i)}(0+)$, $i = 0, 1, \cdots, k$, exist as finite numbers, and $|\tilde{G}^{(k)}(0+)| \neq 0$.

From the above considerations we reach the following conclusion.

THEOREM 2. *Suppose that the selected test inputs and the corresponding outputs satisfy* (38), (39), *and* (40) (*or the conditions in Remark* 3). *Then the nonlinear system identification problem* (11) (*with* $n=0$) *has a unique causal solution expressed by* (35)–(37).

4.2. ε-causal solution. In order to do away with restrictions (39) and (40) we now introduce the following concept.

DEFINITION. A nonlinear operator $V \in B_n^2$ is ε-causal (for some positive $\varepsilon < 1$) if

$$P_t(V(P_t u)) = P_t(V(u)) \quad \text{for } t \geqq \varepsilon. \tag{51}$$

Remark 4. This concept is motivated by the fact that it is not possible to identify a causal system without first observing its input-output behavior over a time interval of nonzero length.

The following conclusion is clear.

THEOREM 3. *Suppose condition* (38) *holds. Then for an arbitrarily small* $\varepsilon > 0$, *the nonlinear system identification problem* (11) (*with* $n=0$) *has a unique* ε-*causal solution given by*:

$$\hat{V}_t(0) = \begin{cases} \sum_{j=1}^{m} c_j(t) \exp\left\{\frac{1}{\rho}\langle u_j, \cdot\rangle_{L^2(0,\varepsilon)}\right\}, & 0 < t \leqq \varepsilon, \\ \sum_{j=1}^{m} \tilde{c}_j(t) \exp\left\{\frac{1}{\rho}\langle P_t u_j, P_t(\cdot)\rangle_{L^2(I)}\right\}, & \varepsilon \leqq t \leqq 1, \end{cases} \tag{52}$$

where

$$\mathbf{c}(t) = G^{-1}\mathbf{y}(t), \tag{53a}$$

$$\tilde{\mathbf{c}}(t) = \tilde{G}^{-1}(t)\mathbf{y}(t), \tag{53b}$$

where G *is an* $m \times m$ *matrix with elements*

$$G_{ij} = \exp\left\{\frac{1}{\rho}(u_i, u_j)_{L^2(0,\varepsilon)}\right\}, \tag{54}$$

and $\tilde{G}(t)$ *is defined by* (37).

5. Nonlinear system identification in the noisy case. If the ouput measurements are corrupted by noise, we model the input-output relation by

$$y = V(u) + \nu, \tag{55}$$

where ν is the projection of the noise into the output space. Then the test input-output data is of the form

$$y_j = V(u_j) + \nu_j, \tag{56}$$

where we assume that

$$\langle \nu_i, \nu_j\rangle_{H_n^2(I)} = q_i \delta_{ij}, \qquad i, j = 1, \cdots, m, \tag{57}$$

q_i being positive constants and δ_{ij} = Kronecker delta.

The nonlinear system identification problem may then be posed as the unconstrained minimization over all $V \in B_n^2$ of the functional

$$\tilde{J}(V) = \|V\|_{B_n^2}^2 + \sum_{j=1}^{m} q_j^{-1} \|V(u_j) - y_j\|_{H_n^2(I)}. \tag{58}$$

The solution is obtained in much the same was as in [2]. It has the same form as (25) (noncausal case) or (35) (causal case) except that the vector $\mathbf{c}(t)$ or $\tilde{\mathbf{c}}(t)$ is calculated by

$$\mathbf{c}(t) \text{ or } \tilde{\mathbf{c}}(t) = (I + Q^{-1}\mathbf{G}(t))^{-1}Q^{-1}y(t), \tag{59}$$

where

$$Q = \operatorname{diag}(q_1, \cdots, q_m) \tag{60}$$

and $\mathbf{G}(t)$ is either $G(t)$ or $\tilde{G}(t)$, defined previously, depending on whether the case is noncausal or causal.

6. Conclusion. We have presented an approach, based on a Fock space framework, for parameter-free nonlinear system identification for a system whose input-output map is expressible by a Volterra series subject to appropriate restrictions, and under the assumption that a set of test input-output pairs is available.

The algorithms described in this paper have been implemented in various computer simulations with encouraging results.

Acknowledgments. We are indebted to an anonymous reviewer and to Professor John Polking for helpful comments on the original version of this paper.

REFERENCES

[1] A. CAPRIHAN AND R. J. P. DEFIGUEIREDO, *The generalized smoothing spline and system identification*, Internal Memorandum, Dept. Electrical Engineering, Rice Univ., Houston, TX, 1971.

[2] R. J. P. DEFIGUEIREDO AND A. CAPRIHAN, *An algorithm for the construction of the generalized smoothing spline with application to system identification*, in Proc. of the 1977 Conference on Information Sciences and Systems, The Johns Hopkins Univ., Baltimore, MD, 1977, pp. 494–500.

[3] A. V. BALAKRISHNAN, *Determination of nonlinear systems from input-output data*, in Proc. Princeton Conference on Identification Problems in Communication and Control, March 1963, pp. 31–49.

[4] H. C. HSIEH, *The least squares estimation of linear and nonlinear system weighting function matrices*, Inform. and Control, 7 (1964), pp. 84–115.

[5] W. L. ROOT, *On system measurement and identification*, in Proc. of Symposium on System Theory, April 1965, Polytechnic Institute of Brooklyn, Brooklyn, NY, 1965, pp. 133–157.

[6] F. J. BEUTLER AND W. L. ROOT, *The operator pseudoinverse in control and system identification*, in Generalized Inverses and Applications, M. F. Nashed, ed., Academic Press, New York, 1976.

[7] J. L. FRANKLIN, *Well posed stochastic extensions of ill-posed linear problems*, J. Math. Anal. Appl., 31 (1970), pp. 682–716.

[8] E. MOSCA, *On a class of ill-posed extension problems and a related gradient iteration*, IEEE Trans. Automat. Control, AC-17 (1972), pp. 459–465.

[9] W. A. PORTER, *Data interpolation, causality structure, and system identification*, Inform. and Control, 29 (1975), pp. 217–233.

[10] ———, *An overview of polynomic system theory*, IEEE Proc., 64 (1976), pp. 18–23.

[11] ———, *Causal realization from input-output pairs*, this Journal, 15 (1977), pp. 120–128.

[12] T. A. W. DWYER, *Partial differential equations in Fischer–Fock spaces for the Hilbert–Schmidt holomorphy type*, Bull. Amer. Math. Soc., 77 (1971), pp. 725–730.

[13] ———, *Holomorphic Fock representations and partial differential equations on countably Hilbert spaces*, Bull. Amer. Math. Soc., 79 (1973), pp. 1045–1050.

[14] V. BARGMANN, *Remarks on a Hilbert space of analytic functions*, Proc. Nat. Acad. Sci. U.S.A., 48 (1962), pp. 199–204.

[15] L. V. ZYLA, *A theory of nonlinear system approximation and identification based on Volterra expansions*, Ph.D. Dissertation, Rice Univ., Houston, TX, 1977.

[16] E. HILLE AND R. S. PHILLIPS, *Functional Analysis and Semi-groups*, AMS Colloquium Publications 31, American Mathematical Society, Providence, RI, 1957.

[17] A. GUICHARDET, *Symmetric Hilbert Spaces and Related Topics*, Lecture Notes in Mathematics, 261, Springer-Verlag, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0008 $01.25/0

NONLINEAR OPTIMAL CONTROL PROBLEMS IN HEAT CONDUCTION*

AVNER FRIEDMAN† AND LI-SHANG JIANG‡

Abstract. We consider two control problems: (a) Maximize the melting in a Stefan problem, given the total heat flux as a control variable; (b) cool the temperature of an endpoint of a rod as much as possible, given the control variable as a certain coefficient in the corresponding parabolic equation. Both problems are solved using the same method.

Key words. optimal control, parabolic equations, Stefan problem, heat conduction

Introduction. In this work we consider optimal control problems for parabolic equations. In the literature one can find models in which the control appears linearly, that is, either as an inhomogeneous term in the parabolic equation or as the initial or boundary data, e.g., [1], [3], [4], [5], [9]. In the present work the control appears in a nonlinear fashion. The existence of an optimal control can be established by a direct argument, using a minimizing sequence. Our interest here is in the actual analysis of the optimal control; in fact, we determine exactly what it is.

The method is based on linearization and it consists of two steps:

(i) First we study optimal ε-perturbation of a given control.

(ii) Next we apply (i) to the optimal control, treating it as an ε-perturbation of another control (which depends on ε).

We shall apply the method to two models:

(a) the Stefan free boundary problem;

(b) a cooling problem for $u_t - u_{xx} + k(x)u = 0$.

In case (a) the control variable is the amount of heat flux used to melt ice; we wish to spread the flux (in time) in such a way as to melt as much ice as possible.

In case (b) the control variable is $k(x)$; we wish to choose it so as to achieve the best cooling result of one endpoint.

Problem (a) is studied in §§ 1, 2 and problem (b) is studied in § 3. The method of this paper should apply to various other models in which the control does not appear in the leading coefficients or in the free boundary condition; see Remark 3.2.

1. The Stefan problem; ε-perturbation. In this section we consider an auxiliary problem that will be used in § 2 in order to solve the optimal control problem (a) mentioned briefly in the introduction.

Consider a one-phase Stefan problem: find u, h satisfying

$$
\begin{aligned}
&u_t - u_{xx} = 0 && \text{if } 0 < x < h(t), \quad 0 < t < T,\\
&u(x, 0) = \phi(x) && \text{if } 0 < x < b \quad (b = h(0) > 0),\\
&-u_x(0, t) = g(t) + \varepsilon\delta(t) && \text{if } 0 < t < T,\\
&u(h(t), t) = 0 && \text{if } 0 < t < T,\\
&\frac{dh}{dt} = -u_x(h(t), t) && \text{if } 0 < t < T.
\end{aligned} \tag{1.1}
$$

* Received by the editors June 16, 1982, and in revised form January 6, 1983. This work was partially supported by the National Science Foundation under grant MCS 7915171.

† Mathematics Department, Northwestern University, Evanston, Illinois, 60201.

‡ Department of Mathematics, Peking University, Peking, China.

Here $\phi(x) \geqq 0$, $g(t) \geqq 0$, $\delta(t) \geqq 0$, $\varepsilon \geqq 0$ and $b > 0$ are given. It is well known that this problem has a unique solution for any (say) continuous data ϕ, g, δ. In fact the solution can be obtained by solving a nonlinear Volterra type integral equation for $u_x(h(t), t)$ (see [2]). By re-examining the proof one easily sees that the proof of existence and uniqueness remains valid even if $g(t)$ and $\delta(t)$ are measures, provided

$$\int_0^T g(t) < \infty, \qquad \int_0^T \delta(t) < \infty.$$

We recall [7], [10] that the free boundary is C^∞ for $t > 0$.

We denote the solution of (1.1) for $\varepsilon > 0$ by u_ε, h_ε. One can establish [8] that

$$\begin{aligned}
&u_\varepsilon(x, t) = u(x, t) + \varepsilon v(x, t) + O(\varepsilon^2), && 0 < x < h(t), \quad 0 < t < T,\\
&\frac{\partial}{\partial x} u_\varepsilon(x, t) = u_x(x, t) + \varepsilon v_x(x, t) + O(\varepsilon^2), && 0 < x < h(t), \quad 0 < t < T,\\
&h_\varepsilon(t) = h(t) + \varepsilon\gamma(t) + O(\varepsilon^2),\\
&\frac{d}{dt} h_\varepsilon(t) = h'(t) + \varepsilon\gamma'(t) + O(\varepsilon^2).
\end{aligned} \tag{1.2}$$

Notice, by comparison, that if $\delta(t) \geqq 0$, then $\gamma(t) \geqq 0$.

We fix $A > 0$.

DEFINITION 1.1. Denote by $\mathscr{A}$ the class of measures $\delta(t)$ $(0 < t < T)$ satisfying

$$\delta(t) \geqq 0, \qquad \int_0^T \delta(t) \leqq A.$$

Set $\gamma(t) = \gamma_\delta(t)$.

Problem (Π_ε). Find $\delta \in \mathscr{A}$ such that

$$\gamma_\delta(T) = \max_{\delta \in \mathscr{A}} \gamma_{\tilde{\delta}}(T).$$

THEOREM 1.1. *There exists a control δ which solves Problem* (Π_ε).

Indeed, any minimizing sequence δ_n has a subsequence for which $\delta_n \to \delta$ (weak convergence of measures) and the corresponding $v = v_{\delta_n}$, $\gamma = \gamma_{\delta_n}$ converge; $\gamma_{\delta_n} \to \gamma$ uniformly, and $\lim \delta_n$, $\lim v_{\delta_n}$, $\lim \gamma_{\delta_n}$ give the maximum.

The main result of this section is the following:

THEOREM 1.2. *The optimal control δ is unique; it coincides with A times the Dirac measure at $t = 0$.*

Thus the best policy for maximizing the amount of melting is to use all the admissible heat flux right at the beginning.

Proof. It is easy to see that $v(x, t)$ satisfies:

$$\begin{aligned}
&v_t - v_{xx} = 0 && \text{if } 0 < x < h(t), \quad 0 < t < T,\\
&v(x, 0) = 0 && \text{if } 0 < x < b,\\
&-v_x(0, t) = \delta(t) && \text{if } 0 < t < T,\\
&v(h(t), t) = \gamma(t) h'(t) && \text{if } 0 < t < T,\\
&\gamma(t) = \int_0^t \delta(s)\, ds - \int_0^{h(t)} v(x, t)\, dx && \text{if } 0 < t < T.
\end{aligned} \tag{1.3}$$

Setting

$$w(x,t)=\int_0^t \delta(s)\,ds-\int_0^x v(y,t)\,dy,$$

we obtain for w the problem

$$\begin{aligned}
&w_t-w_{xx}=0 && \text{if } 0<x<h(t),\quad 0<t<T,\\
&w(x,0)=0 && \text{if } 0<x<b,\\
&w(0,t)=\int_0^t \delta(s)\,ds && \text{if } 0<t<T,\\
&w_x(h(t),t)+h'(t)w(h(t),t)=0 && \text{if } 0<t<T,
\end{aligned}\tag{1.4}$$

and

$$\gamma(t)=w(h(t),t).\tag{1.5}$$

We shall represent w in terms of the Green's function $G(x,t;\xi,\tau)$ $(\tau<t)$ in the domain

$$\Omega\colon 0<x<h(t),\quad 0<t<T;$$

G satisfies

$$\begin{aligned}
&\frac{\partial G}{\partial \tau}+\frac{\partial^2 G}{\partial \xi^2}=0 && \text{in } \Omega\cap\{\tau<t\},\\
&G(x,t;0,\tau)=0,\\
&G_\xi(x,t;h(\tau),\tau)+h'(\tau)G(x,t;h(\tau),\tau)=0,\\
&G(x,t;\xi,\tau)\to\delta_x(\xi) && \text{if } \tau\uparrow t,
\end{aligned}\tag{1.6}$$

where $\delta_x(\xi)$ is the Dirac measure (in ξ) supported at $\xi=x$.

We have

$$w(x,t)=\int_0^t w(0,\tau)G_\xi(x,t;0,\tau)\,d\tau,$$

so that, by (1.5) and the third condition in (1.4),

$$\gamma(t)=\int_0^t F(t,s)\delta(s)\,ds,\tag{1.7}$$

where

$$F(t,s)=\int_s^t G_\xi(h(t),t;0,\tau)\,d\tau.\tag{1.8}$$

We claim

$$\frac{\partial}{\partial s}F(T,s)<0 \quad \text{if } 0\leqq s<T.\tag{1.9}$$

Indeed, using the maximum principle we find that

$$G(x,t;\xi,\tau)>0 \quad \text{if } \xi>0,\quad \tau<t,$$

for any $0<x\leqq h(t)$. Since $G(x, t; 0, \tau)=0$, the maximum principle gives

$$G_\xi(x, t; 0, \tau)>0 \quad (0<x\leqq h(t)).$$

It follows that $F_s(T, s)=-G_\xi(h(T), T, 0, s)<0$.

From (1.9) and (1.7) it is clear that $\gamma(T)$ is maximized if and only if $\delta(s)$ is taken to be the Dirac measure at $s=0$ times the constant A.

2. The general Stefan problem. Consider the Stefan problem for $u(x, t)$, $h(t)$:

$$\begin{aligned}
&u_t-u_{xx}=0 && \text{if } 0<x<h(t), \quad 0<t<T,\\
&u(x, 0)=\phi(x) && \text{if } 0<x<b,\\
&-u_x(0, t)=g(t) && \text{if } 0<t<T,\\
&u(h(t), t)=0 && \text{if } 0<t<T \quad (h(0)=b>0),\\
&h'(t)=-u_x(h(t), t) && \text{if } 0<t<T
\end{aligned} \tag{2.1}$$

with a control function $g(t)$ varying in the control set

$$K=\left\{g(t) \text{ is a nonnegative measure, } \int_0^T g(t)\,dt\leqq A\right\}.$$

We sometimes write $u=u_g$, $h=h_g$.

Problem (Π). Find $g\in K$ such that

$$h_g(T)=\max_{\tilde{g}\in K} h_{\tilde{g}}(T).$$

THEOREM 2.1. *There exists a solution g of Problem* (Π).

The proof is similar to the proof of Theorem 1.1.

The main result of this section is the following:

THEOREM 2.2. *There is a unique solution g of Problem* (Π), *given by* $g(t)=A\delta_0(t)$, *where* $\delta_0(t)$ *is the Dirac measure supported at* $t=0$.

Proof. Suppose the assertion is not true. Then there is an optimal control $g(t)$ and $0<\gamma_1\leqq T$ such that

$$\int_{\gamma_1}^T g(t)\,dt>0.$$

For any small $\delta>0$ there exists an $\eta_\delta>0$ such that

$$\int_{\gamma_1}^{\gamma_1+\eta_\delta} g(t)\,dt=\delta, \qquad \eta_\delta\leqq T.$$

Define

$$l(t)=\begin{cases} g(t)/\delta & \text{in } (\gamma_1, \gamma_1+\eta_\delta),\\ 0 & \text{elsewhere}\end{cases} \tag{2.2}$$

and

$$g_\delta(t)=g(t)-\delta l(t). \tag{2.3}$$

Consider the Stefan problem for $u_{\varepsilon,\delta}$, $h_{\varepsilon,\delta}$:

$$\frac{\partial}{\partial t}u_{\varepsilon,\delta}-\frac{\partial^2}{\partial x^2}u_{\varepsilon,\delta}=0 \qquad \text{if } 0<x<h_{\varepsilon,\delta}(t), \quad 0<t<T,$$

$$
\begin{aligned}
&u_{\varepsilon,\delta}(x,0)=\phi(x) && \text{if } 0<x<b,\\
&-\frac{\partial}{\partial x}u_{\varepsilon,\delta}(0,t)=g_\delta(t)+\varepsilon l(t) && \text{if } 0<t<T,\\
&u_{\varepsilon,\delta}(h_{\varepsilon,\delta}(t),t)=0 && \text{if } 0<t<T,\\
&\frac{d}{dt}h_{\varepsilon,\delta}(t)=-\frac{\partial}{\partial x}u_{\varepsilon,\delta}(h_{\varepsilon,\delta}(t),t) && \text{if } 0<t<T.
\end{aligned}
\tag{2.4}
$$

In view of (2.3) we have

$$u_{\delta,\delta}(x,t)=u(x,t), \qquad h_{\delta,\delta}(t)=h(t),$$

where $u=u_g$, $h=h_g$ is the solution of (2.1) corresponding to the optimal control g.

We now wish to apply the results of § 1 in order to derive a contradiction as $\delta\to 0$. For this we need certain estimates on how the solution of (2.4) behaves as $\delta\to 0$, uniformly in both ε and δ, or actually just for $\varepsilon=\delta$.

Set $u_\delta=u_{0,\delta}$, $h_\delta=h_{0,\delta}$.

LEMMA 2.3.

$$0\leqq h(t)-h_\delta(t)\leqq\delta. \tag{2.5}$$

Proof. By comparison (since $g\geqq g_\delta$)

$$h(t)\geqq h_\delta(t) \quad\text{and}\quad u(x,t)\geqq u_\delta(x,t). \tag{2.6}$$

Integrating the heat equation for u and using the initial and boundary conditions in (2.1) we find that

$$h(t)=\int_0^t g(s)\,ds+\int_0^b \phi(x)\,dx-\int_0^{h(t)} u(x,t)\,dx.$$

A similar formula holds for h_δ. By taking the difference and using (2.6), we find that

$$0\leqq h(t)-h_\delta(t)\leqq\int_0^t (g(s)-g_\delta(s))\,ds=\delta.$$

LEMMA 2.4.

$$|h'(t)-h'_\delta(t)|\leqq C\delta. \tag{2.7}$$

Proof. As already mentioned in the paragraph following (1.1), $h_\delta(t)$ and each of its t-derivatives are bounded independently of δ, if $\varepsilon_0\leqq t\leqq T$ (for any $\varepsilon_0>0$) and $h_\delta(t)$ is continuous in (t,δ) up to $t=0$. Thus Green's function G_δ for the domain

$$\Omega_\delta\colon\ 0<x<h_\delta(t), \qquad 0<t<T$$

exists and the usual estimates hold independently of δ.

