SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

STABILIZABILITY OF LINEAR SYSTEMS OVER A COMMUTATIVE NORMED ALGEBRA WITH APPLICATIONS TO SPATIALLY-DISTRIBUTED AND PARAMETER-DEPENDENT SYSTEMS*

WILLIAM L. GREEN† AND EDWARD W. KAMEN‡

Abstract. The problem of achieving stabilization by using state feedback is considered for linear systems given by a pair of matrices whose entries belong to a real or complex commutative normed algebra. This framework is applicable to various types of linear systems, including spatially-distributed systems, systems depending on parameters, and infinite-dimensional systems. Necessary and sufficient conditions for stabilizability are derived in terms of solutions to an associated Riccati equation defined in the Gelfand-transform domain. Necessary and sufficient conditions for stabilizability are also given in terms of a local rank criterion involving the Gelfand transform of the system coefficients. The results are applied to the problem of positioning a long seismic cable.

Key words. linear systems, linear control, spatially-distributed systems, systems with unknown parameters, feedback control, systems over a normed ring, Riccati equations

1. Introduction.

In this paper we study the problem of stabilization for linear systems whose coefficients belong to a commutative normed algebra. This framework arises in the study of spatially-distributed systems, systems whose coefficients depend on parameters, and infinite-dimensional systems. We begin by describing these applications in some detail, and then we consider the problem of stabilization.

With $\mathbb{Z}$ equal to the set of integers and $\mathbb{R}$ equal to the field of real numbers, let $l^1(\mathbb{Z},\mathbb{R})$ denote the commutative convolution algebra of absolutely summable real-valued functions defined on $\mathbb{Z}$. We may regard a matrix over $l^1(\mathbb{Z},\mathbb{R})$ as an absolutely summable bi-infinite sequence of matrices over $\mathbb{R}$. A pair (F, G) consisting of a $n\times n$ matrix F over $l^1(\mathbb{Z},\mathbb{R})$ and a $n\times m$ matrix G over $l^1(\mathbb{Z},\mathbb{R})$ then defines a type of linear spatially-distributed continuous-time system given by the state equation

$$\frac{dx(t,r)}{dt}=\sum_{j=-\infty}^{\infty}F(r-j)x(t,j)+\sum_{j=-\infty}^{\infty}G(r-j)u(t,j),\qquad t\in\mathbb{R},\quad r\in\mathbb{Z}. \tag{1.1}$$

In (1.1), $x(t,r)\in\mathbb{R}^n$ is the state at time t and spatial point $r\in\mathbb{Z}$, and $u(t,r)\in\mathbb{R}^m$ is the input or control at time t and spatial point r. Representations of the form (1.1) arise in the study of long strings of coupled systems, such as strings of vehicles (see Melzer and Kuo [26] and Chu [6]). They also result from the discretization (with respect to the spatial coordinate) of partial differential equations. An example of such a discretization is the representation for a long seismic cable used in offshore oil exploration

* Received by the editors August 10, 1981, and in final revised form November 15, 1983. This work was supported in part by the U.S. Army Research Office, Research Triangle Park, North Carolina, under contracts DAAG29-80-K0076 and DAAG29-81-K-0166.

† School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332.

‡ Center for Mathematical System Theory, Department of Electrical Engineering, University of Florida, Gainesville, Florida 32611.

(El-Sayed and Krishnaprasad [7]) given by

$$\begin{bmatrix} \dfrac{dq(t,r)}{dt} \\ \dfrac{d^2q(t,r)}{dt^2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -k/M & -c/M \end{bmatrix} \begin{bmatrix} q(t,r) \\ \dfrac{dq(t,r)}{dt} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ k/2M & 0 \end{bmatrix} \begin{bmatrix} q(t,r-1) \\ \dfrac{dq(t,r-1)}{dt} \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & 0 \\ k/2M & 0 \end{bmatrix} \begin{bmatrix} q(t,r+1) \\ \dfrac{dq(t,r+1)}{dt} \end{bmatrix} + \begin{bmatrix} 0 \\ 1/M \end{bmatrix} u(t,r). \tag{1.2}$$

In (1.2), $q(t,r)$ is the position of the rth cable segment relative to some reference, $u(t,r)$ is the control applied to the rth cable segment, and k, c, M are positive constants. As in [7], we omit the boundary conditions, on the grounds that each cable segment is very short when compared with the total length of the cable. Clearly, (1.2) can be written in the form (1.1) by defining

$$F(0) = \begin{bmatrix} 0 & 1 \\ -k/M & -c/M \end{bmatrix}, \qquad F(-1) = F(1) = \begin{bmatrix} 0 & 0 \\ k/2M & 0 \end{bmatrix}, \qquad F(r) = 0, \quad |r| \geqq 2,$$

$$G(0) = \begin{bmatrix} 0 \\ 1/M \end{bmatrix}, \qquad G(r) = 0, \quad |r| \geqq 1.$$

The spatially-distributed representation (1.2) for the seismic cable is useful in problems involving the control of the cable since the control mechanism would be discrete in the spatial coordinate.