We can represent $w=u-u_\delta$ in Ω_δ in the form

$$
\begin{aligned}
w(x,t)=&\int_0^t (g-g_\delta)(\tau)G_\delta(x,t;0,\tau)\,d\tau\\
&-\int_0^t u(h_\delta(\tau),\tau)\frac{\partial}{\partial\xi}G_\delta(x,t;h_\delta(\tau),\tau)\,d\tau.
\end{aligned}
\tag{2.8}
$$

Since $|h_\delta(t)-h(t)|\leqq C\delta$ by (2.5), we have

$$u(h_\delta(\tau),\tau)\leqq C\delta;$$

also

$$G(x, t; 0, \tau) \leqq C_\lambda \quad \text{if } x \geqq \lambda \quad (\forall \lambda > 0).$$

Hence we deduce from (2.8) that

$$|u(x, t) - u_\delta(x, t)| = |w(x, t)| \leqq C_\lambda \delta \quad \text{if } x \geqq \lambda. \tag{2.9}$$

Setting $v_1(t) = u_x(h(t), t)$ we can write an integral equation for v_1 (see [2, Chap. 8])

$$v_1(t) = 2\int_0^t v_1(\tau) G_x(h(t), t; h(\tau), \tau)\, d\tau + 2\int_0^t u(\lambda, \tau) G^\lambda_{x\xi}(h(t), t; \lambda, \tau)\, d\tau,$$

where $G^\lambda(x, t; \xi, \tau)$ is Green's function for the heat operator in the half space $x > \lambda$. A similar integral equation holds for

$$v_2(t) = \frac{\partial}{\partial x} u_\delta(h_\delta(t), t)$$

with $h(t)$, $h(\tau)$ replaced by $h_\delta(t)$, $h_\delta(\tau)$, respectively. Taking the difference and setting

$$V(t) = v_1(t) - v_2(t),$$

we obtain, after using Lemma 2.3 and (2.9),

$$V(t) = 2\int_0^t V(\tau) G^\lambda_x(h(t), t; h(\tau), \tau)\, d\tau$$

$$+ 2\int_0^t v_2(\tau)[G^\lambda_x(h(t), t; h(\tau), \tau) - G^\lambda_x(h_\delta(t), t; h_\delta(\tau), \tau)]\, d\tau + O(\delta).$$

Denote the second integral on the right-hand side by J and take for simplicity $\lambda = 0$. Since

$$G(x, t; \xi, \tau) = K(x, t; \xi, \tau) + K(-x, t; \xi, \tau),$$

where

$$K(x, t; \xi, \tau) = \frac{1}{2\sqrt{\pi}\sqrt{t-\tau}} \exp\left[\frac{|x-\xi|^2}{4(t-\tau)}\right],$$

we easily obtain, using Lemma 2.3,

$$J = 2\int_0^t v_2(\tau)[K(h(t), t; h(\tau), \tau) - K(h_\delta(t), t; h_\delta(\tau), \tau)]\, d\tau + O(\delta) \equiv J_1 + O(\delta).$$

Using the mean value theorem we find that

$$J_1 = 2\int_0^t v_2(\tau) \frac{F(t, \tau)}{(t-\tau)^{3/2}} \int_\tau^t [-[h'(\xi) - h'_\delta(\xi)]]\, d\xi \quad (|F| \leqq C)$$

$$= 2\int_0^t v_2(\tau) \frac{F(t, \tau)}{(t-\tau)^{3/2}} \int_\tau^t V(\xi)\, d\xi\, d\tau$$

$$= \int_0^t H(t, \xi) V(\xi)\, d\xi, \qquad |H(t, \xi)| \leqq \frac{C}{\sqrt{t-\xi}}.$$

Combining the above estimates, we arrive at the inequality

$$|V(t)| \leqq C\int_0^t \frac{V(\tau)}{\sqrt{t-\tau}} + C\delta.$$

Hence $|V(t)| \leqq C\delta$, and (2.7) follows.

Using Lemma 2.4 we shall now complete the proof of Theorem 2.2. Since

$$g(t) = g_\delta(t) + \delta l(t),$$

we can expand

$$h(t) = h_\delta(t) + \delta\gamma(t) + W(t), \qquad u(x,t) = u_\delta(x,t) + \delta v(x,t) + Z(x,t) \tag{2.10}$$

in analogy to (1.2) with $g(t)$, $\delta(t)$ replaced by $g_\delta(t)$, $l(t)$ and ε replaced by δ; W, Z, γ and v depend of course on δ.

We shall prove below that

$$|W(t)| \leqq C\delta^2. \tag{2.11}$$

Suppose for the moment that (2.11) is true. We modify the control $l(t)$ replacing it by $\tilde{l}(t) = \delta_0(t)$ (i.e., the Dirac measure with support at $t = 0$), and set

$$\tilde{g}(t) = g_\delta(t) + \delta\tilde{l}(t).$$

Then in analogy to (2.10), (2.11)

$$\tilde{h}(t) = h_\delta(t) + \delta\tilde{\gamma}(t) + \tilde{W}(t), \qquad \tilde{u}(x,t) = u_\delta(x,t) + \delta\tilde{v}(x,t) + \tilde{Z}(x,t) \tag{2.12}$$

and

$$|\tilde{W}(t)| \leqq C\delta^2. \tag{2.13}$$

From the proof of Theorem 1.2 with $g(t)$, $\delta(t)$ replaced by $g_\delta(t)$, $l(t)$, we have

$$\gamma(T) < \tilde{\gamma}(t) - \alpha\delta \quad \text{for some } \alpha > 0, \tag{2.14}$$

where α is independent of δ; α depends on γ_1 (in the definition of $l(t)$). Indeed, we use here (1.7)–(1.9) and observe that, for any $\varepsilon_0 > 0$,

$$F_s(T,s) \leqq -c < 0 \quad \text{if } 0 \leqq s \leqq T - \varepsilon_0,$$

where c is independent of δ.

Combining (2.10)–(2.14) we see that, for small δ,

$$h(T) < \tilde{h}(T) - \alpha\delta + C\delta^2 < \tilde{h}(T),$$

a contradiction to the maximality of $h(T)$.

We shall now prove (2.11); the proof of (2.13) is similar.

The function Z satisfies

$$\begin{aligned}
&Z_t - Z_{xx} = 0 && \text{if } 0 < x < h_\delta(t), \quad 0 < t < T,\\
&Z(x,0) = 0 && \text{if } 0 < x < b,\\
&Z_x(0,t) = 0 && \text{if } 0 < t < T,\\
&Z(h_\delta(t),t) = u(h_\delta(t),t) - \delta\gamma(t)h'_\delta(t) && \text{if } 0 < t < T.
\end{aligned} \tag{2.15}$$

We compute

$$\begin{aligned}
u(h_\delta(t),t) - \delta\gamma(t)h'_\delta(t) &= -\int_{h_\delta(t)}^{h(t)} u_x(\xi,t)\,d\xi - \delta\gamma(t)h'_\delta(t)\\
&= \int_{h_\delta(t)}^{h(t)} [-u_x(\xi,t) - (-u_x(h(t),t))]\,d\xi + h'(t)(h(t) - h_\delta(t) - \delta\gamma(t))\\
&\quad + \delta\gamma(t)(h'(t) - h'_\delta(t)),
\end{aligned}$$

or

$$Z(h_\delta(t), t) = h'(t)W(t) + \int_{h_\delta(t)}^{h(t)} [-u_x(\xi, t) + u_x(h(t), t)]\, d\xi + \delta\gamma(t)(h'(t) - h'_\delta(t)). \tag{2.16}$$

Next, noting that

$$h(t) = \int_0^t g(s)\, ds + \int_0^b \phi(x)\, dx - \int_0^{h(t)} u(x, t)\, dx,$$

$$h_\delta(t) = \int_0^t g_\delta(s)\, ds + \int_0^b \phi(x)\, dx - \int_0^{h_\delta(t)} u_\delta(x, t)\, dx,$$

and subtracting the respective sides, we get, after using the last equation in (1.3),

$$W(t) = -\int_0^{h_\delta(t)} Z(x, t)\, dx - \int_{h_\delta(t)}^{h(t)} u(x, t)\, dx. \tag{2.17}$$

We introduce the function

$$S(x, t) = -\int_0^x Z(\xi, t)\, d\xi.$$

It satisfies

$$\begin{aligned} &S_t - S_{xx} = 0 \quad \text{if } 0 < x < h_\delta(t), \quad 0 < t < T, \\ &S(x, 0) = 0 \quad \text{if } 0 < x < b, \\ &S(0, t) = 0 \quad \text{if } 0 < t < T, \end{aligned} \tag{2.18}$$

and, by (2.17),

$$S(h_\delta(t), t) = W(t) + \int_{h_\delta(t)}^{h(t)} u(x, t)\, dx. \tag{2.19}$$

Since $S_x(h_\delta(t), t) = -Z(h_\delta(t), t)$, if we make use of (2.16), (2.19) we find that

$$S_x + h'(t)S_x = R(t) \quad \text{on } x = h_\delta(t), \tag{2.20}$$

where

$$R(t) = h'(t) \int_{h_\delta(t)}^{h(t)} u(x, t)\, dx - \int_{h_\delta(t)}^{h(t)} d\xi \int_\xi^{h(t)} \frac{\partial^2 u(\eta, t)}{\partial \eta^2}\, d\eta - \delta\gamma(t)(h'(t) - h'_\delta(t)).$$

Using Lemmas 2.3, 2.4 we easily see that

$$|R(t)| \leqq C\delta^2. \tag{2.21}$$

Since S is a solution of (2.18), (2.20), we deduce from (2.21) (for instance by considering the function S/δ^2) that

$$|S(x, t)| \leqq C\delta^2.$$

Recalling (2.19) and Lemma 2.3, we immediately get (2.11). This completes the proof of Theorem 2.2.

Remark. Suppose we change the control set K in problem (Π), taking

$$K = \left\{ g \in L^\infty(0, T);\ 0 \leqq g(t) \leqq M,\ \int_0^T g(t)\, dt \leqq A \right\},$$

where $MT > A$. Then the optimal control is given by

$$g(t) = \begin{cases} M & \text{if } 0 \leqq t \leqq \dfrac{A}{M}, \\ 0 & \text{if } t > \dfrac{A}{M}. \end{cases}$$

Indeed, the proof is the same as for Theorem 2.2.

3. A cooling problem. Consider the parabolic problem

$$\begin{aligned} &u_t - u_{xx} + k(x)u = 0 && \text{if } 0 < x < 1, \quad 0 < t < T, \\ &u_x(0, t) = 0 && \text{if } 0 < t < T, \\ &u(1, t) = 0 && \text{if } 0 < t < T, \\ &u(x, 0) = 1 && \text{if } 0 < x < 1, \end{aligned} \tag{3.1}$$

where $k(x)$ is a control function. We take the control set to be

$$K = \left\{ k(x) \in L^\infty(0, 1);\ 0 \leqq k(x) \leqq M,\ \int_0^1 k(x)\, dx = \theta,\ k(x) \text{ monotone increasing} \right\},$$

where θ and M are given, and $M > \theta > 0$.

We consider a problem of best cooling:

Problem (Π_K). Find $\tilde{k} \in K$ such that

$$u_k(0, T) = \min_{k \in K} u_{\tilde{k}}(0, T),$$

where u_k, $u_{\tilde{k}}$ are the solutions u corresponding to k and $\tilde{k}$ respectively.

Thus the optimal control is the one which yields the smallest temperature at $x = 0$, $t = T$.

The existence of a solution to this problem can be established by taking minimizing sequences. Our main interest here is in characterizing the optimal control.

THEOREM 3.1. *Problem* (Π_K) *has a unique solution* k^*, *given by* $k^*(x) \equiv \theta$.

Proof. Suppose the assertion is not true. Then there exists an optimal control k, $k \neq k^*$. It follows that for any sufficiently small δ there exist nonnegative functions $l_\delta(x)$, $\tilde{l}_\delta(x)$ with support in intervals $\Sigma = \{0 \leqq x \leqq a\}$, $\tilde{\Sigma} = \{b \leqq x \leqq 1\}$ such that $b - a \geqq c > 0$,

$$\int_\Sigma l_\delta(x)\, dx = \int_{\tilde{\Sigma}} \tilde{l}_\delta(x)\, dx = 1$$

and

$$k(x) + \delta l_\delta(x) - \delta \tilde{l}_\delta(x) \text{ is increasing;} \tag{3.2}$$

c is independent of δ. For simplicity we write $l = l_\delta$, $\tilde{l} = \tilde{l}_\delta$.

For any positive ε consider the problem:

$$
\begin{aligned}
&\frac{\partial}{\partial t}u_{\varepsilon,\delta}-\frac{\partial^2}{\partial x^2}u_{\varepsilon,\delta}+(k(x)+\varepsilon l(x))u_{\varepsilon,\delta}=0 \quad \text{if } 0<x<1, \quad 0<t<T,\\
&\frac{\partial}{\partial x}u_{\varepsilon,\delta}(0,t)=0 \qquad\qquad \text{if } 0<t<T,\\
&u_{\varepsilon,\delta}(1,t)=0 \qquad\qquad \text{if } 0<t<T,\\
&u_{\varepsilon,\delta}(x,0)=1 \qquad\qquad \text{if } 0<x<1.
\end{aligned} \tag{3.3}
$$

We set

$$u_{\varepsilon,\delta}=v+\varepsilon w+z \tag{3.4}$$

and substitute this into (3.3). Comparing powers of ε^0, ε we find that (3.3) holds if

$$
\begin{aligned}
&v_t-v_{xx}+kv=0 && \text{for } 0<x<1, \quad 0<t<T,\\
&v_x(0,t)=0, \quad v(1,t)=0 && \text{for } 0<t<T, \qquad (3.5)\\
&v(x,0)=1 && \text{for } 0<x<1;\\
&w_t-w_{xx}+kw=-l(x)v && \text{for } 0<x<1, \quad 0<t<T,\\
&w_x(0,t)=0, \quad w(1,t)=0 && \text{for } 0<t<T, \qquad (3.6)\\
&w(x,0)=0 && \text{for } 0<x<1;
\end{aligned}
$$

and

$$
\begin{aligned}
&z_t-z_{xx}+(k+\varepsilon l)z=-\varepsilon^2 l(x)w && \text{for } 0<x<1, \quad 0<t<T,\\
&z_x(0,t)=0, \quad z(1,t)=0 && \text{for } 0<t<T, \qquad (3.7)\\
&z(x,0)=0 && \text{for } 0<x<1.
\end{aligned}
$$

Denotc by $G(x,t;\xi,\tau)$ the Green's function for (3.5), that is,

$$
\begin{aligned}
&G_{\xi\xi}+G_\tau+kG=0 \quad (\tau<t),\\
&G_\xi(x,t;0,\tau)=0, \quad G(x,t;1,\tau)=0,\\
&G(x,t;\xi,\tau)\to\delta_0(\xi-x) \quad \text{as } \tau\to t.
\end{aligned}
$$

Then we can write

$$v(x,t)=\int_0^1 G(x,t;\xi,0)\,d\xi \tag{3.8}$$

and

$$w(x,t)=\int_0^1\int_0^t G(x,t;\xi,\tau)[-l(\xi)v(\xi,\tau)]\,d\xi\,d\tau=-\int_0^1 F(x,\xi,\tau)l(\xi)\,d\xi, \tag{3.9}$$

where

$$F(x,\xi,t)=\int_0^t d\tau\int_0^1 G(x,t;\xi,\tau)G(\xi,\tau;\eta,0)\,d\eta. \tag{3.10}$$

We introduce also the Green's function $G_\varepsilon(x, t; \xi, \tau)$ for (3.5) with $k(x)$ replaced by $k(x)+\varepsilon l(x)$. Then we can represent the solution z of (3.7) in the form

$$z(x,t) = -\varepsilon^2 \int_0^1 \int_0^t G_\varepsilon(x, t; \xi, \tau) l(\xi) w(\xi, \tau)\, d\xi\, d\tau; \tag{3.11}$$

also,

$$w(x,t) = -\int_0^t \int_\Sigma G(x, t; \xi, \tau) l(\xi) v(\xi, \tau)\, d\xi\, d\tau. \tag{3.12}$$

We observe that if the coefficients $k(x)$ and $l(x)$ are Hölder continuous then G and G_ε can be constructed by the parametrix method. Furthermore, they are majorized by the fundamental solutions, so that, if $|k(x)| \leqq C$, $|k(x)+\varepsilon l(x)| \leqq C$, then

$$G(x, t; \xi, \tau) \leqq \frac{C}{\sqrt{t-\tau}}, \qquad G_\varepsilon(x, t; \xi, \tau) \leqq \frac{C}{\sqrt{t-\tau}}. \tag{3.13}$$

For L^∞ coefficients k and l, the Green's function can be obtained by approximating with Green's functions for smooth coefficients k_m, l_m ($k_m \to k$, $l_m \to l$); from the parametrix method we obtain (3.13) for the approximating Green's functions with C independent of m; hence (3.13) holds even without the assumption of Hölder continuity.

We now specialize to $\varepsilon = \delta$. Then $|k(x)| \leqq M$, $|\delta l(x)| \leqq M$ and thus (3.13) holds. Using this in (3.9), (3.11) we obtain, since $\int l(x) = 1$,

$$|w(x,t)| \leqq C, \tag{3.14}$$

$$|z(x,t)| \leqq C\delta^2. \tag{3.15}$$

Setting $F(\xi) = F(0, \xi, T)$, we claim that

$$F'(\xi) < 0. \tag{3.16}$$

Indeed,

$$F'(\xi) = \int_0^T d\tau \int_0^1 G_\xi(0, T; \xi, \tau) G(\xi, \tau; \eta, 0)\, d\eta$$

$$+ \int_0^T d\tau \int_0^1 G(0, T; \xi, \tau) G_\xi, (\xi, \tau; \eta, 0)\, d\eta = J_1 + J_2.$$

By (3.8)

$$J_2 = \int_0^T G(0, T; \xi, \tau) v_\xi(\xi, \tau)\, d\xi,$$

where v is the solution of (3.5). By the maximum principle $v_\xi(1, t) \leqq 0$. Since

$$(v_\xi)_t - (v_\xi)_{\xi\xi} + k v_\xi = -k' v \leqq 0,$$

we can apply the maximum principle to v_ξ and deduce that $v_\xi < 0$ in $0 < \xi < 1$, $0 < t < T$. Consequently, $J_2 < 0$.

Treating in the same way solutions $U = U_m$ of

$$-U_\tau - U_{\xi\xi} + k(\xi) U = 0 \qquad (0 < \xi < 1, 0 < \tau < T),$$

$$U_\xi(0, \tau) = 0, \quad U(1, \tau) = 0 \qquad (0 < \tau < T),$$

$$U(\xi, T) = \varphi_m(\xi) \to \delta_0(\xi) \qquad \text{as } m \to \infty$$

with $\varphi_m'(\xi) \leqq 0$, we see that

$$G_\xi(0, T; \xi, \tau) < 0 \quad (0 < \xi < 1, 0 < \tau < T),$$

and consequently also $J_1 < 0$. It follows that (3.16) holds. In fact the above proof yields also the estimate

$$F'(\xi) < -C < 0, \tag{3.17}$$

where C is a constant independent of δ.

Let $k_\delta(x) = k(x) + \delta l(x)$. Replacing k by k_δ in the above analysis and l by $-\tilde{l}$ (notice that $k_\delta'(x) \geqq 0$, by (3.2)) we obtain a solution

$$\tilde{u}_{\varepsilon,\delta} = v + \varepsilon\tilde{w} + \tilde{z}$$

analogous to (3.4), and $\tilde{u}_{\delta,\delta}$ is the solution u of (3.1) corresponding to the control $\tilde{k} = k_\delta + \delta\tilde{l}$ which belongs to K (recall (3.2)). In view of (3.17),

$$\tilde{w}(0, T) < w(0, T) - \alpha\delta \quad (\alpha > 0),$$

where α is a constant independent of δ. Taking $\varepsilon = \delta$ and making use of (3.15) and of the corresponding estimate for $\tilde{z}$, we conclude that

$$\tilde{u}_{\delta,\delta}(0, T) < u(0, T) - \alpha\delta + C\delta^2 < u(0, T),$$

which is a contradiction to the minimality of $u(0, T)$.

Remark 3.1. The method of this section does not seem to apply to the corresponding cooling problem for

$$u_t - (k(x)u_x)_x = 0 \quad (0 < x < 1, 0 < t < T),$$

where k varies in the set:

$$\{k \text{ measurable, meas } \{k(x) = 1\} = \theta, \text{ meas } \{k(x) = 2\} = 1 - \theta\}.$$

Partial results for this problem were obtained by Joel Friedman [6].

Remark 3.2. The method of this paper extends to other free boundary problems with one-dimensional spacial variable x (instead of the Stefan problem), to other objective functionals (such as a weighted average of the temperature $\int_0^1 \alpha(x)u(x, T)\,dx$ in the rod problem), and to hyperbolic equations such as $u_{tt} - u_{xx} = k(x)u$ ($k(x)$ is the control). Problems with x in R^n will require deeper analysis of the ε-perturbation problem.

REFERENCES

[1] H. O. FATTORINI, *Time optimal control of solutions of operational differential equations*, this Journal, 2 (1964), pp. 54–59.

[2] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[3] ———, *Optimal control for parabolic equations*, J. Math. Anal. Appl., 18 (1967), pp. 479–491.

[4] ———, *Optimal control in Banach spaces*, J. Math. Anal. Appl., 19 (1967), pp. 35–55.

[5] ———, *Optimal control in Banach spaces with fixed end-points*, J. Math. Anal. Appl., 24 (1968), pp. 161–181.

[6] J. FRIEDMAN, *Optimal control for cooling problems* (unpublished), 1980.

[7] L.-S. JIANG, *Existence and differentiability of the solution of a two-phase Stefan problem for quasilinear parabolic equations*, Chinese Math. Acta, 6 (1965), pp. 481–496.

[8] P. JOCHUM, *Differentiable dependence upon the data in a one-phase Stefan problem*, Math. Meth. Appl. Sci., 2 (1980), pp. 73–90.

[9] J. L. LIONS, *Contrôle optimal de systèmes gouvernés par des équations aux dérivées partielles*, Dunod, Paris, 1968.

[10] D. G. SCHAEFFER, *A new proof of the infinite differentiability of the free boundary in the Stefan problem*, J. Differential Equations, 20 (1976), pp. 266–269.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0009 $01.25/0

PROPERTIES OF RELAXED TRAJECTORIES FOR A CLASS OF NONLINEAR EVOLUTION EQUATIONS ON A BANACH SPACE*

N. U. AHMED†

Abstract. In this paper, we consider a class of nonconvex control problems in a Banach space, which is convexified by introducing measure-valued controls. We study some topological properties of the set of trajectories of the original system and that of the relaxed system. It is shown that, under certain reasonable assumptions, the set of trajectories of the original system is dense (in an appropriate topology) in the set of relaxed trajectories. As a corollary of this result it follows that the attainable set of the original system is dense in that of the relaxed system. These results are directly useful in terminal optimization problems and time-optimal control problems. For illustration, we present two examples in § 6.

Key words. nonlinear evolution equations, measure-valued controls, original and relaxed systems, Cesàri property, original and relaxed trajectories, attainable set, terminal control, time-optimal control

1. Introduction. It is known that convexity conditions play a central role in the study of existence of optimal controls. In fact if the tangent bundle (also known as the orientor field or the velocity field) satisfies the convexity condition or more precisely the Cesari property [1], [2], [3], [4], [5], [6], [7], [8], in the general case, then under some additional regularity assumptions [1], one can prove the existence of optimal controls. Both Kuratowski and Cesari properties have been widely used in the literature for this purpose [1], [2], [5], [6], [7], [8], [18].

If an original optimal control problem does not satisfy the convexity condition one may introduce Radon measures or, in particular, probability measures as generalized controls. This convexifies the tangent bundle of the controlled system thereby making it possible to prove the existence of optimal (generalized) controls for the relaxed system. Once this is done, a natural question arises as to whether the relaxed optimal trajectory can be approximated closely by a trajectory of the original control problem. This question can be answered by showing if or not the set of original trajectories is dense (in an appropriate topology) in the set of relaxed trajectories.

For finite dimensional problems and for systems governed by functional equations, this question has been studied in the literature [9], [10], [11].

In this paper we wish to study this question for a class of controlled nonlinear evolution equations on a Banach space. Under the usual convexity condition optimal control of this class of systems was studied by Ahmed and Teo [1], [2]. Using the Cesari property, the existence of optimal controls for a Lagrange problem was also shown in Ahmed and Teo [1]. In this paper we remove the convexity hypothesis, introduce the generalized controls and prove the density of the set of original trajectories in the set of relaxed trajectories.

2. Basic notation. Let H be a separable Hilbert space and E a dense linear subspace of H carrying the structure of a reflexive Banach space with H^* and E^* denoting the corresponding (topological) duals. It is assumed that the injection $E \hookrightarrow H$ is continuous and E is dense in H. Identifying H^* with H we have $E \subset H \subset E^*$. Let $\langle y, x\rangle_{E^*-E}$ denote the duality product of an element y of E^* with an element x of E. For scalar products in H we use the notation $(y, x)_H$. The norms will be denoted by $|\cdot|_G$ for $G = (E, H, E^*)$. If $y \in H$ and $x \in E$ then $\langle y, x\rangle_{E^*-E} = (y, x)_H$. The Banach space E, furnished with the weak topology, will be denoted by either E_w or (E, τ_{E^*}). Let

*Received by the editors January 15, 1982, and in revised form July 6, 1982. This work was supported in part by the National Science and Engineering Council of Canada under grant no. 7109.

† Department of Electrical Engineering, University of Ottawa, Ottawa, Ontario, Canada K1N 6N5.

$I=(0, T)$ be any bounded interval in $R_+=[0, \infty)$ and denote by $L_p(I, G)$ the equivalence classes of strongly measurable functions on I with values in the Banach space G and furnished with the norm topology

$$\|x\|_{L_p(I,G)} \equiv \begin{cases} \left(\int_I |x(t)|_G^p \, dt\right)^{1/p} & \text{for } 1 \leqq p < \infty, \\ \text{ess sup } \{|x(t)|_G, t \in I\} & \text{for } p = \infty. \end{cases}$$

Here G stands for any of the Banach spaces E, H or E^* and $1 \leqq p \leqq \infty$. For $1<p$, $q<\infty$ and G a reflexive Banach space, the spaces $L_p(I, G)$ and $L_q(I, G^*)$ are also reflexive and $(L_p(I, G))^* = L_q(I, G^*)$ provided $p^{-1}+q^{-1}=1$. We are mostly interested in the Banach spaces $L_p(I, E)$ and $L_p(I, H)$ and the corresponding duals $L_q(I, E^*)$ and $L_q(I, H)$. For a Banach space G, $C(I, G)$ denotes the space of (strongly) continuous functions on I with values in G. Furnished with the sup norm $\|x\|_{C(I,G)} \equiv \sup \{|x(t)|_G, t \in I\}$, $C(I, G)$ is a Banach space. For $G_w = (G, \tau_{G^*})$, $C(I, G_w)$ denotes the topological vector space of functions defined on I with values in the Banach space G and continuous in the weak topology. This is a locally convex complete topological vector space.

3. Admissible controls. Let B be a Polish space and Γ a closed subset of B. Let $\mathcal{U}$ denote the set of all measurable functions $\{u\}$ defined on I with values $u(t) \in \Gamma$ a.e. We call $\mathcal{U}$ the class of original controls.

Let $M(\Gamma)$ denote the space of all probability measures on the Borel σ-field of Γ and $C_b(\Gamma)$ the space of all real-valued bounded continuous functions on Γ. In $M(\Gamma)$ we define a base of neighbourhoods:

$$\{N(\nu, F, \varepsilon): \nu \in M(\Gamma), F \text{ finite subsets of } C_b(\Gamma) \text{ and } \varepsilon > 0\},$$

where

$$N(\nu, F, \varepsilon) \equiv \left\{\mu \in M(\Gamma): \left|\int_\Gamma g \, d\mu - \int_\Gamma g \, d\nu\right| < \varepsilon, g \in F\right\}.$$

This base of neighbourhoods defines a topology on $M(\Gamma)$ and with this topology $M(\Gamma)$ becomes a Hausdorff topological space. A net $\{\mu^\alpha\} \in M(\Gamma)$ converges weakly to an element $\mu \in M(\Gamma)$ if and only if, for every $g \in C_b(\Gamma)$

$$\int_\Gamma g \, d\mu^\alpha \to \int_\Gamma g \, d\mu.$$

Since $M(\Gamma) \subset C_b^*(\Gamma)$, the topology defined above is also known as the (weak star) w^*-topology and the associated convergence as the w^*-convergence. This topology is metrizable and $M(\Gamma)$ with this topology becomes a separable metric space [12, Thm. 6.2, p. 43]. Further if Γ is compact, then $M(\Gamma)$ is a compact metric space [12, Thm. 6.4, p. 44]. Since Γ is a closed subset of the Polish space B, $M(\Gamma)$ is topologically complete, that is, $M(\Gamma)$ is homeomorphic to a complete metric space and hence $M(\Gamma)$ is also a Polish space. Let $\mathcal{M}$ denote the family of all weakly measurable functions $\{\mu\}$ defined on I with values $\mu_t \in M(\Gamma)$ for all $t \in I$. By weak measurability of μ we mean that the numerical valued function

$$t \to \int_\Gamma g \, d\mu_t$$

is Lebesgue measurable for each $g \in C_b(\Gamma)$.

4. The controlled system and basic assumptions. We consider that the controlled system is governed by the following nonlinear evolution equation

$$\frac{dx}{dt} = A(t)x + f(t, x, u), \quad t \in I, \qquad x(0) = x_0 \tag{1}$$

in the Banach space E, where $u \in \mathcal{U}$, $\{A(t), t \in I\}$ is a family of densely defined linear operators, not necessarily bounded in H, with domains $D(A(t)) \subset E$ and range $R(A(t)) \subset E^*$ and, in general, $f: I \times E \times \Gamma \to E^*$. We call this the original system.

Similary for a $\mu \in \mathcal{M}$, we consider the evolution equation

$$\frac{dx}{dt} = A(t)x + \int_\Gamma f(t, x, \sigma)\, d\mu_t(\sigma), \quad t \in I, \qquad x(0) = x_0. \tag{2}$$

We call this the relaxed system.

For $1 < p, q < \infty$ with $p^{-1} + q^{-1} = 1$, let L denote the operator determined by

$$D(L) \equiv \{x \in L_p(I, E): x(t) \in D(A(t)) \cap D(A^*(t)) \text{ for } t \in I,$$
$$\text{and } A(t)x(t), A^*(t)x(t), \dot{x}(t) \in C^0(I, E^*) \cap L_q(I, E^*)\}$$

with

$$(Lx)(t) \equiv \left(\frac{d}{dt} + A(t)\right)x, \quad t \in I \quad \text{for } x \in D(L).$$

We assume throughout the paper that L is densely defined as a linear operator from $L_p(I, E)$ to $L_q(I, E^*)$ and that the strong and weak extensions of L from $L_p(I, E)$ to $L_q(I, E^*)$ coincide (that is $L_s = L_w$) [13], [1], [2].

For $u \in \mathcal{U}$ and $x_0 \in D(A(t))$, $t \in I$, an element $x_u \in L_p(I, E)$ is said to be a strong solution of the evolution equation (1) if $x_u(0) = x_0$ and $(L_s x)(t) = f(t, x_u(t), u(t))$ a.e. on I. We denote this family of solutions by $X \equiv \{x_u: u \in \mathcal{U}\}$ and call it the set of original trajectories.

Similarly for a generalized control $\mu \in \mathcal{M}$ and $x_0 \in D(A(t))$, $t \in I$, an element $x_\mu \in L_p(I, E)$ is a strong solution of the evolution equation (2) if $x_\mu(0) = x_0$ and $(L_s x_\mu)(t) = \int_\Gamma f(t, x_\mu(t), \sigma)\, d\mu_t(\sigma)$ a.e. on I. We denote this family of solutions by $X_r \equiv \{x_\mu: \mu \in \mathcal{M}\}$ and call it the set of relaxed trajectories.

We assume throughout the presentation, unless stated otherwise, that the operators A and f satisfy the following conditions.

Assumption (A1). $\{A(t), t \in I\}$ is a family of densely defined linear operators in H (not necessarily bounded) with domains $D(A(t)) \subset E$ and range $R(A(t)) \subset E^*$ for $t \in I$.

Assumption (A2). $\langle A(t)e, e\rangle_{E^*-E} \leqq 0$ for all $e \in D(A(t)) \subset E$, $t \in I$, and the strong and weak extensions of L from $L_p(I, E)$ to $L_q(I, E^*)$ coincide.

Assumption (F1). The function $t \to \langle f(t, e, \sigma), g\rangle_{E^*-E}$ is measurable on I for arbitrary $e, g \in E$ and $\sigma \in B$ and $f: I \times E \times B \to E^*$ is demicontinuous in the sense that whenever $t_n \to t$ in I, $\xi_n \underset{n}{\to} \xi$ in E and $v_n \to v$ in B

$$\langle f(t_n, \xi_n, v_n), e\rangle_{E^*-E} \to \langle f(t, \xi, v), e\rangle_{E^*-E}$$

for each $e \in H$.

Assumption (F2).

$$\langle f(t, x, \sigma) - f(t, y, \sigma), x - y\rangle_{E^*-E} \leqq 0$$

for all $x, y \in E$ and $\sigma \in \Gamma$.

Assumption (F3). There exists an $h \in L_q(I, R_+)$ and $\alpha \geqq 0$ such that

$$|f(t, x, \sigma)|_{E^*} \leqq h(t) + \alpha |x|_E^{p/q} \quad \text{a.e.}$$

for each $x \in E$ and for all $\sigma \in \Gamma$.

Assumption (F4). There exists an $h_1 \in L_1(I, R_+)$, $\beta > 0$ such that

$$\langle f(t, x, \sigma), x \rangle_{E^*-E} \leqq h_1(t) - \beta |x|_E^p \quad \text{a.e.}$$

for each $x \in E$ and for all $\sigma \in \Gamma$.

LEMMA 4.1. *Under the assumptions* (A1)–(A2) *and* (F1)–(F4):

(i) *For each (ordinary) control* $u \in \mathcal{U}$ *and initial state* $x_0 \in E$, *the evolution equation* (1) *has a unique strong solution* $x \in L_p(I, E)$.

(ii) *For each (relaxed) control* $\mu \in \mathcal{M}$ *and initial state* $x_0 \in E$, *the evolution equation* (2) *has a unique strong solution* $x \in L_p(I, E)$.

Further, in either case, $x \in C(I, E_w) \cap C(\bar{I}, H)$ *and* $x(t) \in D(A(t))$ *a.e.*

This result essentially follows from a general result due to Browder [13, Thm. 5, p. 54] and has been used in control theory [2, Thm. 3.1, p. 62], [1, Thm. 2.5.2, p. 111]. Here we have stated the result under slightly stronger condition (Assumption (F1)) than necessary for the proof of only existence of solutions. This is essential for our purpose.

Remark 4.1. The result (ii) of Lemma 4.1 also follows from Browder's general result by simply requiring that the mapping $\tilde{f}(t, x, \mu): I \times E \times M(\Gamma) \to E^*$, defined by

$$\tilde{f}(t, x, \mu) = \int_\Gamma f(t, x, \sigma)\, d\mu(\sigma),$$

satisfy the properties (F2)–(F4) uniformly with respect to $\mu \in M(\Gamma)$.

5. Properties of ordinary and relaxed trajectories. For each $(t, x) \in \Delta \equiv \{(t, x) \in I \times E: x \in D(A(t))\}$ the set $R(t, x)$, defined by

$$R(t, x) \equiv \{e^* \in E^*: e^*(\phi) = \langle A(t)x, \phi \rangle_{E^*-E} + \langle f(t, x, \sigma), \phi \rangle_{E^*-E} \quad \text{for some } \sigma \in \Gamma \text{ and for all } \phi \in D(A^*(t))\},$$

determines the tangent bundle (velocity field) for the original system (1). Similarly for each $(t, x) \in \Delta$ the set $R_r(t, x)$ given by

$$R_r(t, x) \equiv \{e^* \in E^*: e^*(\phi) = \langle A(t)x, \phi \rangle_{E^*-E} + \int_\Gamma \langle f(t, x, \sigma), \phi \rangle\, d\mu(\sigma) \quad \text{for some } \mu \in M(\Gamma) \text{ and for all } \phi \in D(A^*(t))\}$$

determines the velocity field for the relaxed system (2). Set

$$R_0(t, x) \equiv \text{Cl Co}\, R(t, x) \tag{3}$$

and consider the differential inclusions

$$\dot{x}(t) \in R_0(t, x(t)), \quad t \in I, \qquad x(0) = x_0 \tag{4}$$

and

$$\dot{x}(t) \in R_r(t, x(t)), \quad t \in I, \qquad x(0) = x_0. \tag{5}$$

The set X_r given by

$$X_r \equiv \{x \in C(I, E_w) \cap D(L): x(0) = x_0 \text{ and } \dot{x}(t) \in R_r(t, x(t)) \text{ a.e. in } I\},$$

defines the family of trajectories generated by the differential inclusion (5). As we shall see later, the set X_r is, in fact, the admissible trajectories of the relaxed dynamical system (2) (Corollary 5.1). Similarly we define the set X_0 as being the admissible trajectories of the differential inclusion (4) corresponding to the tangent bundle R_0.

Denote by $A(A^*)$ the function $t \to A(t)(A^*(t))$ from I into $\mathscr{L}(E, E^*)$, where $\mathscr{L}(E, E^*)$ denotes the space of linear operators (not necessarily bounded) from E to E^*. Let us define

$$D(A) \equiv \{g \in L_p(I, E): (Ag)(t) \equiv A(t)g(t) \text{ is defined a.e. on } I \text{ and } Ag \in L_q(I, E^*)\}$$

and similarly

$$D(A^*) \subset L_p(I, E).$$

In the sequel we need the following result.

LEMMA 5.1. *Suppose the operators A and f satisfy the assumptions* (A1)–(A2) *and* (F1)–(F4) *respectively. Then the sets X, X_r and X_0 are all conditionally sequentially compact subsets of $C(I, E_w)$.*

Proof. The proof follows from arguments similar to those given in [2, Thm. 4.1].

Note. The assumption: $D(A^*)$ is a set of category II in $L_p(I, E)$ [2, Lemma 4.2, p. 64]: is not necessary. It was used merely to prove that the set $\{\dot{x}, x \in X\}$ is a bounded subset of $L_q(I, E^*)$. But it follows from direct computation using the facts that (i) the set of trajectories X are strong solutions to (1) contained in a bounded subset of $C(I, E_w) \cap L_p(I, E)$ and hence the set $\{\tilde{A}x, x \in X\}$, $\tilde{A}$ denoting the weak or strong extension on A in $L_p(I, E)$, is a bounded subset of $L_q(I, E^*)$; and (ii) the growth condition (F3) implies that f maps bounded subsets of $L_p(I, E)$ into bounded subsets of $L_q(I, E^*)$.

With the help of the above preparatory results we can now prove our main results.

THEOREM 5.1. *Suppose that the assumptions of Lemma* 5.1 *hold and that Γ is a compact subset of the Polish space B. Then*:

(i) *for each $(t, x) \in \Delta$, $R_r(t, x)$ is a closed convex subset of E^*;*

(ii) $X_r = X_0$;

and

(iii) X_r *(hence X_0) is a sequentially compact subset of $C(I, E_w)$.*

Proof. (i) Since $M(\Gamma)$ is the space of probability measures, $R_r(t, x)$, for $(t, x) \in \Delta$, is obviously convex. The closure follows from the facts that $M(\Gamma)$ is compact whenever Γ is so and that $\sigma \to \langle f(t, x, \sigma), e\rangle_{E^*-E}$ is continuous on Γ for each $(t, x) \in \Delta$ and $e \in E$.

(ii) For the proof of this it suffices to show that $R_r(t, x) = R_0(t, x)$ for $(t, x) \in \Delta$. Since Γ is a compact subset of a Polish space B, $M(\Gamma)$ is compact and, being the space of probability measures, is obviously convex. Thus it follows from the Krein–Milman Theorem [14, Thm. 4, p. 440], [1. Thm. 1.1.15, p. 11] that ext $M(\Gamma)$ (≡the set of extremal points of $M(\Gamma)$) is nonempty. Let δ_σ denote the Dirac measure on Γ with support concentrated at the point $\sigma \in \Gamma$. Define $M_0 = \{\delta_\sigma: \sigma \in \Gamma\}$. Clearly $M_0 \subset M(\Gamma)$ and $M_0 = \text{ext } M(\Gamma)$. Again by the Krein–Milman theorem Cl Co $M_0 = M(\Gamma)$. We show that this implies that $R_r(t, x) = R_0(t, x)$ for $(t, x) \in \Delta$. First, we show that $R_0(t, x) \subset R_r(t, x)$. Indeed, for any $\nu \in \text{Co } M_0$, there exist an integer n, numbers $\{\alpha_i: i = 1, 2, \cdots, n\}$ with $\alpha_i \geqq 0$, $\sum \alpha_i = 1$ and $\{\sigma_i: i = 1, 2, \cdots, n\} \in \Gamma$ such that $\nu = \sum_{i=1}^n \alpha_i \delta_{\sigma_i}$ and for $e \in E$,

$$\int_\Gamma \langle f(t, x, \sigma), e\rangle \, d\nu = \sum_{i=1}^n \alpha_i \langle f(t, x, \sigma_i), e\rangle.$$

Since $\nu \in M(\Gamma)$ also, the set

$$\operatorname{Co} R(t,x) \equiv \Big\{ e^* \in E^* : e^*(e) = \langle A(t)x, e\rangle + \int_\Gamma \langle f(t,x,\sigma), e\rangle \, d\nu \quad \text{for some } \nu \in \operatorname{Co} M_0 \text{ and for all } e \in D(A^*(t)) \Big\}$$

is clearly contained in $R_r(t,x)$, and, since $R_r(t,x)$ is closed and convex,

$$\operatorname{Cl}\operatorname{Co} R(t,x) \equiv R_0(t,x) \subset R_r(t,x).$$

Now we show that $R_r(t,x) \subset R_0(t,x)$. Let $e^* \in R_r(t,x)$; then, by definition, there exists a $\mu \in M(\Gamma)$ such that, for all $e \in D(A^*(t))$,

$$e^*(e) = \langle A(t)x, e\rangle + \int_\Gamma \langle f(t,x,\sigma), e\rangle \, d\mu(\sigma). \tag{6}$$

Since $\operatorname{Cl}\operatorname{Co} M_0 = M(\Gamma)$, there exists a sequence $\{\nu^n\} \in \operatorname{Co} M_0$ such that $\nu^n \to \mu$ weakly in $M(\Gamma)$ and hence

$$\int_\Gamma \langle f(t,x,\sigma), e\rangle \, d\nu^n \to \int_\Gamma \langle f(t,x,\sigma), e\rangle \, d\mu. \tag{7}$$

Let $e_n^* \in E^*$ correspond to ν^n defined by

$$e_n^*(e) = \langle A(t)x, e\rangle + \int_\Gamma \langle f(t,x,\sigma), e\rangle \, d\nu^n$$

for all $e \in D(A^*(t))$. Clearly $e_n^* \in \operatorname{Co} R(t,x)$, and, since, by virtue of (6) and (7), $e_n^* \to e^*$ in the w^*-topology on E^*, we conclude that $e^* \in \operatorname{Cl}\operatorname{Co} R(t,x)$. Thus $R_r(t,x) \subset R_0(t,x)$ for arbitrary $(t,x) \in \Delta$. This proves that $R_r(t,x) = R_0(t,x)$ and hence the differential inclusions (4) and (5) are equivalent and consequently $X_r = X_0$.

(iii) By Lemma 5.1 X_r (hence X_0) is a conditionally sequentially compact subset of $C(I, E_w)$. Therefore it suffices to prove its closure. Let $\{x_n\} \in X_r$ and suppose $x_n \to x^*$ in $C(I, E_w)$. As in Ahmed and Teo [2, Thm. 5.2, p. 73] we can show that (1) $\dot{x}^*(t)$ exists a.e. on I with $\dot{x}^* \in L_q(I, E^*)$, (2) $x^*(t) \in D(A(t))$ a.e. on I and (3) $\dot{x}^*(t) \in R_r(t, x(t))$ a.e. on I. (Note that for the proof of the last inclusion (3) we need closure and convexity of $R_r(t, x(t))$ which follow from part (i) of this theorem and consequently the assumption (F5) in [2] is not required here.) As a consequence, there exists, for almost all $t \in I$, a $\mu_t \in M(\Gamma)$ such that

$$\langle \dot{x}^*(t), e\rangle = \langle A(t)x^*(t), e\rangle + \int_\Gamma \langle f(t, x^*(t), \sigma), e\rangle \, d\mu_t \tag{8}$$

for all $e \in D(A^*(t))$. Let μ denote the mapping $t \to \mu_t$ on I. The crucial question is whether or not we can select a measurable substitute μ^* for μ (that is $\mu^* \in \mathcal{M}$) such that (8) holds with μ^* replacing μ. For each $t \in I$, for which $\dot{x}^*(t)$, $A(t)x^*(t)$ and $\sigma \to f(t, x^*(t), \sigma)$ exist, define the set

$$G^*(t) \equiv \Big\{ \nu \in M(\Gamma) : \langle \dot{x}^*(t), e\rangle = \langle A(t)x^*(t), e\rangle + \int_\Gamma \langle f(t, x^*(t), \sigma), e\rangle \, d\nu \text{ for all } e \in D(A^*(t)) \Big\}. \tag{9}$$

This defines a set-valued mapping G^* from I into $2^{M(\Gamma)}$, and the problem of existence of a $\mu^* \in \mathcal{M}$ reduces to the problem of existence of a measurable selection of G^*. We

show that G^* has indeed a measurable selection. For this we note that since Γ is a closed subset of a Polish space B, $M(\Gamma)$ is a separable metric space with a metric compatible with its topology. Thus if we can show that $t \to G^*(t)$ is a measurable multifunction (set-valued function) with closed or complete values, then by virtue of a selection theorem due to Himmelberg, Jacobs and Van Vleck [15, Thm. 1, p. 278], [1, Thm. 1.4.4, p. 40] we shall be able to conclude that there exists a (weakly) measurable function μ^* from I into $M(\Gamma)$ such that $\mu_t^* \in G^*(t)$ for $t \in I$. This would then imply that $x^* \in X_r$ and hence that X_r is closed. That G^* has closed (hence complete) values $G^*(t)$, follows from the facts that, for any arbitrary $e \in E$, the function $\sigma \to \langle f(t, x^*(t), \sigma), e\rangle$ is continuous on Γ and that $M(\Gamma)$ is compact. It remains to show that G^* is a measurable multifunction. By virtue of the equivalence of conditions (i) and (v) [15, Thm. 1, p. 278], [1, Tm. 1.4.3, p. 39] it suffices to show that, for every $\varepsilon > 0$, there exists a closed set $I_\varepsilon \subset I$ such that the graph $\gamma(G_\varepsilon^*)$ of the set-valued map G_ε^*, which is the restriction of G^* to I_ε, is a closed subset of $I_\varepsilon \times M(\Gamma)$. Indeed, since $x^* \in L_p(I, E)$ and $\dot{x}^*$, Ax^* and $\{f(\cdot, x^*(\cdot), \sigma), \sigma \in \Gamma\} \subset L_q(I, E^*)$, for every $\varepsilon > 0$, there exists a closed set $I_\varepsilon \subset I$ with Lebesgue measure $l(I \backslash I_\varepsilon) < \varepsilon$ such that on $I_\varepsilon \dot{x}^*$, Ax^* and $f(\cdot, x^*(\cdot), \sigma)$ are continuous E^*-valued functions and x^* is a continuous E-valued function. Let $(t^n, \mu^n) \in \gamma(G_\varepsilon^*)$ such that $t^n \to t^0$ and $\mu^n \xrightarrow{w} \mu^0$; we show that $(t^0, \mu^0) \in \gamma(G_\varepsilon^*)$ or equivalently, $\mu^0 \in G_\varepsilon^*(t^0)$. Since I_ε is closed, $t^0 \in I_\varepsilon$ and further, due to continuity of the restrictions

$$
\begin{aligned}
&\{\dot{x}^*(t), A(t)x^*(t), t \in I_\varepsilon\}, \\
&\langle \dot{x}^*(t^n), e\rangle \to \langle \dot{x}^*(t^0), e\rangle \quad \text{and} \quad \langle A(t^n)x^*(t^n), e\rangle \to \langle A(t^0)x^*(t^0), e\rangle
\end{aligned} \tag{10}
$$

for each $e \in E$. It remains to verify that

$$
\int_\Gamma \langle f(t^n, x^*(t^n), \sigma), e\rangle \, d\mu^n \to \int_\Gamma \langle f(t^0, x^*(t^0), \sigma), e\rangle \, d\mu^0. \tag{11}
$$

Clearly for every $e \in E$,

$$
\begin{aligned}
&\left| \int_\Gamma \langle f(t^n, x^*(t^n), \sigma), e\rangle \, d\mu^n - \int_\Gamma \langle f(t^0, x^*(t^0), \sigma), e\rangle \, d\mu^0 \right| \\
&\leqq \left| \int_\Gamma \{\langle f(t^n, x^*(t^n), \sigma), e\rangle - \langle f(t^0, x^*(t^0), \sigma), e\rangle\} \, d\mu^n \right| \\
&\quad + \left| \int_\Gamma \langle f(t^0, x^*(t^0), \sigma), e\rangle \, d\mu^n - \int_\Gamma \langle f(t^0, x^*(t^0), \sigma), e\rangle \, d\mu^0 \right|.
\end{aligned} \tag{12}
$$

Since on I_ε, x^* is a continuous E-valued function and $\{t^n, t^0\} \in I_\varepsilon$ and I_ε is closed, it follows from (F1) and the compactness of Γ that

$$
\langle f(t^n, x^*(t^n), \sigma), e\rangle_{E^*-E} \to \langle f(t^0, x^*(t^0), \sigma), e\rangle_{E^*-E} \tag{13}
$$

uniformly in σ on Γ.

Recalling that $\{\mu^n\}$ is a sequence of bounded positive (probability) measures, it follows from the above fact that, for every $\varepsilon > 0$, there exists an $n_1 = n_1(\varepsilon, e)$ such that

$$
\left| \int_\Gamma \{\langle f(t^n, x^*(t^n), \sigma), e\rangle - \langle f(t^0, x^*(t^0), \sigma), e\rangle\} \, d\mu^n \right| < \frac{\varepsilon}{2} \tag{14}
$$

for all $n > n_1$. Since $\mu^n \xrightarrow{w} \mu^0$, there exists an $n_2 = n_2(\varepsilon, e)$ such that the last term in

(12) is less than $\varepsilon/2$ for $n > n_2$. Hence for all $n > n_1 \vee n_2$ we have

$$\left| \int_\Gamma \langle f(t^n, x^*(t^n), \sigma), e \rangle \, d\mu^n - \int_\Gamma \langle f(t^0, x^*(t^0), \sigma), e \rangle \, d\mu^0 \right| < \varepsilon$$

and, $e \in E$ and $\varepsilon > 0$ being arbitrary, we conclude that (11) holds. Thus it follows from (10) and (11) that $\mu^0 \in G_\varepsilon^*(t^0)$ and hence the graph $\gamma(G_\varepsilon^*)$ is a closed subset of $I_\varepsilon \times M(\Gamma)$ and this implies that $t \to G^*(t)$ is a measurable multifunction as required to complete the proof.

As a corollary to the above theorem, we have the following result.

COROLLARY 5.1. *The relaxed system*

$$\frac{dx}{dt}(t) = A(t)x(t) + \int_\Gamma f(t, x(t), \sigma) \, d\mu_t(\sigma) \quad a.e.\ in\ I,$$

$$x(0) = x_0, \qquad \mu \in \mathcal{M},$$

is equivalent to the differential inclusion

$$\dot{x}(t) \in R_r(t, x(t)) \quad a.e.\ in\ I,$$

$$x(0) = x_0,$$

in the sense that $\phi \in C(I, E_w) \cap L_p(I, E)$ is a solution of the latter if and only if it is a solution of the former for some control $\mu \in \mathcal{M}$. Further, since $R_r = R_0$, the relaxed system (2) *and the differential inclusions* (4) *and* (5) *are all equivalent.*

With the help of the above result we can now prove our main result which states that any relaxed trajectory can be approximated as closely as desired by a trajectory of the original system (1). For this we make use of a result due to Datko [16, Lemma 7, p. 23].

Let Y be a topological space and F a measurable set-valued mapping defined on I with nonempty values $F(t) \in 2^Y$. By the integration $\int_I F(t)\, dt$ of the function F, we mean that

$$\int_I F(t) \, dt \equiv \left\{ \int_I f(t) \, dt, f \text{ any measurable selection of } F \right\}.$$

LEMMA 5.2. *Let Y be a separable reflexive Banach space and let* $\mathrm{cb}(Y)$ *denote the class of closed bounded subsets of Y. Let $K : I \to \mathrm{cb}(Y)$ and suppose there exists a $g \in L_r(I, R)$ for some $1 < r < \infty$ such that*

$$\sup \{y^*(x), x \in K(t)\} \leqq g(t) \quad a.e.\ on\ I$$

for all $y^* \in Y^*$ with $|y^*|_{Y^*} \leqq 1$. Then, for any measurable set $J \subset I$ with Lebesgue measure $l(J) \neq 0$,

$$\int_J \mathrm{Cl\,Co}\, K(t) \, dt = \mathrm{Cl} \int_J K(t) \, dt \tag{15}$$

and further the set-valued mapping $t \to \mathrm{Cl\,Co}\, K(t)$ is measurable.

This result is due to Datko [16, Lemma 7, p. 23]. We are now prepared to prove our main result.

Note. For the sake of notational convenience we shall use the symbol $A = \{A(t), t \in I\}$ to denote both the original operator and its extension.

THEOREM 5.2. *Suppose the assumptions of Theorem* 5.1 *hold, and that, for each $\xi \in E$ and $t \in I$, the set*

$$F(t, \xi) \equiv \{e^* \in E : e^* = f(t, \xi, \sigma), \sigma \in \Gamma\}$$

is a closed subset of E^*. Then the set of original trajectories X is dense in the set of relaxed trajectories X_r $(=X_0)$ with respect to the usual topology on $C(\bar{I}, H)$.

Proof. Let x be an arbitrary element of X_r, then by definition $\dot{x}(t) \in R_r(t, x(t))$ a.e. and $x(0) = x_0$. Let $y(v)$ denote an element of X corresponding to an admissible original control $v \in \mathcal{U}$. Since $x, y \in L_p(I, E)$ and $\dot{x}, \dot{y} \in L_q(I, E^*)$ and $E \subset H \subset E^*$, with E being dense in H, it follows [1, Thm. 1.2.15, p. 27] that $x, y \in C(\bar{I}, H)$.

Thus

$$\frac{d}{dt}|x(t) - y(t, v)|_H^2 = 2\langle \dot{x}(t) - \dot{y}(t, v), x(t) - y(t, v)\rangle_{E^*-E}$$

a.e. on I. Since y is a strong solution of the original system (1) we can rewrite the above equality as

$$\frac{d}{dt}|x(t) - y(t, v)|_H^2 = 2\langle \dot{x}(t) - A(t)y(t, v) - f(t, y(t, v), v(t)), x(t) - y(t, v)\rangle. \tag{16}$$

Further, being a strong solution of the relaxed system, $x(t) \in D(A(t))$ a.e. and $Ax \in L_q(I, E^*)$. Thus we can rearrange (16) as,

$$\begin{aligned}\frac{d}{dt}|x(t) - y(t, v)|_H^2 = {} & 2\langle \dot{x}(t) - A(t)x(t) - f(t, x(t), v(t)), x(t) - y(t, v)\rangle \\ & + 2\langle A(t)(x(t) - y(t, v)), x(t) - y(t, v)\rangle \\ & + 2\langle f(t, x(t), v(t)) - f(t, y(t, v), v(t)), x(t) - y(t, v)\rangle,\end{aligned}$$

from which, by virtue of (A2) and (F2), we obtain

$$\frac{d}{dt}|x(t) - y(t, v)|_H^2 \leqq 2\langle \dot{x}(t) - A(t)x(t) - f(t, x(t), v(t)), x(t) - y(t, v)\rangle \tag{17}$$

for almost all $t \in I$ and any $v \in \mathcal{U}$.

For the given x, define the multifunction K with values $K(t) \equiv F(t, x(t))$, $t \in I$. It follows from (F3) that there exists a function $g \in L_q(I, R)$ such that

$$\sup\{e^*(e), e^* \in K(t)\} \leqq g(t) \quad \text{a.e.} \tag{18}$$

for any $e \in E$ with $|e|_E \leqq 1$.

Further it follows from the hypothesis of the theorem that $K(t)$ is a closed subset of E^* for every t for which it is defined. Thus, for $Y = E^*$ and $r = q$, K satisfies the hypotheses of Lemma 5.2 and consequently, for any measurable subset $J \subset I$,

$$\int_J \mathrm{Cl\,Co}\, K(t)\, dt = \mathrm{Cl} \int_J K(t)\, dt. \tag{19}$$

By Theorem 5.1, $R_r = R_0$, and $X_r = X_0$, and hence

$$R_r(t, x(t)) = R_0(t, x(t)) = A(t)x(t) + \mathrm{Cl\,Co}\, K(t) \quad \text{a.e.}$$

and

$$\dot{x}(t) \in R_0(t, x(t)) \quad \text{a.e.}$$

Thus, $\dot{x}(t) - A(t)x(t) \in \mathrm{Cl\,Co}\, K(t)$ a.e. on I and denoting $\dot{x} - Ax$ by f^* we have $f^*(t) \in \mathrm{Cl\,Co}\, K(t)$ a.e. and that $f^* \in L_q(I, E^*)$. Since $g \in L_q(I, R)$ (see (18)), for every $\varepsilon > 0$, we can find an integer $m = m(\varepsilon)$, and partition the interval $\bar{I} = [0, T]$ into intervals $\{I_j = [t_{j-1}, t_j], j = 1, 2, \cdots, m, t_0 = 0, t_m = T\}$ of, say, equal lengths T/m such

that $\bar{I} = \cup I_j$, and

$$\int_{I_j} |g(t)|^q \, dt < \left(\frac{q\varepsilon}{2^{q+2}}\right) \tag{20}$$

for all $j \in \{1, 2, \cdots, m\}$.

Since $f^*(t) \in \operatorname{Cl} \operatorname{Co} K(t)$ a.e., it follows from (19) that, for each interval I_j, there exists a sequence of measurable functions $\{f_j^n, n = 1, 2, \cdots\}$, for each $j \in \{1, 2, \cdots, m\}$, such that

$$f_j^n(t) \in K(t) \quad \text{a.e. on } I_j$$

and

$$\int_{I_j} \langle f^*(t), e(t)\rangle_{E^*-E} \, dt = \lim_{n\to\infty} \int_{I_j} \langle f_j^n(t), e(t)\rangle_{E^*-E} \, dt \tag{21}$$

for each $e \in L_p(I, E)$ and $j \in \{1, 2, \cdots, m\}$. Define $\{f^n\}$ such that $f^n(t) = f_j^n(t)$ for $t \in I_j$ and $j \in \{1, 2, \cdots, m\}$. Clearly $f^n \in L_q(I, E^*)$. Since $f^n(t) \in K(t)$ a.e. and K is a measurable set-valued mapping with closed values, and, by hypothesis (F1), $\sigma \to \langle f(t, x(t), \sigma), e\rangle$, for $e \in E$, is continuous and Γ is compact it follows that the multifunction G_n, defined by

$$G_n(t) \equiv \{\sigma \in \Gamma : \langle f^n(t), e\rangle = \langle f(t, x(t), \sigma), e\rangle, \, e \in E\},$$

is nonempty, measurable and has closed (hence compact) values. Thus, by the selection theorem due to Kuratowski and Ryll–Nardzewski [1, Thm. 1.4.5, p. 40], G_n has a measurable selection $u^n \in \mathcal{U}$ such that $f^n(t) = f(t, x(t), u^n(t))$ a.e. on I. Integrating (17) over the interval $[0, t]$, $t \in I$, with v replaced by u^n, we have

$$|x(t) - y^n(t)|_H^2 \leqq 2 \int_0^t \langle f^*(\theta) - f^n(\theta), x(\theta) - y^n(\theta)\rangle_{E^*-E} \, d\theta \tag{22}$$

for $t \in I$ where $y^n \equiv y(u^n)$ and $f^* \equiv \dot{x} - Ax$. Since $\{y^n\} \in X$ and X is a conditionally sequentially compact subset of $C(I, E_w)$ (Lemma 5.1), there exists a subsequence of $\{y^n\}$, relabelled as $\{y^n\}$, and a $y^0 \in C(I, E_w)$ such that $y^n \to y^0$ in $C(I, E_w)$. Further, $\{y^n\}$ being a sequence of strong solutions of the system (1) all contained in $L_p(I, E)$ with limit y^0, the set $\{Ay^n, w\text{-}\lim Ay^n\}$ is contained in a bounded subset of $L_q(I, E^*)$. Hence it follows from (F3) and the evolution equation (1) that the set of distributional derivatives $\{\dot{y}^n, \dot{y}^0\}$ is contained in a bounded subset of $L_q(I, E^*)$. Therefore X is also a conditionally sequentially compact subset of $L_p(I, E)$ and there exists a subsequence of the sequence $\{y^n\}$, relabelled as $\{y^n\}$, such that $y^n \to y^0$ strongly in $L_p(I, E)$ and consequently $\eta^n \equiv (x - y^n) \to (x - y^0) \equiv \eta^0$ strongly in $L_p(I, E)$.

On the other hand it follows from (21) that $f^n \to f^*$ weakly in $L_q(I, E^*)$. Thus, for any $j \in \{1, 2, \cdots, m\}$,

$$\lim_n \int_{I_j} \langle f^* - f^n, \eta^n\rangle_{E^*-E} \, dt = 0$$

and consequently for the given $\varepsilon > 0$ there exists an integer $n^* = n^*(\varepsilon, m(\varepsilon))$ such that for $n \geqq n^*$

$$\left| \int_{I_j} \langle f^* - f^n, \eta^n\rangle \, dt \right| < \frac{\varepsilon}{4m} \tag{23}$$

for all $j \in \{1, 2, \cdots, m\}$. Since, for an arbitrary $t \in I$, there is a $j \in \{1, 2, \cdots, m\}$ such

that $t \in I_j$ we have, using (22) and (23),

$$|x(t)-y^n(t)|_H^2 \leqq \left(\frac{\varepsilon}{2}\right)+2\left|\int_{t_{j-1}}^{t}\langle f^*-f^n, \eta^n\rangle\, d\theta\right|$$

for all $n \geqq n^*$.

Therefore for any $\varepsilon > 0$, there exists an integer $n(\varepsilon) = n_0$ and hence an ordinary control $u^0 = u^{n_0} \in \mathcal{U}$ and a $y^0 = y(u^0)$ such that

$$\|x-y^0\| \equiv \sup\{|x(t)-y^0(t)|_H, t \in \bar{I}\} < \varepsilon.$$

This completes the proof of the theorem.

Since the trajectories of both the original and relaxed systems are elements of $C(\bar{I}, H)$, it makes sense to talk of attainable sets in H. For $\tau \in [0, T]$, let

$$\mathcal{A}(\tau) \equiv \{y \in H: y = x_u(\tau), u \in \mathcal{U}\}$$

denote the attainable set at time τ of the original system (1) and

$$\mathcal{A}_r(\tau) \equiv \{y \in H: y = x_\mu(\tau), \mu \in \mathcal{M}\}$$

that of the relaxed system (2). Then, as a consequence of the previous result (Theorem 5.2), we have the following:

COROLLARY 5.2. *For each* $\tau \in [0, T]$, *the set* $\mathcal{A}(\tau)$ *is strongly dense in* $\mathcal{A}_r(\tau)$.

Remark 5.1. For each $\tau \in [0, T]$, the (relaxed) attainable set $\mathcal{A}_r(\tau)$ is a weakly sequentially compact subset of $E \subset H$.

The result presented in this paper are directly useful in terminal optimization problems and time-optimal control problems [2].

6. Examples. In this section we present two examples of evolution equations to which our general theory applies. The first one is an example of a parabolic system with a strongly nonlinear perturbation term introduced by Browder with controls considered in [3], [4]. The second is an example of a general class of first order systems of linear partial differential equations, the symmetric positive systems of Friedrichs, with a nonlinear perturbation term [13, p. 79], [17, p. 73].

Example 1. *Strongly nonlinear parabolic system.* Let $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ be a multi-index with $\{\alpha_i\}$ nonnegative integers and define $|\alpha| = \sum_{i=1}^n \alpha_i$. Let $p \geqq 2$ and $q \equiv p/(p-1)$, and let $W^{m,p} \equiv W^{m,p}(\Omega)$ denote the usual Sobolev space, where Ω is an open bounded connected subset of R^n with smooth boundary $\partial\Omega$, with the usual norm

$$\|\phi\|_{W^{m,p}} = \left(\sum_{|\alpha| \leqq m} \|D^\alpha \phi\|_{L_p(\Omega)}^p\right)^{1/p}, \qquad m = 0, 1, 2, \cdots.$$

Let E be a closed subspace of $W^{m,p}$ having the structure of a reflexive Banach space so that $C_0^\infty \subset W_0^{m,p} \subset E \subset W^{m,p}$.

Let B and $\Gamma \subset B$ be as defined in § 3.

We introduce the nonlinear operators f through the following Dirichlet form:

$$b^v(t, \phi, \psi) \equiv \sum_{|\alpha| \leqq m} \langle F_\alpha(t, \cdot; \phi, D_1\phi, \cdots, D_m\phi; v), D^\alpha \psi\rangle_{L_q(\Omega)-L_p(\Omega)}$$

for $\phi, \psi \in E$ and $v \in B$, where $\langle g, h\rangle_{L_q-L_p} \equiv \int_\Omega g \cdot h\, dy$.

We assume that the functions $\{F_\alpha, |\alpha| \leqq m\}$ satisfy the following properties:

(F1)′ For each α, $F_\alpha(t, y; \eta; v)$ is measurable in $(t, y) \in I \times \Omega \equiv Q$ and continuous in the variables (t, η, v); $t \in I$, $\eta \equiv \{\eta_\alpha, |\alpha| \leqq m\}$, and $v \in B$.

(F2)′ For each $(t, y) \in Q$, and each pair $\eta(\eta') \equiv \{\eta_\alpha(\eta'_\alpha), |\alpha| \leqq m\}$,

$$\sum_{|\alpha| \leqq m} (F_\alpha(t, y; \eta; v) - F_\alpha(t, y; \eta'; v))(\eta_\alpha - \eta'_\alpha) \geqq 0$$

for all $v \in \Gamma$.

(F3)′ For a fixed $p > 1$, there exists a constant c (possibly dependent on Γ) and a function $g \in L_q(Q)(q^{-1} + p^{-1} = 1)$ such that for almost all $(t, y) \in Q$ and each $\eta \equiv \{\eta_\alpha, |\alpha| \leqq m\}$

$$|F_\alpha(t, y; \eta; v)| \leqq c \sum_{|\alpha| \leqq m} |\eta_\alpha|^{p-1} + g(t, y) \quad \text{a.e. in } Q$$

uniformly in $v \in \Gamma$.

(F4)′ There exists a positive number $c_1 > 0$ and a function $h \in L_1(Q) = L_1(I, L_1(\Omega))$ such that

$$\sum_{|\alpha| \leqq m} F_\alpha(t, y; \eta; v)\eta_\alpha \geqq c_1 \sum_{|\alpha| \leqq m} |\eta_\alpha|^p - h(t, y) \quad \text{a.e. in } Q$$

uniformly with respect to $v \in \Gamma$.

It is not difficult to verify that under the above assumptions, for each $v \in B$, the Dirichlet form b^v is bounded in $W^{m,p}$ for each $t \in I$. Further for each fixed $t \in I$, $v \in B$, and $\phi \in W_0^{m,p}$, $\psi \to b^v(t, \phi, \psi)$ is a continuous linear (antilinear in the complex case) form on $W_0^{m,p}$; hence there exists a function $\tilde{f}^v : I \times W_0^{m,p} \to W^{-m,q}$ such that $\langle \tilde{f}^v(t, \phi), \psi \rangle_{W^{-m,q} - W^{m,p}} = b^v(t, \phi, \psi)$. For the nonlinear operator f (see (1)), we may choose $E = W_0^{m,p}$ and

$$f(t, \phi, v) \equiv -\tilde{f}^v(t, \phi).$$

Under the given assumptions (F1)′–(F4)′, it is easy to verify that f, (as defined above) considered as a mapping from $I \times E \times B \to E^*$, satisfies our basic assumptions (F1)–(F4) of § 4. We consider this f as the nonlinear perturbation in (1).

For the linear operator $\{A(t), t \in I\}$, consider the bilinear form

$$a(t, \phi, \psi) \equiv \sum_{\{|\alpha|, |\beta|\} \leqq s} \langle a_{\alpha\beta}(t, \cdot) D^\beta \phi, D^\alpha \psi \rangle_{L_q - L_p},$$

where s (a positive integer) $\geqq m$, $\phi, \psi \in W^{s,p}$; and for all α, β, $a_{\alpha\beta}(t, \cdot) \in L_r(\Omega)$ with $r = p/(p-2)$, $p \geqq 2$. We may assume that the functions $\{a_{\alpha\beta}\}$ are continuous in the first variable and measurable in the second. Clearly for each $t \in I$, $a(t, \cdot, \cdot)$ is a continuous bilinear form on $W^{s,p} \times W^{s,p}$ and defines a bounded linear operator $A_0(t)$ from $W_0^{s,p}$ to $W^{-s,q}$ so that for each $t \in I$ and $\phi, \psi \in W_0^{s,p}$

$$\langle A_0(t)\phi, \psi \rangle_{W^{-s,q} - W^{s,p}} \equiv a(t, \phi, \psi), \qquad t \in I.$$

Assuming that

$$\sum_{\{|\alpha|, |\beta|\} \leqq s} a_{\alpha,\beta}(t, y)\xi^\alpha \xi^\beta \geqq 0,$$

for all $(t, y) \in Q \equiv I \times \Omega$, where $\xi^\gamma \equiv \xi_1^{\gamma_1} \cdot \xi_2^{\gamma_2} \cdots \xi_n^{\gamma_n}$, $\gamma = \alpha$ or β, it is clear that $a(t, \phi, \phi) \geqq 0$ for $\phi \in W^{s,p}$. For the operator $A = \{A(t), t \in I\}$ (see (1)) we choose $A(t) \equiv -A_0(t)$, $t \in I$ with domain

$$D(A(t)) \equiv \{\phi \in E : A_0(t)\phi \in E^*, D^\alpha \phi|_{\partial\Omega} = 0, |\alpha| \leqq s - 1\}.$$

For $s = m$, $\{A(t), t \in I\}$ is a family of bounded linear operators from E to E^* and for $s > m$ it is a family of unbounded operators with domains $D(A(t)) \subset E$ and range

$R(A(t)) \subset E^*$. We consider this operator A as the linear part in (1) and take $L_2(\Omega)$ for H. Clearly for $p \geqq 2$, $E \subset H \subset E^*$ and A satisfies the basic assumptions of § 4.

Using the operators A and f as defined above we obtain the differential equation $(\partial\psi/\partial t) = A(t)\psi + f(t, \psi, u)$, $u \in \mathcal{U}$, which is a special case of the general evolution equation (1). Note that for the operator $A(t)$, $t \in I$, one can choose any elliptic partial differential operator satisfying the basic assumptions. Also the choice of $E(W_0^{m,p} \subset E \subset W^{m,p})$ partly determines the boundary conditions.

Example 2. *First order hyperbolic systems.* We consider the quasilinear hyperbolic system of the form

$$\frac{\partial\psi}{\partial t} + \sum_{j=1}^{n} A_j(t, y)D_j\psi + B(t, y)\psi + F(t, y; \psi; u) = 0, \qquad (t, y) \in Q: \equiv I \times \Omega,$$

$$\psi(0, y) = \psi_0(y), \qquad y \in \Omega,$$

$$\psi(t, y) = 0, \qquad t \in I, \quad y \in \partial\Omega,$$

where $D_j\psi \equiv \partial\psi/\partial y_j$ with ψ an r-vector-valued function on $I \times \Omega$, Ω an open bounded set in R^n and $I = (0, T)$.

As in the previous example we proceed with the nonlinear term. F is a function defined on $I \times \Omega \times R^n \times B \to R^r$ which satisfies the following properties:

(F1)″ $F: I \times \Omega \times R^r \times B \to R^r$ is continuous.

(F2)″ For each $(t, y) \in I \times \Omega$ and each pair $\xi, \xi' \in R^r$

$$(F(t, y; \xi; v) - F(t, y; \xi'; v), \xi - \xi')_{R^r} \geqq 0 \quad \text{for all } v \in \Gamma.$$

(F3)″ There exists a function $g \in L_q(Q)$ and a constant $c \geqq 0$ such that

$$|F(t, y; \xi; v)| \leqq c|\xi|^{p-1} + g(t, y) \quad \text{a.e. in } Q$$

uniformly in $v \in \Gamma$.

(F4)″ There exists a function $h \in L_1(Q)$ and a number $c_1 > 0$ such that

$$(F(t, y; \xi; v), \xi)_{R^r} \geqq c_1|\xi|^p - h(t, y) \quad \text{a.e. in } Q$$

uniformly in $v \in \Gamma$.

We choose $E = L_p(\Omega)$, $E^* = L_q(\Omega)$ and $H = L_2(\Omega)$ with $p \geqq 2$, $p^{-1} + q^{-1} = 1$. Clearly $E \subset H \subset E^*$ since Ω is bounded. Define

$$f(t, \phi, v) \equiv -F(t, \cdot; \phi(\cdot); v(\cdot))$$

for $t \in I$, $\phi \in E$ and $v \in B$. Under the assumptions (F1)″–(F4)″ it is easy to verify that f maps $I \times E \times B$ into E^* and satisfies the basic assumptions (F1)–(F4) of § 4.

For the linear operator $\{A(t), t \in I\}$ we take the symmetric positive system of Friedrichs [13, p. 79], [17, p. 73]

$$A_0(t)\phi \equiv \sum_{j=1}^{n} A_j(t, \cdot)D_j\phi(\cdot) + B(t, \cdot)\phi(\cdot),$$

where ϕ is an r-vector-valued function on Ω and A_j, $j = 1, 2, \cdots, n$ and B are $(r \times r)$ matrix-valued functions on $Q \equiv I \times \Omega$ which are assumed to satisfy the following properties:

(A1)″ B is a matrix-valued C^0 function on $\bar{Q}$ while each A_j, $j = 1, 2, \cdots, n$, is a matrix-valued C^1 function on $\bar{Q}$.

(A2)″ The operator A_0 is hermitian nonnegative in the sense that

$$\operatorname{Re}\{(B(t, y)\xi, \xi) - \tfrac{1}{2}\textstyle\sum((D_jA_j)(t, y)\xi, \xi)\} \geqq 0$$

for all complex r-dimensional vectors ξ and $(t, y) \in Q$.

Define $\{A(t), t \in I\}$ by setting

$$D(A(t)) \equiv \{\phi \in E: A_0(t)\phi \in E^*, \phi|_{\partial\Omega} = 0\}$$

with $A(t)\phi = -A_0(t)\phi$ for $\phi \in D(A(t))$. Since $D(A(t))$ contains C_0^∞, it is dense in E. Combining these we obtain the evolution equation $(\partial\psi/\partial t) = A(t)\psi + f(t, \psi, u)$, $u \in \mathcal{U}$ as a special case of (1).

Remark 6.1. It is clear from the above examples that the choice of the state space E depends, among other factors, on the form of the abstract function $x \to f(t, x, v)$. Similarly the choice of B (the space where the controls take their values) depends on the mapping $v \to f(t, x, v)$. For example, if the functions F_α (Example 1) and F (Example 2) depend on v along with its spatial derivatives of order up to k $(0 \leqq k < \infty)$ we may choose for B a Sobolev space $W^{k,s}(\Omega, R^d) \equiv \{v \in L_s(\Omega, R^d): \|D^\alpha v\|_{L_s} < \infty, |\alpha| \leqq k\}$ for a suitable $1 \leqq s < \infty$. Clearly these are Polish spaces. We may also choose for B any closed bounded convex subset of $W_0^{k,s}$, $1 < s < \infty$, furnished with the weak topology. With respect to the weak topology B is a compact Hausdorff space and, since the dual $W^{-k,s'}(1/s + 1/s' = 1)$ of $W_0^{k,s}$ is separable, this topology is metrizable [14, Thm. V.5.2, p. 426] and hence B is a Polish space. If F_α and F do not depend on the spatial derivatives of v, one may then choose for B a closed bounded convex subset of $L_\infty(\Omega, R^d)$. In this case B is w^*-compact and, since L_1 is separable, this topology is also metrizable [14, Thm. V.5.1, p. 426] and hence B is a Polish space. In the case of $L_1(\Omega, R^d)$, we can choose a weakly compact subset $B \subset L_1(\Omega, R^d)$. Since the weak topology is metrizable [14, Thm. V.6.3] B is again a Polish space. In fact we can choose for B any separable Fréchet space since it is Polish. Thus there are many choices within the framework of Polish spaces mainly determined by the function f and specific control constraints.

Acknowledgment. The author would like to thank Professor Berkovitz and the anonymous reviewers for valuable comments that led to substantial improvement of the results of the paper.

REFERENCES

[1] N. U. AHMED AND K. L. TEO, *Optimal Control of Distributed Parameter Systems*, North-Holland, New York, Oxford, 1981.

[2] ———, *Optimal control of systems governed by a class of nonlinear evolution equations in a reflexive Banach space*, J. Optim. Theory Appl., 25 (1978), pp. 57–81.

[3] N. U. AHMED, *Optimal control of a class of strongly nonlinear parabolic systems*, J. Math. Anal. Appl., 16 (1977), pp. 188–207.

[4] ———, *Some comments on optimal control of a class of strongly nonlinear parabolic systems*, J. Math. Anal. Appl., 68 (1979), pp. 595–598.

[5] L. CESARI, *Geometric and analytic views in existence theorems for optimal control in Banach spaces*, I: *Distributed controls*, J. Optim. Theory Appl., 14 (1974), pp. 505–520.

[6] ———, *Geometric and analytic views in existence theorems for optimal control in Banach spaces*, II: *Distributed and boundary controls*, J. Optim. Theory Appl., 15 (1975), pp. 467–497.

[7] ———, *Geometric and analytic views in existence theorems for optimal control in Banach spaces*, III: *Weak solution*, J. Optim. Theory Appl., 19 (1976), pp. 185–214.

[8] LEONARD D. BERKOVITZ, *Existence and lower closure theorems for abstract control problems*, this Journal, 12 (1974), pp. 27–42.

[9] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[10] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York and London, 1972.

[11] R. V. GAMKRELIDZE, *Principles of Optimal Control Theory*, Plenum Press, New York and London, 1978.
[12] T. PARATHASARATHY, *Probabilities on Metric Space*, Academic Press, New York and London, 1967.
[13] F. E. BROWDER, *Nonlinear initial value problems*, Ann. Math., 81 (1965), pp. 51–87.
[14] N. DUNFORD AND J. T. SCHWARZ, *Linear Operators, Part* I, John Wiley, New York, 1958.
[15] C. J. HIMMELBERG, M. Q. JACOBS AND E. S. VAN VLECK, *Measurable multifunctions, selectors, Filippov's implicit functions lemma*, J. Math. Anal. Appl., 25 (1969), pp. 276–284.
[16] RICHARD DATKO, *Measurability properties of set-valued mappings in a Banach space*, this Journal, 8 (1970), pp. 226–238.
[17] H. TANABE, *Equations of Evolution*, Pitman, London, San Francisco, Melbourne, 1975.
[18] N. U. AHMED, *Relaxed controls in infinite dimensional systems*, Proc. IEEE-CDC (1982), pp. 1073–1077.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

0363-0129/83/2106-0010 $01.25/0

BOUNDARY STABILIZATION OF LINEAR ELASTODYNAMIC SYSTEMS*

JOHN LAGNESE†

Abstract. It is proved that an elastic medium which occupies a bounded region in three dimensional Euclidean space can be uniformly stabilized by means of traction forces applied on a portion of the boundary of the medium (the remaining portion being clamped) provided the geometry of the boundary is suitably restricted. It is also shown that in the absence of such restrictions the medium can still be strongly stabilized by such traction forces.

Key words. elastodynamic systems, boundary control, uniform stabilization, strong stabilization

1. Introduction and statement of results. The main purpose of this paper is to show that an elastic medium which occupies a bounded region Ω in three dimensional Euclidean space can be uniformly stabilized by means of traction forces applied on a portion of the boundary of the medium, provided the geometry of the boundary is suitably restricted. However, we will also show that in the absence of such restrictions the medium can still be strongly stabilized by such traction forces.

If $u=(u_1, u_2, u_3)$ denotes the coordinates of the displacement at time t of the particle which in the nondeformed state has coordinates $x=(x_1, x_2, x_3)$, then if the displacements are small the governing dynamical system is

$$\rho(x)u_{i,tt}-\sigma_{ij,j}+q(x)u_i=0, \qquad x\in\Omega, \quad t>0, \quad i=1,2,3. \tag{1.1}$$

Subscripts following a comma denote differentiation, e.g., $u_{i,t}=\partial u_i/\partial t$, $\sigma_{ij,k}=\partial\sigma_{ij}/\partial x_k$, and the summation convention is used throughout. $\rho(x)>0$ is the local density of the medium and $q(x)u_i(q(x)\geqq 0)$ represents a restoring force proportional to the displacement u_i. σ_{ij}, the stress tensor, is related to the strain tensor

$$\varepsilon_{kl}=\tfrac{1}{2}(u_{k,l}+u_{l,k})$$

by the relation

$$\sigma_{ij}=a_{ijkl}\varepsilon_{kl}.$$

The a_{ijkl} are the coefficients of elasticity. These will depend on x when the material is inhomogeneous, but we shall assume that they do not depend on time. They have the symmetry properties

$$a_{ijkl}=a_{klij}=a_{jilk}. \tag{1.2}$$

These coefficients, as well as $\rho(x)$ and $q(x)$, are assumed to be of class $C^1(\bar{\Omega})$. If we set

$$c_{ijkl}=\tfrac{1}{2}(a_{ijkl}+a_{ijlk})$$

then (1.1) may be written

$$\rho(x)u_{i,tt}-\frac{\partial}{\partial x_j}(c_{ijkl}u_{k,l})+q(x)u_i=0.$$

The c_{ijkl} clearly have the same symmetry properties as the a_{ijkl} and we suppose they satisfy an ellipticity condition

$$c_{ijkl}\xi_{ij}\xi_{kl}>c_0\xi_{ij}\xi_{ij} \tag{1.3}$$

* Received by the editors February 12, 1982, and in revised form October 25, 1982.
† Department of Mathematics, Georgetown University, Washington, D.C. 20057.

for some $c_0 > 0$, all $x \in \bar{\Omega}$ and all real second order tensors ξ_{ij}. It follows at once that

$$a_{ijkl}\xi_{ij}\xi_{kl} \geqq c_0 \xi_{ij}\xi_{ij} \tag{1.4}$$

for all *symmetric* second order tensors ξ_{ij}.

The boundary $\partial\Omega$ of Ω is assumed to be class C^2 and to consist of two disjoint parts, Γ_0 and Γ_1, with $\Gamma_1 \neq \varnothing$ and relatively open in $\partial\Omega$. Γ_0 will be assumed to be either empty or to have a nonempty interior, and the elastic medium is assumed to be fixed there, that is,

$$u_i = 0, \qquad x \in \Gamma_0, \quad t > 0, \quad i = 1, 2, 3. \tag{1.5}$$

On Γ_1 traction forces are specified:

$$\sigma_{ij} n_j = F_i, \qquad x \in \Gamma_1, \quad t > 0, \quad i = 1, 2, 3, \tag{1.6}$$

where $n = (n_1, n_2, n_3)$ is the unit normal vector pointing into the exterior of Ω. It is by means of these forces that the system (1.1), (1.5), (1.6) is to be stabilized.

Associated with smooth solutions of (1.1), (1.5), (1.6) is the *energy functional*

$$E(u, t) = \tfrac{1}{2} \int_\Omega [\rho u_{i,t} u_{i,t} + q u_i u_i + \sigma_{ij} \varepsilon_{ij}] \, dx$$

$$= \tfrac{1}{2} \int_\Omega [\rho u_{i,t} u_{i,t} + q u_i u_i + a_{ijkl} \varepsilon_{kl} \varepsilon_{ij}] \, dx$$

in which the integrand is evaluated at time t. A short calculation shows that

$$\frac{d}{dt} E(u, t) = \int_\Omega \frac{\partial}{\partial x_j} (u_{i,t} \sigma_{ij}) \, dx = \int_{\Gamma_1} u_{i,t} F_i \, d\sigma.$$

Consequently, the linear feedback law

$$F = -b \frac{\partial u}{\partial t}, \tag{1.7}$$

where b is a nonnegative function, stabilizes the system in the sense that the energy is nonincreasing. (One could just as well choose the slightly more general feedback law $F_i = -b_i u_{i,t}$, $i = 1, 2, 3$, $\forall b_i \geqq 0$. All of our results extend to this situation with only trivial modifications in statements and proofs.) Our main result is the following.

THEOREM 1.1. *Assume that*

$$b \in C^1(\bar{\Gamma}_1), \qquad b(x) \geqq b_0 > 0 \quad \text{on } \Gamma_1, \tag{1.8}$$

and that there is a vector field $f = (f_1, f_2, f_3)$ of class $C^2(\bar{\Omega})$ such that

$$f \cdot n \leqq 0 \text{ on } \Gamma_0, \tag{1.9}$$

$$f \cdot n > 0 \text{ on } \bar{\Gamma}_1, \tag{1.10}$$

$$f \cdot \nabla \rho \geqq \alpha \rho \text{ in } \Omega \text{ for some } \alpha > -1, \tag{1.11}$$

$$\Lambda(\xi) \geqq (2 + \alpha) c_{ijkl} \xi_{ij} \xi_{kl} \text{ in } \Omega \text{ for all second order tensors } \xi = [\xi_{ij}], \text{ where} \tag{1.12}$$

$$\Lambda(\xi) = (2 f_{j,m} c_{imkl} - f_m c_{ijkl,m}) \xi_{ij} \xi_{kl}.$$

Then there are positive constants C, δ, such that

$$E(u, t) \leqq C e^{-\delta t} E(u, 0), \qquad t \geqq 0, \tag{1.13}$$

for every solution of (1.1), (1.5), (1.6), (1.7) *for which* $E(u, 0) < \infty$.

The condition $E(u, 0) < +\infty$ means that $u_i(\cdot, 0) \in H^1(\Omega)$, $u_{i,t}(\cdot, 0) \in L^2(\Omega)$, and $u_i(\cdot, 0) = 0$ on Γ_0 in the sense of traces if $\Gamma_0 \neq \emptyset$, $i = 1, 2, 3$. In § 2 it will be proved that for such initial data the problem (1.1), (1.5), (1.6), (1.7), has a unique weak solution which satisfies $E(u, \cdot) \in C([0, \infty))$.

Conditions (1.9), (1.10) mean that flows of f (i.e., solutions of $\dot{x} = f(x)$) which begin in $\bar{\Omega}$ remain in $\bar{\Omega}$. Conditions (1.11) and (1.12) comprise a nontrapping hypothesis relative to these flows, and may be interpreted as meaning that the distance between flows (measured in a suitable metric depending on the coefficients of (1.1)) increases with time, and thus excludes regions which would cause such flows to be "pinched" together. We refer to [11, § 4] for a complete discussion of related geometric conditions associated with the ordinary wave operator.

The condition (1.10) does not seem necessary from a physical standpoint and it would be useful to eliminate it since (1.9) and (1.10) together force $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \emptyset$ if $\partial\Omega$ is to be as smooth as required above. On the other hand, both this degree of smoothness and the condition $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \emptyset$ are needed in the proof to assure that (1.1), (1.5), (1.6), (1.7) has classical pointwise solutions, as it is known that singularities in solutions can occur only at corners of $\partial\Omega$ or at points of $\bar{\Gamma}_0 \cap \bar{\Gamma}_1$. However, there are specific geometries in which $\partial\Omega$ is only piecewise smooth and also $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 \neq \emptyset$, yet such singularities do not occur (at least for certain specific elastodynamic systems). Our results would then also apply to such special situations still assuming, of course, (1.8)–(1.12).

A simple situation in which all hypotheses of Theorem 1.1 are met is the following. Let $\Omega = \Omega_1 - \bar{\Omega}_0$, where Ω_0 and Ω_1 are bounded regions in R^3 with C^2 boundaries Γ_0 and Γ_1, respectively. Assume that (i) $\bar{\Omega}_0 \subset \Omega_1$, (ii) Ω_0 is starshaped and Ω_1 strongly starshaped with respect to a common point $x_0 \in \Omega_0$, and (iii) the coefficients ρ and a_{ijkl} in (1.1) are constants. Conditions (1.9)–(1.12) will then hold with $f(x) = x - x_0$.

One may completely eliminate the geometric restrictions of Theorem 1.1 and also weaken condition (1.8) and still obtain the following strong stabilization result.

THEOREM 1.2. *Assume that* $b \in C^1(\bar{\Gamma}_1)$, $b \geqq 0$, $b \not\equiv 0$. *If* u *is a solution of* (1.1), (1.5), (1.6), (1.7) *with* $E(u, 0) < +\infty$, *then*

$$E(u, \infty) \doteq \lim_{t \to \infty} E(u, t) = 0.$$

When $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \emptyset$, this last result can be deduced from the general stabilization results of Benchimol [2] (cf. Slemrod [10]). However, these results imply only *weak stabilization* in the general case, again because of the possible occurence of singular solutions, although it may be possible to augment the arguments in [2] to obtain Theorem 1.2. However, we will use a totally different approach based on Proposition 3.1 below, a result which also plays a key role in the proof of Theorem 1.1. Results analogous to Theorem 1.1 were obtained in [7] for solutions of scalar wave equations in R^n, and for solutions of the ordinary wave in [3], [4], [8]. The proof presented here follows a line of reasoning similar to that in [7], but also requires ad hoc arguments specific to the system under consideration and substantially different from those in [7].

Finally, we note that uniform stabilization of evolutionary systems such as the one considered here is closely related to *exact controllability* of solutions of such systems, and such controllability results could easily be deduced from Theorem 1.1 using a technique originally devised by D. L. Russell [9]. Because the argument is now standard (cf. [3]) we will not pursue further the exact controllability question.

Theorems 1.1 and 1.2 are proved in § 3. In the next section an existence and regularity theory will be developed for solutions of (1.1), (1.5), (1.6) with F given by the feedback law (1.7).

2. Existence, uniqueness, and regularity of solutions. We consider the problem

$$\rho u_{i,tt} - \sigma_{ij,j} + q u_i = 0, \qquad x \in \Omega, \quad t > 0, \quad i = 1, 2, 3, \tag{2.1}$$

$$u = 0, \quad x \in \Gamma_0, \quad t > 0, \tag{2.2}$$

$$\sigma_{ij} n_j + b u_{i,t} = 0, \qquad x \in \Gamma_1, \quad t > 0, \quad i = 1, 2, 3, \tag{2.3}$$

$$u(x, 0) = u_0(x), \qquad u_t(x, 0) = v_0(x), \quad x \in \Omega. \tag{2.4}$$

The coefficients have the smoothness assumed in § 1 and satisfy

$$\rho(x) \geqq \rho_0 > 0, \quad q(x) \geqq 0, \quad b(x) \geqq 0,$$

and the a_{ijkl} possess the symmetry and ellipticity properties (1.2) and (1.4).

In order to formulate the definition of a solution to (2.1)–(2.4) we introduce the following real Hilbert spaces. For $m \geqq 1$ an integer,

$$H = (L^2(\Omega))^3, \quad V_{\Gamma_0}^m = (H_{\Gamma_0}^m(\Omega))^3, \quad \Sigma_{\Gamma_1} = (L^2(\Gamma_1))^3,$$

where $H_{\Gamma_0}^m(\Omega)$ consists of real functions in $H^m(\Omega)$ which vanish on Γ_0. The scalar product and norm in H are denoted by $(\cdot, \cdot)_{\Gamma_1}$ and $|\cdot|$, respectively, and those in Σ_{Γ_1} by $(\cdot, \cdot)_{\Gamma_1}$ and $|\cdot|_{\Gamma_1}$. The norms of u in $H_{\Gamma_0}^m(\Omega)$, $L^2(\Omega)$ and $L^2(\Gamma_1)$ are

$$\left(\sum_{|\alpha| \leqq m} \int_\Omega |D^\alpha u|^2 \, dx \right)^{1/2}, \quad \left(\int_\Omega |u|^2 \rho \, dx \right)^{1/2}, \quad \left(\int_{\Gamma_1} |u|^2 \, d\sigma \right)^{1/2},$$

respectively. The norm on $V_{\Gamma_0}^m$ will be denoted by $\|\cdot\|_m$. When $m = 1$ we shall write V_{Γ_0} and $\|\cdot\|$, respectively, in place of $V_{\Gamma_0}^1$ and $\|\cdot\|_1$. Finally we introduce the "finite energy" space E defined as follows: $E = V_{\Gamma_0} \times H$ algebraically, and we endow E with the inner product

$$((u^{(1)}, u^{(2)}), (v^{(1)}, v^{(2)}))_E = \int_\Omega (c_{ijkl} u_{k,l}^{(1)} v_{i,j}^{(1)} + q u^{(1)} \cdot v^{(1)} + \rho u^{(2)} \cdot v^{(2)}) \, dx.$$

(Strictly speaking, $(\cdot, \cdot)_E$ is only a semi-inner product if $\Gamma_0 = \varnothing$ and $q(x) \equiv 0$, in which case E is defined to be the space of equivalence classes of $V_{\Gamma_0} \times H$ modulo zero energy states $u^{(1)} = a + b \wedge x$, $u^{(2)} = 0$.) E is a Hilbert space and, if $\Gamma_0 \neq \varnothing$ or if $q(x) > 0$ on a set in Ω of positive measure, it follows from Korn's inequality that E is topologically equivalent to $V_{\Gamma_0} \times H$ endowed with the usual product norm.

The problem (2.1)–(2.4) will be treated as an initial value problem for an evolution equation in E by setting $u^{(1)} = u$, $u^{(2)} = u_t$. Then (2.1), (2.4) can be written

$$U_t - AU = 0, \qquad t > 0, \tag{2.5}$$

$$U(0) = U_0, \tag{2.6}$$

where $U = (u^{(1)}, u^{(2)})'$, $U_0 = (u_0, v_0)'$, ($'$ denotes the transpose of a vector) and A is the linear operator in E defined by

$$D(A) = \{(u^{(1)}, u^{(2)}) \in E \mid u^{(1)} \in V_{\Gamma_0}^2,$$
$$u^{(2)} \in V_{\Gamma_0}, \, c_{ijkl} u_{k,l}^{(1)} n_j = -b u_i^{(2)} \text{ on } \Gamma_1, \, i = 1, 2, 3\},$$

$$AU = \begin{pmatrix} 0 & I \\ (1/\rho)\mathscr{A} & 0 \end{pmatrix} U, \qquad U \in D(A),$$

and

$$\mathscr{A} u^{(1)} = \left(\frac{\partial}{\partial x_j} (c_{ijkl} u_{k,l}^{(1)}) - q u_i^{(1)} \right)_{i=1}^3.$$

A short calculation shows that if $U \in D(A)$ and $V \in V_{\Gamma_0} \times V_{\Gamma_0}$,

$$(AU, V)_E = \int_\Omega [c_{ijkl}(u^{(2)}_{k,l} v^{(1)}_{i,j} - u^{(1)}_{k,l} v^{(2)}_{i,j}) + q(u^{(2)} \cdot v^{(1)} - u^{(1)} \cdot v^{(2)})]\, dx$$

$$- \int_{\Gamma_1} b u^{(2)} \cdot v^{(2)}\, d\sigma \tag{2.7}$$

and therefore A is *dissipative*, that is,

$$(AU, U)_E \leqq 0, \qquad U \in D(A).$$

If it were also true that $\mathrm{Re}\,(I - A) = E$, it would follow from the Lumer–Phillips theorem that A is the infinitesimal generator of a strongly continuous semigroup e^{tA}, $t \geqq 0$, of contractions on E, and the unique weak solution of (2.5), (2.6) would then be given by $e^{tA} U_0$, $U_0 \in E$. This is certainly the case when $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \varnothing$ but not so when $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 \neq \varnothing$, since then the problem $(I - A)U = F \in E$ can have weak solutions U which are not in $D(A)$ (singularities can occur at points of $\bar{\Gamma}_0 \cap \bar{\Gamma}_1$). However, we will prove the following.

PROPOSITION 2.1. *A has a dissipative extension $\bar{A}$ with $D(\bar{A}) \subset V_{\Gamma_0} \times V_{\Gamma_0}$ and* $\mathrm{Rg}\,(I - \bar{A}) = E$.

Proof. Let $\mathcal{H} = V_{\Gamma_0} \times V_{\Gamma_0}$, and $\mathcal{H}'$ denote the dual of $\mathcal{H}$ relative to E, so that $\mathcal{H} \subset E \subset \mathcal{H}'$ with each space dense in the one which follows it and with continuous injections. For U, V in $\mathcal{H}$, we define $B(U, V)$ as the right side of (2.7):

$$B(U, V) = \int_\Omega [c_{ijkl}(u^{(2)}_{k,l} v^{(1)}_{i,j} - u^{(1)}_{k,l} v^{(2)}_{i,j}) + q(u^{(2)} \cdot v^{(1)} - u^{(1)} \cdot v^{(2)}]\, dx$$

$$- \int_{\Gamma_1} b u^{(2)} \cdot v^{(2)}\, d\sigma.$$

$B(U, V)$ is a continuous bilinear form on $\mathcal{H}$, so there is an operator $\tilde{A} \in \mathcal{L}(\mathcal{H}, \mathcal{H}')$ such that

$$B(U, V) = \langle \tilde{A} U, V \rangle$$

for all U, V in $\mathcal{H}$, where $\langle F, G \rangle$ denotes the scalar product of $F \in \mathcal{H}'$ and $G \in \mathcal{H}$ in the $\mathcal{H}' - \mathcal{H}$ duality. We now define an operator $\bar{A}$ as follows:

$$D(\bar{A}) = \{U \in \mathcal{H} \mid \tilde{A} U \in E\}, \qquad \bar{A} U = \tilde{A} U \quad \text{for } U \in D(\bar{A}).$$

Then

$$B(U, V) = (\bar{A} U, V)_E$$

for all $U \in D(\bar{A})$, $V \in \mathcal{H}$, so $\bar{A}$ is a dissipative extension of A. We now show that $\mathrm{Rg}\,(I - \bar{A}) = E$.

To do so, let $F = (f^{(1)}, f^{(2)}) \in E$ and consider first the equation $(I - A)U = F$. This is equivalent to

$$u^{(1)} \in V^2_{\Gamma_0}, \qquad u^{(2)} \in V_{\Gamma_0},$$

$$u^{(1)} - u^{(2)} = f^{(1)}, \qquad \rho u^{(2)} - \mathcal{A} u^{(1)} = \rho f^{(2)},$$

$$c_{ijkl} u^{(1)}_{k,l} n_j = -b u^{(2)}_i \quad \text{on } \Gamma_1, \quad i = 1, 2, 3.$$

This system is easily uncoupled and yields the following problem for $u^{(1)}$:

$$u^{(1)} \in V^2_{\Gamma_0}, \quad \rho u^{(1)} - \mathscr{A} u^{(1)} = \rho(f^{(1)} + f^{(2)}),$$
$$c_{ijkl} u^{(1)}_{k,l} + b u^{(1)}_i = b f^{(1)}_i \quad \text{on } \Gamma_1, \quad i = 1, 2, 3.$$

Let $\psi \in V^2_{\Gamma_0}$ be chosen so that

$$c_{ijkl} \psi_{k,l} n_j + b\psi_i = b f^{(1)}_i \quad \text{on } \Gamma_1$$

and set $w = u^{(1)} - \psi$. The problem for w is then

$$w \in V^2_{\Gamma_0},$$
$$\rho w - \mathscr{A} w = \rho(f^{(1)} + f^{(2)}) - (\rho\psi - \mathscr{A}\psi) \doteq g,$$
$$c_{ijkl} w_{k,l} n_j + b w_i = 0 \quad \text{on } \Gamma_1, \quad i = 1, 2, 3.$$

A variational formulation of this problem is

$$w \in V_{\Gamma_0},$$

$$\int_\Omega (c_{ijkl} w_{k,l} v_{i,j} + \rho w \cdot v + q w \cdot v)\, dx + \int_{\Gamma_1} b w \cdot v \, d\sigma = \int_\Omega g \cdot v \, dx \quad \forall v \in V_{\Gamma_0}. \tag{2.8}$$

Because of the ellipticity condition (1.3) and since $b \geqq 0$ on Γ_1, we can use the Lax–Milgram theorem to conclude that (2.8) has a unique solution in V_{Γ_0}. Then setting $u^{(1)} = w + \psi$, $u^{(2)} = u^{(1)} - f^{(1)}$, we obtain

$$u^{(1)} \in V_{\Gamma_0}, \quad u^{(2)} \in V_{\Gamma_0}, \quad u^{(2)} = u^{(1)} - f^{(1)}, \tag{2.9}$$

$$\int_\Omega (c_{ijkl} u^{(1)}_{k,l} v_{i,j} + \rho u^{(2)} \cdot v + q u^{(1)} \cdot v)\, dx + \int_{\Gamma_1} b u^{(2)} \cdot v \, d\sigma = \int_\Omega \rho f^{(2)} \cdot v \, dx \quad \forall v \in V_{\Gamma_0}. \tag{2.10}$$

From (2.9) we have

$$\int_\Omega (c_{ijkl} u^{(1)}_{k,l} v^{(1)}_{i,j} + q u^{(1)} \cdot v^{(1)})\, dx - \int_\Omega (c_{ijkl} u^{(2)}_{k,l} v^{(1)}_{i,j} + q u^{(2)} \cdot v^{(1)})\, dx$$
$$= \int_\Omega (c_{ijkl} f^{(1)}_{k,l} v^{(1)}_{i,j} + q f^{(1)} \cdot v^{(1)})\, dx.$$

Setting $v = v^{(2)}$ in (2.10) and then adding to it the last expression yields

$$(U, V)_E - B(U, V) = (F, V)_E \quad \forall V \in \mathscr{H},$$

that is, $(I - \bar{A})U = F$.

We can now easily deduce the following existence, uniqueness and regularity results for (2.1)–(2.4).

THEOREM 2.1 (*existence-uniqueness*). *Assume that $u_0 \in V_{\Gamma_0}$, $v_0 \in H$. Then there is a unique function u such that*

$$u \in C([0, \infty); V_{\Gamma_0}), \qquad u(0) = u_0,$$
$$u_t \in C([0, \infty); H), \qquad u_t(0) = v_0,$$
$$\sqrt{b}\, \gamma_0 u \in H^1(0, T; \Sigma_{\Gamma_1}),$$

and

$$\int_0^T \int_\Omega [c_{ijkl} u_{k,l} \phi_{i,j} + qu \cdot \phi - \rho u_t \cdot \phi_t]\, dx\, dt + \int_0^T \int_{\Gamma_1} b(\gamma_0 u)_t \cdot (\gamma_0 \phi)\, d\sigma\, dt = 0 \tag{2.11}$$

for every $T > 0$ *and every* ϕ *in the space*

$$D_T = \{\phi \mid \phi \in L^2((0, T); V_{\Gamma_0}),\ \phi_t \in L^2((0, T); H),\ \phi(0) = \phi(T) = 0\}.$$

Remark. $\gamma_0 u$ denotes the trace of u on $\partial\Omega$.

THEOREM 2.2 (*regularity*). *Assume that* $u_0 \in V^2_{\Gamma_0}$, $v_0 \in V_{\Gamma_0}$ *satisfy the compatibility condition*

$$c_{ijkl}(u_0)_{k,l} n_j + b(v_0)_i = 0 \quad \text{on } \Gamma_1, \quad i = 1, 2, 3.$$

Then the solution to (2.1)–(2.4) *satisfies*

$$u_t \in C([0, \infty); V_{\Gamma_0}), \tag{2.12}$$

$$u_{tt} \in C([0, \infty); H), \tag{2.13}$$

$$\sqrt{b}\, \gamma_0 u \in H^2((0, T); \Sigma_{\Gamma_1}) \tag{2.14}$$

and

$$\int_\Omega [c_{ijkl} u_{k,l} v_{i,j} + (\rho u_{tt} + qu) \cdot v]\, dx + \int_{\Gamma_1} b(\gamma_0 u)_t \cdot (\gamma_0 v)\, d\sigma = 0 \tag{2.15}$$

for every $v \in V_{\Gamma_0}$. *If, in addition,* $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \varnothing$, *then*

$$u \in C([0, \infty); V^2_{\Gamma_0}) \tag{2.16}$$

and satisfies (2.1)–(2.4) *in the pointwise sense.*

Proof of Theorem 2.1. $\bar{A}$ is the generator of a strongly continuous semigroup $e^{t\bar{A}}$, $t \geqq 0$, of contractions on E. Suppose $U_0 = (u_0, v_0)' \in D(\bar{A})$. Then for $t \geqq 0$, $U(t) = e^{t\bar{A}} U_0$ belongs to $D(\bar{A})$, is strongly continuously differentiable in E and is the unique solution to $U(0) = U_0$,

$$(U'(t), V)_E - B(U(t), V) = 0 \quad \forall V \in \mathscr{H}, \quad t \geqq 0. \tag{2.17}$$

Setting $V = (v, 0)$, $v \in V_{\Gamma_0}$, we obtain

$$\int_\Omega \{c_{ijkl}[(u_t^{(1)})_{k,l} - u^{(2)}_{k,l}] v_{i,j} + q(u_t^{(1)} - u^{(2)}) \cdot v\}\, dx = 0$$

and so $u^{(2)} = u_t^{(1)}$. Then setting $V = (0, \phi)$ with $\phi \in D_T$ we obtain (2.11) after an integration by parts in t, where $u = u^{(1)}$. Also, setting $V = U$ in (2.17) and integrating in t we obtain

$$\tfrac{1}{2}|U(t)|_E^2 - \int_0^t B(U(t), U(t))\, dt = \tfrac{1}{2}|U_0|_E^2,$$

or

$$\tfrac{1}{2}|U(t)|_E^2 + \int_0^t \int_{\Gamma_1} b|(\gamma_0 u)_t|^2\, d\sigma\, dt = \tfrac{1}{2}|U_0|_E^2.$$

Thus $\sqrt{b}\gamma_0 u \in H^1((0, T); \Sigma_{\Gamma_1})$. This proves Theorem 2.1 if $U_0 \in D(\bar{A})$, and the case $U_0 \in E$ is now easily handled by approximating U_0 with elements from $D(\bar{A})$.

Proof of Theorem 2.2. The assumptions on u_0, v_0 mean that $U_0 \in D(A) \subset D(\bar{A})$, and so we have $U_t(t) = e^{t\bar{A}}(AU_0)$, that is, $W = U_t$ is the unique weak solution of $W_t - \bar{A}W = 0$, $W(0) = AU_0$. Equations (2.12)–(2.15) then follow from Theorem 2.1. If also $\bar{\Gamma}_0 \cap \bar{\Gamma}_1 = \varnothing$, then $\bar{A} = A$ so that $U(t) \in D(A)$, $t \geqq 0$ and $AU \in C([0, \infty); E)$. Equation (2.16) then follows from standard a priori estimates for solutions of elliptic boundary value problems [1].

3. Decay of solutions. In this section Theorems 1.1 and 1.2 will be proved. The proofs will be given under the assumption that either $\Gamma_0 \neq \varnothing$ or that meas $\{x \in \Omega \mid q(x) > 0\} > 0$. The argument given below has to be modified slightly along the lines of [7] to handle the case $\Gamma_0 = \varnothing$ and $q(x) = 0$, due to the fact that (2.1)–(2.3) then admits nontrivial solutions with zero energy.

The proofs of both Theorems 1.1 and 1.2 are based on the following result.

PROPOSITION 3.1. *If $b \geqq 0$ on Γ_1, $b \not\equiv 0$, then for every $\varepsilon > 0$ there is a number C_ε such that for every $\beta > 0$ and for every solution of* (2.1)–(2.4) *for which $E(u, 0) < +\infty$,*

$$\int_0^\infty \int_\Omega e^{-2\beta t} u_i u_i \, dx \, dt \leqq C_\varepsilon E(u, 0) + \varepsilon \int_0^\infty \int_\Omega e^{-2\beta t} u_{i,t} u_{i,t} \, dx \, dt. \tag{3.1}$$

This proposition will be proved at the end of this section.

Proof of Theorem 1.1. We first suppose that $E(u_t, 0) < +\infty$, that is, $(u(\cdot, 0), u_t(\cdot, 0)) \in D(A)$. According to Theorem 2.2, u then satisfies (1.1), (1.5)–(1.7) in the pointwise sense on $\Omega \times (0, \infty)$.

Our starting point is the identity

$$0 = (\rho u_{i,tt} - \sigma_{ij,j} + q u_i)(2t u_{i,t} + 2(f \cdot \nabla u_i) + (f_{m,m} - 1)u_i).$$

It can be verified that this is equivalent to

$$\frac{\partial}{\partial t} Q_1(x, t) = \frac{\partial}{\partial x_j} R_j(x, t) + Q_2(x, t), \tag{3.2}$$

where

$$Q_1 = t(\rho u_{i,t} u_{i,t} + \sigma_{ij}\varepsilon_{ij} + q u_i u_i) + 2(f \cdot \nabla u_i)\rho u_{i,t} + (f_{m,m} - 1)\rho u_i u_{i,t},$$

$$Q_2 = -(f \cdot \nabla \rho) u_{i,t} u_{i,t} + (f \cdot \nabla q) u_i u_i + 2q u_i u_i - \sigma_{ij} u_i f_{m,mj}$$

$$-[2\sigma_{ij} u_{i,m} f_{m,j} - 2\sigma_{ij}\varepsilon_{ij} - a_{ijkl,m}\varepsilon_{ij}\varepsilon_{kl} f_m],$$

and

$$R_j = 2t u_{i,t}\sigma_{ij} + 2(f \cdot \nabla u_i)\sigma_{ij} + (f_{m,m} - 1)\sigma_{ij} u_i + \rho f_j u_{i,t} u_{i,t} - \sigma_{im}\varepsilon_{im} f_j - q u_i u f_j.$$

We define the functional

$$Q(t) = (\alpha - 1) t E(u, t) + \int_\Omega Q_1(x, t) \, dx.$$

Then from (3.5)

$$\dot{Q}(t) = (\alpha - 1)E(u, t) + (\alpha - 1)t \frac{d}{dt} E(u, t) + \int_\Omega [R_{j,j}(x, t) + Q_2(x, t)] \, dx. \tag{3.3}$$

We shall prove below that there are positive numbers C, δ and t_0, independent of u,

such that

$$\dot{Q}(t) \leqq -\delta E(u,t) + C\int_\Omega u_i u_i\,dx, \qquad t \geqq t_0. \tag{3.4}$$

We may then obtain the conclusion of Theorem 1.1 as follows. Multiply (3.4) by $e^{-2\beta t}$ and integrate from t to $\infty (t \geqq t_0)$ to obtain

$$\begin{aligned}\int_t^\infty e^{-2\beta s}E(u,s)\,ds + 2\beta\int_t^\infty e^{-2\beta s}Q(s)\,ds + Q(s)\,e^{-2\beta s}\Big|_t^\infty \\ \leqq C\int_0^\infty\int_\Omega e^{-2\beta t}u_i u_i\,dx\,dt,\end{aligned} \tag{3.5}$$

where C is independent of β. For t sufficiently large we have

$$0 \leqq Q(t) \leqq CtE(u,0).$$

In fact, from Korn's inequality [5, Thm. 3.1] and [5, Thm. 3.3],

$$\begin{aligned}\int_\Omega |2(f\cdot\nabla u_i)\rho u_{i,t} + (f_{m,m}-1)\rho u_i u_{i,t}|\,dx \\ \leqq C\int_\Omega (u_{i,j}u_{i,j} + u_i u_i + u_{i,t}u_{i,t})\,dx \\ \leqq C\int_\Omega (\varepsilon_{ij}\varepsilon_{ij} + q u_i u_i + u_{i,t}u_{i,t})\,dx \\ \leqq C\int_\Omega (a_{ijkl}\varepsilon_{ij}\varepsilon_{kl} + q u_i u_i + u_{i,t}u_{i,t})cx \\ \leqq CE(u,t) \leqq CE(u,0).\end{aligned}$$

It follows from (3.5) that

$$\int_{t_1}^\infty e^{-2\beta s}E(u,s)\,ds \leqq CE(u,0) + \int_0^\infty\int_\Omega e^{-2\beta t}u_i u_i\,dx\,dt \tag{3.6}$$

for a suitable $t_1 > 0$, where C depends on t_1 but not on β. Since

$$\int_0^{t_1} e^{-2\beta s}E(u,s)\,ds \leqq t_1 E(u,0)$$

we may replace t_1 in (3.6) by zero. We now invoke Proposition 3.1 with a small enough ε and obtain

$$\int_0^\infty e^{-2\beta s}E(u,s)\,ds \leqq CE(u,0).$$

Since C does not depend on β it follows that

$$\int_0^\infty E(u,s)\,ds \leqq CE(u,0). \tag{3.7}$$

This has been proved under the assumption $E(u_t, 0) < +\infty$, but may easily be extended to initial data satisfying $E(u,0) < +\infty$ by approximating such data by smoother data. The exponential decay (1.13) follows from (3.7) in a standard way (cf. [7]) because of the semigroup property of the map $S(t)$: $(u(\cdot,0), u_t(\cdot,0)) \to (u(\cdot,t), u_t(\cdot,t))$.

To complete the proof of Theorem 1.1 we must still verify (3.4). Using Gauss' theorem we obtain from (3.3)

$$\begin{aligned}\dot{Q}(t) &= (\alpha-1)E(u,t)+\int_\Omega Q_2(x,t)\,dx\\ &\quad+\int_{\Gamma_0}[2(f\cdot\nabla u_i)\sigma_{ij}n_j-\sigma_{im}\varepsilon_{im}(f\cdot n)]\,d\sigma\\ &\quad+\int_{\Gamma_0}[(\alpha+1)tu_{i,t}\sigma_{ij}n_j+2(f\cdot\nabla u_i)\sigma_{ij}n_j\\ &\qquad+(f_{m,m}-1)\sigma_{ij}u_in_j+\rho u_{i,t}u_{i,t}(f\cdot n)-\sigma_{im}\varepsilon_{im}(f\cdot n)-qu_iu_i(f\cdot n)]\,d\sigma.\end{aligned}\tag{3.8}$$

The integral over Γ_0 is nonpositive for the following reason. On Γ_0, $u=0$ hence $u_{i,j}=(\partial u_i/\partial n)n_j$ there. Therefore

$$\begin{aligned}2(f\cdot\nabla u_i)\sigma_{ij}n_j-\sigma_{im}\varepsilon_{im}(f\cdot n) &= 2(f\cdot n)\sigma_{ij}u_{i,j}-\sigma_{im}\varepsilon_{im}f\cdot n\\ &= (f\cdot n)\sigma_{ij}\varepsilon_{ij}\end{aligned}$$

because of the symmetry of σ_{ij}. The last expression is nonpositive on Γ_0 because of (1.9) and the ellipticity condition (1.4). Thus

$$\dot{Q}(t)\leqq(\alpha-1)E(u,t)+\int_\Omega Q_2(x,t)\,dx+\int_{\Gamma_1}[\cdots]\,d\sigma,$$

where $[\cdots]$ denotes the integrand in the integral over Γ_1 on the right side of (3.8).

Next we estimate $\int_\Omega Q_2\,dx$. Using (1.11) and (1.12) we have

$$\begin{aligned}\int_\Omega Q_2\,dx &= \int_\Omega[-(f\cdot\nabla\rho)u_{i,t}u_{i,t}-\Lambda(Du)+2\sigma_{ij}\varepsilon_{ij}]\,dx\\ &\quad+\int_\Omega[(f\cdot\nabla q)+2q]u_iu_i\,dx-\int_\Omega\sigma_{ij}u_if_{m,mj}\,dx\\ &\leqq\int_\Omega[-\alpha\rho u_{i,t}u_{i,t}-\alpha\sigma_{ij}\varepsilon_{ij}-\alpha qu_iu_i]\,dx\\ &\quad-\int_\Omega\sigma_{ij}u_if_{m,mj}\,dx+C\int_\Omega u_iu_i\,dx\\ &\leqq-2\alpha E(u,t)+C\int_\Omega u_iu_i\,dx-\int_\Omega\sigma_{ij}u_if_{m,mj}\,dx.\end{aligned}$$

Thus

$$\dot{Q}(t)\leqq-(\alpha+1)E(u,t)+C\int_\Omega u_iu_i\,dx-\int_\Omega\sigma_{ij}u_if_{m,mj}\,dx+\int_{\Gamma_1}[\cdots]\,d\sigma.$$

By the Schwarz inequality and (1.4)

$$|\sigma_{ij}u_if_{m,mj}|\leqq C(\varepsilon_{kl}\varepsilon_{kl})^{1/2}(u_iu_i)^{1/2}\leqq\delta\varepsilon_{kl}\varepsilon_{kl}+C_\delta u_iu_i\leqq\frac{\delta}{c_0}\sigma_{ij}\varepsilon_{ij}+C_\delta u_iu_i$$

for any $\delta>0$. Choosing δ sufficiently small gives

$$\dot{Q}(t)\leqq-\tfrac{1}{2}(\alpha+1)E(u,t)+C\int_\Omega u_iu_i\,dx+\int_{\Gamma_1}[\cdots]\,d\sigma.$$

To estimate the remaining boundary integral we use the boundary condition (2.3). Thus

$$\int_{\Gamma_1}[\cdots]\,d\sigma=\int_{\Gamma_1}[-(\alpha+1)tbu_{i,t}u_{i,t}-2b(f\cdot\nabla u_i)u_{i,t}-b(f_{m,m}-1)u_iu_{i,t}$$
$$+(f\cdot n)\rho u_{i,t}u_{i,t}-(f\cdot n)\sigma_{im}\varepsilon_{im}-qu_iu_i(f\cdot n)]\,d\sigma.$$

The last term in the integral is nonpositive because of (1.10). The second term is estimated as follows:

$$\left|\int_{\Gamma_1}2b(f\cdot\nabla u_i)u_{i,t}\,d\sigma\right|\leqq\delta\int_{\Gamma_1}u_{i,j}u_{i,j}\,d\sigma+C_\delta\int_{\Gamma_1}u_{i,t}u_{i,t}\,d\sigma$$
$$\leqq\frac{\delta}{c_0}\int_{\Gamma_1}c_{ijkl}u_{i,j}u_{k,l}\,d\sigma+C_\delta\int_{\Gamma_1}u_{i,t}u_{i,t}\,d\sigma$$
$$=\frac{\delta}{c_0}\int_{\Gamma_1}\sigma_{ij}\varepsilon_{ij}\,d\sigma+C_\delta\int_{\Gamma_1}u_{i,t}u_{i,t}\,d\sigma.$$

As for the third term, we have

$$\left|\int_{\Gamma_1}b(f_{m,m}-1)u_{i,t}u_i\,d\sigma\right|\leqq\delta\int_{\Gamma_1}u_iu_i\,d\sigma+C_\delta\int_{\Gamma_1}u_{i,t}u_{i,t}\,d\sigma$$
$$\leqq\delta C\int_\Omega(u_iu_i+u_{i,j}u_{i,j})\,dx+C_\delta\int_{\Gamma_1}u_{i,t}u_{i,t}\,d\sigma$$
$$\leqq\delta C\int_\Omega(\varepsilon_{ij}\varepsilon_{ij}+qu_iu_i)\,dx+C_\delta\int_{\Gamma_1}u_{i,t}u_{i,t}\,d\sigma$$
$$\leqq\delta CE(u,t)+C_\delta\int_{\Gamma_1}u_{i,t}u_{i,t}\,d\sigma.$$

By choosing δ sufficiently small we obtain

$$\dot{Q}(t)\leqq-\lambda E(u,t)+C\int_\Omega u_iu_i\,dx+\int_{\Gamma_1}[-(\alpha+1)tb+C_1]u_{i,t}u_{i,t}\,d\sigma$$
$$+\int_{\Gamma_1}(\lambda-(f\cdot n))\sigma_{ij}\varepsilon_{ij}\,d\sigma,$$

where $0<\lambda<\inf_{\Gamma_1}(f\cdot n)$. The last integral is therefore nonpositive and the remaining integral over Γ_1 is nonpositive for all sufficiently large t. This proves (3.4) and completes the proof of Theorem 1.1.

Proof of Theorem 1.2. Again we may suppose without loss of generality that $E(u_t,0)<+\infty$. Then, as we shall prove below,

$$\int_0^\infty e^{-2\beta t}E(u,t)\,dt\leqq C_\varepsilon(E(u,0)+E(u_t,0))+\varepsilon\int_0^\infty\int_\Omega e^{-2\beta t}u_{i,tt}u_{i,tt}\,dx\,dt. \tag{3.9}$$

Theorem 1.2 follows from (3.9). In fact, since $(d/dt)E(u,t)\leqq0$ we have

$$E(u,t)=E(u,\infty)+F(u,t),$$

where $F(u,t)\geqq0$. From (3.9)

$$\frac{1}{2\beta}E(u,\infty)\leqq C_\varepsilon(E(u,0)+E(u_t,0))+\varepsilon\int_0^\infty\int_\Omega e^{-2\beta t}u_{i,tt}u_{i,tt}\,dx\,dt.$$

As was proved in Theorem 2.2, the function $w = \partial u/\partial t$ also satisfies (2.1)–(2.3), hence $E(w, t) \leqq E(u_t, 0)$ and the last integral does not exceed $\varepsilon CE(u_t, 0)/\beta$ for some constant C independent of ε and β. Therefore

$$E(u, \infty) \leqq 2\beta C_\varepsilon (E(u, 0) + E(u_t, 0)) + \varepsilon CE(u_t, 0).$$

Letting $\beta \downarrow 0$ we get

$$E(u, \infty) \leqq \varepsilon CE(u_t, 0) \quad \forall \varepsilon > 0,$$

that is, $E(u, \infty) = 0$.

To prove (3.9) we first apply Proposition 3.1 to the solution $w = \partial u/\partial t$ to obtain

$$\int_0^\infty \int_\Omega e^{-2\beta t} u_{i,t} u_{i,t}\, dx\, dt \leqq C_\varepsilon E(u_t, 0) + \varepsilon \int_0^\infty \int_\Omega e^{-2\beta t} u_{i,tt} u_{i,tt}\, dx\, dt. \tag{3.10}$$

From (2.15) we have

$$a(u(t), v) + (u''(t), v) + (b(\gamma_0 u)'(t), \gamma_0 v)_{\Gamma_1} = 0 \quad \forall v \in V_{\Gamma_0}, \quad t > 0, \tag{3.11}$$

where

$$a(u, v) = \int_\Omega (c_{ijkl} u_{k,l} v_{i,j} + qu \cdot v)\, dx.$$

Replace v by $e^{-2\beta t} u(t)$ and integrate from 0 to ∞ to obtain

$$\begin{aligned}
&\int_0^\infty e^{-2\beta t} a(u(t), u(t))\, dt \\
&\quad = -\int_0^\infty e^{-2\beta t} (u''(t), u(t))\, dt - \int_0^\infty e^{-2\beta t} (b(\gamma_0 u)', \gamma_0 u)_{\Gamma_1}\, dt \\
&\quad = \int_0^\infty e^{-2\beta t} |u'(t)|^2\, dt - \int_0^\infty e^{-2\beta t} \frac{\partial}{\partial t} [(u'(t), u(t)) + \tfrac{1}{2} |\sqrt{b}\gamma_0 u|_{\Gamma_1}^2]\, dt.
\end{aligned}$$

The last integral is evaluated by parts and may be written

$$[e^{-2\beta t}((u'(t), u(t)) + \tfrac{1}{2}|\sqrt{b}\gamma_0 u|_{\Gamma_1}^2)]_0^\infty + 2\beta \int_0^\infty e^{-2\beta t} [(u'(t), u(t)) + \tfrac{1}{2} |\sqrt{b}\gamma_0 u(t)|_{\Gamma_1}^2]\, dt.$$

The limit at ∞ vanishes since the factor multiplying $e^{-2\beta t}$ does not exceed

$$\begin{aligned}
C(|u'(t)|^2 + |u(t)|^2 + \|u(t)\|^2) &\leqq C(|u'(t)|^2 + a(u(t), u(t))) \\
&\leqq CE(u, t) \leqq CE(u, 0).
\end{aligned}$$

It follows that

$$\begin{aligned}
\int_0^\infty e^{-2\beta t} a(u(t), u(t))\, dt &\leqq \int_0^\infty e^{-2\beta t} |u'(t)|^2\, dt + [(u_0, v_0) + \tfrac{1}{2} |\sqrt{b}\gamma_0 u|_{\Gamma_1}^2] \\
&\quad + 2\beta \int_0^\infty e^{-2\beta t} |(u'(t), u(t))|\, dt \\
&\leqq C \int_0^\infty e^{-2\beta t} (|u'(t)|^2 + |u(t)|^2)\, dt + E(u, 0)
\end{aligned}$$

if $\beta \leqq 1$, where C is independent of β. (If $\beta > 1$, the left side of (3.9) does not exceed

$\frac{1}{2}E(u, 0)$.) Since

$$\int_0^\infty e^{-2\beta t}E(u, t)\,dt = \tfrac{1}{2}\int_0^\infty e^{-2\beta t}\left[\int_\Omega \rho u_{i,t}u_{i,t}\,dx + a(u(t), u(t))\right]dt,$$

(3.9) follows readily from the last inequality, (3.1) and (3.10). This completes the proof of Theorem 1.2.

Proof of Proposition 3.1. As in the proofs of Theorems 1.1 and 1.2, we may assume $E(u_t, 0) < +\infty$, without loss of generality. Then u satisfies (3.11), in which v may clearly be chosen to be complex-valued. That is

$$a(u(t), \bar{v}) + (u''(t), \bar{v}) + (b(\gamma_0 u)'(t), \gamma_0\bar{v})_{\Gamma_1} = 0 \tag{3.12}$$

for every v in the complex space V_{Γ_0} defined as in § 2 but using complex $H^1(\Omega)$ space. $\bar{v} = (\bar{v}_1, \bar{v}_2, \bar{v}_3)$ denotes the complex conjugate of $v = (v_1, v_2, v_3)$.

Let $T > 0$ be fixed and $\phi \in C^\infty(R)$ such that $\phi(0) = \phi'(0) = 0$ and $\phi(t) = 1$ for $t \geqq T$, and set $w = \phi u$. Then $w(0) = w'(0) = 0$ and $w(t) = u(t)$ for $t \geqq T$. From (3.12) we see that w satisfies

$$a(w(t), \bar{v}) + (w''(t), \bar{v}) + (b(\gamma_0 w)'(t), \bar{v})_{\Gamma_1} = (g, \bar{v}) + (h, \bar{v})_{\Gamma_1} \tag{3.13}$$

where

$$g = 2\phi' u' + \phi'' u \in L^2(0, \infty; V_{\Gamma_0}),$$
$$h = b\phi'(\gamma_0 u) \in L^2(0, \infty; (H^{1/2}(\Gamma))^3)$$

satisfy $g = h = 0$ for $t \geqq T$, $h = 0$ on Γ_0. It follows that

$$\int_0^\infty dt \int_\Omega g_i g_i\,dx = \int_0^T dt \int_\Omega g_i g_i\,dx \leqq CE(u, 0), \tag{3.14}$$

$$\int_0^\infty \|h_i\|^2_{H^{1/2}(\Gamma)}\,dt \leqq CE(u, 0), \qquad i = 1, 2, 3. \tag{3.15}$$

Let ω be a complex parameter with $\operatorname{Im}\omega < 0$ and $\hat{w}$ be the Fourier transform of w:

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

The integral converges in V_{Γ_0} for each such ω. If in (3.12) we replace v by $e^{i\omega t}v$ and integrate in t from 0 to ∞ it follows that $\hat{w}$ satisfies

$$a(\hat{w}, \bar{v}) - \omega^2(\hat{w}, \bar{v}) + i\omega(b\gamma_0\hat{w}, \gamma_0\bar{v})_{\Gamma_1} = (\hat{g}, \bar{v}) + (\hat{h}, \bar{v})_{\Gamma_1} \quad \forall v \in V_{\Gamma_0},$$

since $w(0) = w'(0) = 0$, where $\hat{g} \in V_{\Gamma_0}$, $\hat{h} \in (H^{1/2}_{\Gamma_0}(\Gamma))^3$ are the Fourier transforms of g and h, respectively. We shall write the last equation as

$$a_\omega(\hat{w}, v) = (\hat{g}, \bar{v}) + (\hat{h}, \bar{v})_{\Gamma_1},$$

where for w, v in V_{Γ_0},

$$a_\omega(w, v) \doteq a(w, \bar{v}) - \omega^2(w, \bar{v}) + i\omega(b\gamma_0 w, \gamma_0\bar{v})_{\Gamma_1}.$$

We now consider, for general complex values of ω and arbitrary $g \in (L^2(\Omega))^3$, $h \in (H^{1/2}_{\Gamma_0}(\Gamma))^3$, the problem

$$\begin{aligned} &w \in V_{\Gamma_0}, \\ &a_\omega(w, v) = (g, \bar{v}) + (h, \bar{v})_{\Gamma_1} \quad \forall v \in V_{\Gamma_0}, \end{aligned} \tag{3.16}$$

where $L^2(\Omega)$, $H_{\Gamma_0}(\Gamma)$ are now complex spaces. Thus, for example,

$$|g|^2 = (g, \bar{g}) = \int_\Omega g(x)\bar{g}(x)\rho\, dx.$$

Proposition 3.1 will follow from the following result.

LEMMA 3.1. *Equation* (3.16) *has a unique solution* w *for each* ω *in some neighborhood* $\mathcal{N}$ *of the half-plane* $\operatorname{Im}\omega \leqq 0$. *For* $\omega \in \mathcal{N}$ *one has the estimate*

$$\|w\| \leqq C(\omega)(|g| + \|h\|_{1/2,\Gamma_1}), \tag{3.17}$$

where $C(\omega)$ *is bounded on bounded sets in* $\mathcal{N}$.

The proof of the Proposition 3.1 may now be completed as follows. It suffices to assume $0 < \beta < \beta_\varepsilon$ for some $\beta_\varepsilon > 0$. Write $\omega = \alpha - i\beta$ for $\beta > 0$ and small. By Plancherel's theorem

$$\int_0^\infty e^{-2\beta t}|w(\cdot, t)|^2\, dt = \frac{1}{2\pi}\int_{-\infty}^\infty |\hat{w}(\cdot, \omega)|^2\, d\alpha,$$

$$\int_0^\infty e^{-2\beta t}|w'(\cdot, t)|^2\, dt = \frac{1}{2\pi}\int_{-\infty}^\infty |\omega|^2|\hat{w}(\cdot, \omega)|^2\, d\alpha.$$

Let $A > 0$ be so large that $A^{-2} < \varepsilon$. By Lemma 3.1 there is a number $\delta = \delta(A) > 0$ such that if $|\alpha| \leqq A$ and $|\beta| \leqq \delta$ the estimate (3.17) holds for $\hat{w}$. Thus

$$\int_{-A}^A |\hat{w}(\cdot, \omega)|^2\, d\alpha \leqq C(A)\left[\int_{-\infty}^\infty |\hat{g}(\cdot, \omega)|^2\, d\alpha + \int_{-\infty}^\infty \|\hat{h}(\cdot, \omega)\|^2_{1/2,\Gamma_1}\, d\alpha\right]$$

$$= C_\varepsilon\left[\int_{-\infty}^\infty e^{-2\beta t}|g(\cdot, t)|^2\, dt + \int_{-\infty}^\infty e^{-2\beta t}\|h(\cdot, t)\|^2_{1/2,\Gamma_1}\, dt\right] \leqq C_\varepsilon E(u, 0),$$

where we have used (3.14), (3,15). Also

$$\int_{|\alpha|>A} |\hat{w}(\cdot, \omega)|^2\, d\alpha \leqq \int_{|\alpha|>A} \frac{\alpha^2}{A^2}|\hat{w}(\cdot, \omega)|^2\, d\alpha \leqq \varepsilon\int_{-\infty}^\infty |\omega|^2|\hat{w}(\cdot, \omega)|^2\, d\alpha$$

$$= 2\pi\varepsilon\int_0^\infty e^{-2\beta t}|w'(\cdot, t)|^2\, dt.$$

Therefore

$$\int_0^\infty e^{-2\beta t}|w(\cdot, t)|^2\, dt = \frac{1}{2\pi}\left(\int_{|\alpha|\leqq A} + \int_{|\alpha|>A}\right)|\hat{w}(\cdot, \omega)|^2\, d\alpha$$

$$\leqq C_\varepsilon E(u, 0) + \varepsilon\int_0^\infty e^{-2\beta t}|w'(\cdot, t)|^2\, dt, \tag{3.18}$$

which is almost the result in Proposition 3.1, except that w appears in the integrals instead of u. But using the fact that $w = \phi u$ and that $\phi(t) = 1$ for $t \geqq T$ one easily obtains

$$\int_0^\infty e^{-2\beta t}|u(\cdot, t)|^2\, dt \leqq CE(u, 0) + \int_0^\infty e^{-2\beta t}|w(\cdot, t)|^2\, dt,$$

$$\int_0^\infty e^{-2\beta t}|w'(\cdot, t)|^2\, dt \leqq C\left[E(u, 0) + \int_0^\infty e^{-2\beta t}|u'(\cdot, t)|^2\, dt\right]$$

from which (3.1) follows, in view of (3.18).

Proof of Lemma 3.1. The proof is similar to the proof of [7, Lemma 1], although that result is weaker and pertains only to solutions of the ordinary wave equation. One considers $a_\omega(w, v)$ as a sesquilinear form in H with dense domain V_{Γ_0}. Then a_ω is a holomorphic family of type (a) [6, p. 395] in a half plane $\mathcal{P}$: $\operatorname{Im}\omega < \delta$ for some $\delta > 0$. This means that (i) $a_\omega(v, v)$ is holomorphic in $\omega \in \mathcal{P}$ for each $v \in V_{\Gamma_0}$, and (ii) each a_ω, $\omega \in \mathcal{P}$, is sectorial and closed with constant dense domain. Clearly only (ii) has to be checked.

Writing $\omega = \alpha - i\beta$ we calculate (note $a(v, \bar{v})$ is real)

$$\operatorname{Re} a_\omega(v, v) = a(v, \bar{v}) - (\alpha^2 - \beta^2)|v|^2 + \beta|\sqrt{b}\gamma_0 v|^2_{\Gamma_1}, \tag{3.19}$$

$$\operatorname{Im} a_\omega(v, v) = 2\alpha\beta|v|^2 + \alpha|\sqrt{b}\gamma_0 v|^2_{\Gamma_1}. \tag{3.20}$$

Since

$$|\sqrt{b}\gamma_0 v|^2_{\Gamma_1} \leqq C\|v\|^2 \leqq C_1 a(v, \bar{v}),$$

if $\beta > -\delta$ with δ small we have

$$\operatorname{Re} a_\omega(v, v) + \alpha^2|v|^2 \geqq C\|v\|^2, \tag{3.21}$$

$$|\operatorname{Im} a_\omega(v, v)| \leqq C(\omega)[\operatorname{Re} a_\omega(v, v) + \alpha^2|v|^2]. \tag{3.22}$$

Equations (3.21) and (3.22) imply that the values of $a_\omega(v, v)$ lie in the sector $|\arg(\zeta + \alpha^2)| \leqq \theta$, where $0 < \theta < \pi/2$ satisfies $\tan\theta = C(\omega)$, and that a_ω is closed.

Associated with a_ω is an m-sectorial operator $\mathcal{A}_\omega$ in H such that

$$(\mathcal{A}_\omega w, \bar{v}) = a_\omega(w, v)$$

for every $v \in V_{\Gamma_0}$ and $w \in D(\mathcal{A}_\omega)$, and one has $D(\mathcal{A}_\omega) \subset D(a_\omega) = V_{\Gamma_0}$. Since a_ω is holomorphic in the half plane $\mathcal{P}$: $\operatorname{Im}\omega < \delta$, $\mathcal{A}_\omega$ is a holomorphic family of operators in $\mathcal{P}$ [6, Thm. 4.2, p. 395]. Furthermore from (3.12) and the compactness of the injection $V_{\Gamma_0} \to H$, it follows that $\mathcal{A}_\omega$ has a compact resolvent for each $\omega \in \mathcal{P}$. We now invoke [6, Thm. 1.10, p. 371] and conclude that either zero is an eigenvalue of each $\mathcal{A}_\omega$, or else $\mathcal{A}_\omega^{-1}$ exists as a bounded operator on H for all $\omega \in \mathcal{P}$ with the possible exception of a finite number of values in each compact subset of $\mathcal{P}$. As we shall show, zero cannot be an eigenvalue of $\mathcal{A}_\omega$ if $\operatorname{Im}\omega \leqq 0$. From this it follows that $\mathcal{A}_\omega^{-1}$ exists fror all ω in some neighbourhood $\mathcal{N}$ of $\operatorname{Im}\omega \leqq 0$.

Thus suppose $-\beta = \operatorname{Im}\omega \leqq 0$ and that $w \in V_{\Gamma_0}$ satisfies

$$a_\omega(w, v) = 0 \quad \forall v \in V_{\Gamma_0}.$$

If $\alpha = 0$ it follows from (3.19) that $a(w, \bar{w}) = 0$, hence $w \equiv 0$. If $\alpha \neq 0$ we have from (3.20)

$$\alpha\beta|w|^2 + |\sqrt{b}\gamma_0 w|^2_{\Gamma_1} = 0$$

hence, if $\beta > 0$, $w \equiv 0$ once again. If $\beta = 0$, then since $b \not\equiv 0$ on Γ_1 there is a point $x_0 \in \Gamma_1$ an open ball S_ρ in R^3 centered at x_0 such that $w = 0$ on $\Gamma_1 \cap S_\rho$ and, from (3.16),

$$a(w, \bar{v}) - \alpha^2(w, \bar{v}) = 0 \quad \forall v \in V_{\Gamma_0}. \tag{3.23}$$

Choose ρ so small that S_ρ contains no point of Γ_0, and extend w into $\Omega_\rho \doteq \Omega \cup S_\rho$ by setting $w = 0$ in $\Omega_\rho - \bar{\Omega}$. Then $w \in (H^1(\Omega_\rho))^3$ since $w = 0$ on $\Gamma_1 \cap S_\rho$. If $v \in (C_0^\infty(\Omega_\rho))^3$ its restriction to Ω is in V_{Γ_0} hence, from (3.23),

$$\int_{\Omega_\rho} w_k \left[\frac{\partial}{\partial x_l}(c_{ijkl}\bar{v}_{i,j}) + (\alpha^2\rho - q)\bar{v}_k\right] dx = 0.$$

Thus w is a weak solution in Ω_ρ of the elliptic system

$$\frac{\partial}{\partial x_j}(c_{ijkl}w_{k,l})+(\alpha^2\rho-q)w_i=0, \qquad i=1,2,3. \tag{3.24}$$

But $w=0$ in the open set $\Omega_\rho-\bar{\Omega}$ hence, by the unique continuation property of solutions of (3.24), $w\equiv 0$ in Ω_ρ.

Thus $\mathscr{A}_\omega^{-1}$ exists in a neighborhood $\mathscr{N}$ of $\operatorname{Im}\omega\leqq 0$. If $\omega\in\mathscr{N}$ and $|\alpha|=|\operatorname{Re}\omega|\leqq\omega_0<+\infty$ we have from (3.21)

$$\|v\|\leqq C(\omega_0)[|\mathscr{A}_\omega v|+|v|] \quad \forall v\in D(\mathscr{A}_\omega).$$

But, since the null space of $\mathscr{A}_\omega$ is $\{0\}$, the term $|v|$ on the right may be dropped by a standard argument. This completes the proof when $h\equiv 0$. In the general case we use the trace theorem to obtain a function $w\in(H^2(\Omega))^3$ such that, on Γ, $w=0$, $\sigma_{ij}n_j=h_i$ $(i=1,2,3)$, and

$$\sum_{i=1}^{3}\|w_i\|_{H^2(\Omega)}\leqq C\|h\|_{1/2,\Gamma_1}. \tag{3.25}$$

Let Z be the unique solution in $D(\mathscr{A}_\omega)$ of $\mathscr{A}_\omega Z=\tilde{g}$ where

$$\tilde{g}=g_i+\frac{1}{\rho}[\sigma_{ij,j}+(\omega^2\rho-q)w_i].$$

By what has already been proved,

$$\|Z\|\leqq C(\omega)|\tilde{g}|\leqq C(\omega)(|g|+\|h\|_{1/2,\Gamma_1}), \tag{3.26}$$

where $C(\omega)$ is bounded on bounded sets in $\mathscr{N}$. Set $u=Z+w\in V_{\Gamma_0}$. Then if $v\in V_{\Gamma_0}$ we have

$$\begin{aligned}a_\omega(u,v)&=(\mathscr{A}_\omega Z,\bar{v})+a_\omega(w,v)\\&=(g,\bar{v})-\int_\Omega[\sigma_{ij}\bar{v}_{i,j}+(q-\omega^2\rho)w_i\bar{v}_i]\,dx\\&\quad+\int_{\Gamma_1}\sigma_{ij}n_j\bar{v}_i\,d\sigma+a_\omega(w,v)=(g,\bar{v})+(h,\bar{v})_{\Gamma_1}.\end{aligned}$$

Thus u is the unique solution of (3.16) and from (3.25), (3.26) we have the estimate

$$\|u\|\leqq\|Z\|+\|w\|\leqq C(\omega)(|g|+\|h\|_{1/2,\Gamma_1}).$$

This completes the proof of Lemma 3.1.

REFERENCES

[1] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* II, Comm. Pure Appl. Math., XVII (1964), pp. 35–92.

[2] C. D. BENCHIMOL, *A note on weak stabilization of contraction semigroups*, this Journal, 16 (1978), pp. 373–379.

[3] G. CHEN, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249–274.

[4] ———, *A note on boundary stabilization of the wave equation*, this Journal, 19 (1981), pp. 106–113.

[5] G. DUVAUT AND J. L. LIONS, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.

[6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[7] J. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, to appear.

[8] 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, 77A (1977), pp. 97–127.

[9] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189–211.

[10] M. SLEMROD, *Stabilization of boundary control systems*, J. Differential Equations, 22 (1976), pp. 402–415.

[11] W. STRAUSS, *Dispersal of waves vanishing on the boundary of an exterior domain*, Comm. Pure Appl. Math., 28 (1975), pp. 265–278.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 21, No. 6, November 1983

ERRATUM: ADMISSIBLE INPUT ELEMENTS FOR SYSTEMS IN HILBERT SPACE AND A CARLESON MEASURE CRITERION*

L. F. HO† AND D. L. RUSSELL‡

Page 638 of the paper as printed repeats the text of p. 630. The correct version of p. 638 is as follows:

Since μ is a Carleson measure and (4.11) holds,

$$\mu(E_s) \leqq \sum_{n=1}^{\infty} \mu(S_n) \leqq A \sum_{n=1}^{\infty} \frac{|J_n|}{2}$$

$$= \frac{5A}{2} \sum_{n=1}^{\infty} |I_n| \leqq \frac{5A}{2s} \|\psi\|_{L^1(-\infty,\infty)}.$$

PROPOSITION 4.4. *Let $\phi \in H_0^2$ with boundary function $\phi_0(i\cdot) \in L^2(-\infty, \infty)$. Let $\phi(z)$ be defined by* (4.6). *Then, if μ is a Carleson measure on $\{z | \mathrm{Re}\,(z) > 0\}$,*

$$\int_{\mathrm{Re}\,(z)>0} (\tilde{\phi}(z))^2 \, d\mu(z) \leqq 10A \int_{-\infty}^{\infty} |\phi_0(it)|^2 \, dt. \tag{4.12}$$

Proof. For each $r > 0$ let

$$\psi_r(t) = \begin{cases} \phi_0(it) & \text{if } |\phi_0(it)| > r, \\ 0 & \text{otherwise.} \end{cases}$$

From $\phi_0(i\cdot) \in L^2(-\infty, \infty)$, we conclude that the support of ψ_r is a subset $\sum_r$ of $(-\infty, \infty)$ of finite (Lebesgue) measure. Then $\psi_r \in L^2(\sum_r) \subset L^1(\sum_r)$, and we conclude, since ψ_r vanishes outside $\sum_r$, that $\psi_r \in L^1(-\infty, \infty)$. Moreover

$$\int_0^{\infty} \|\psi_r\|_{L^1(-\infty,\infty)} \, dr = \int_0^{\infty} \int_{\sum_r} |\phi_0(it)| \, dt \, dr$$

$$= \int_{-\infty}^{\infty} \int_0^{|\phi_0(it)|} dr |\phi_0(it)| \, dt = \int_{-\infty}^{\infty} |\phi_0(it)|^2 \, dt \tag{4.13}$$

$$= \|\phi_0(i\cdot)\|^2_{L^2(-\infty,\infty)}.$$

Let $\alpha(s) = \mu(E_s)$. Then we can see that

$$\int_{\mathrm{Re}\,(z)>0} (\tilde{\phi}(z))^2 \, d\mu(z) = -\int_0^{\infty} s^2 \, d\alpha(s) = s \int_0^{\infty} s\alpha(s) \, ds. \tag{4.14}$$

From the definition (4.8) of $\tilde{\psi}$ it is clear that for any two such functions, ψ_1, ψ_2, we have

$$(\psi_1 + \psi_2)(z) \leqq \tilde{\psi}_1(z) + \tilde{\psi}_2(z).$$

---

* This Journal, 21 (1983), pp. 614–640.

† Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019.
‡ Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

Hence

$$\begin{aligned}\tilde{\phi}(z) &= (\psi_r + (\tilde{\phi}_0(i\cdot) - \psi_r))(z) \\ &\leqq \tilde{\psi}_r(z) + \widetilde{(\phi_0(i\cdot) - \psi_r)}(z) \\ &\leqq \tilde{\psi}_r(z) + r\end{aligned} \tag{4.15}$$

since $|\phi(it) - \psi_r(t)|$ is either equal to 0 or is $\leqq r$. Let

$$F_s = \{z \,|\, \tilde{\psi}_r(z) > s\}.$$

Suppose $z \in E_{2r}$. Then $\tilde{\phi}(z) > 2r$ and (4.15) gives

$$\tilde{\psi}_r(z) \geqq \tilde{\phi}(z) - r > r,$$

and we conclude $z \in F_r$. Thus

$$E_{2r} \subset F_r.$$