A pair (F, G) of matrices over $l^1(\mathbb{Z}, \mathbb{R})$ also defines a linear spatially-distributed discrete-time system given by

$$x(k+1,r) = \sum_{j=-\infty}^{\infty} F(r-j)x(k,j) + \sum_{j=-\infty}^{\infty} G(r-j)u(k,j), \qquad k, r \in \mathbb{Z}. \tag{1.3}$$

Equation (1.3) could be the representation for a long string of coupled discrete-time systems. Systems of the form (1.3) may arise by discretizing in time a spatially-distributed continuous-time system given by (1.1). For example, consider the state equation (1.2) for the seismic cable. We can write the solution to (1.2) in the form

$$x(t,r) = e^{F(0)(t-\lambda)}x(\lambda, r)$$

$$+ \int_{\lambda}^{t} e^{F(0)(t-\tau)}[F(1)x(\tau, r-1) + F(-1)x(\tau, r+1) + G(0)u(\tau, r)]\, d\tau, \qquad t > \lambda, \tag{1.4}$$

where

$$x(t,r) = \begin{bmatrix} q(t,r) & \dfrac{dq(t,r)}{dt} \end{bmatrix}'$$

(prime denotes the transpose operation). Now given a fixed real number $T > 0$, we can discretize (1.4) in the usual manner, which yields the following discrete-time approximation of the cable:

$$x(kT+T, r) = e^{F(0)T}x(kT, r) + C_T x(kT, r-1) + C_T x(kT, r+1) + D_T u(kT, r), \tag{1.5}$$

where

$$C_T = \int_0^T e^{F(0)\tau} F(1)\, d\tau, \qquad D_T = \int_0^T e^{F(0)\tau} G(0)\, d\tau. \tag{1.6}$$

Clearly, (1.5) can be written in the form (1.3) with the entries of F and G belonging to the algebra $l^1(\mathbb{Z}, \mathbb{R})$. This representation can be utilized to generate control laws that are discrete in both the time variable and the spatial variable. In § 5, we use the representation (1.5) in the study of a cable positioning problem.

In the application to spatially-distributed systems, the variable k in (1.3) is the discrete time variable and the variable r is the discrete spatial variable. If both k and r are interpreted to be spatial variables, (1.3) could be the state representation for a two-dimensional digital filter or "array processor", that is, a system which processes arrays of data (see Kamen [18], [19] and Kamen and Green [20], [21]).

Now let $\mathscr{C}(\Omega, \mathbb{R})$ denote the commutative algebra consisting of all real-valued continuous functions defined on a compact subset Ω of $\mathbb{R}^N$, where N is a fixed positive integer. As in the work of Byrnes [3], [4], [5], a pair (F, G) of matrices over $\mathscr{C}(\Omega, \mathbb{R})$ defines a linear continuous-time system whose coefficients depend continuously on N parameters. This system is given by the collection of differential equations

$$\frac{dx(t, \omega)}{dt} = F(\omega)x(t, \omega) + G(\omega)u(t, \omega), \qquad \omega \in \Omega. \tag{1.7}$$

Systems specified by (1.7) appear in applications where one or more of the system coefficients are sensitive to operating conditions such as temperature. Representations of the form (1.7) also result from the linearization of nonlinear systems with respect to nominal operating points specified in terms of a set of parameters. An example is the satellite problem (see Brockett [2, pp. 14–15]) which is linearized with respect to a nominal radius and nominal angular velocity.

A pair (F, G) of matrices over $\mathscr{C}(\Omega, \mathbb{R})$ also defines a parameter-dependent linear discrete-time system, given by the collection of difference equations

$$x(k+1, \omega) = F(\omega)x(k, \omega) + G(\omega)u(k, \omega), \qquad \omega \in \Omega. \tag{1.8}$$

Again, (1.8) may result by discretizing in time a continuous-time parameter-dependent system given by (1.7).

A common feature of the spatially-distributed systems and the parameter-dependent systems defined above is that they are specified in terms of a pair of matrices with entries in a commutative algebra ($l^1(\mathbb{Z}, \mathbb{R})$ in the former case and $\mathscr{C}(\Omega, \mathbb{R})$ in the latter case). It is well known that $l^1(\mathbb{Z}, \mathbb{R})$ is a commutative Banach algebra with the usual l^1 norm and that $\mathscr{C}(\Omega, \mathbb{R})$ is also a commutative Banach algebra with the usual sup norm. Both of these algebras have identities, so each class of systems defined above is given in terms of matrices defined over a commutative Banach algebra with identity.

It is thus natural for us to consider a pair (F, G) of matrices defined over an arbitrary commutative normed algebra B_0 with identity, and this brings us to our third class of systems. We claim that whenever B_0 is infinite-dimensional as a linear space, we may interpret such a pair (F, G) as a linear infinite-dimensional discrete-time system. To see this, first note that there exists a Banach space Y such that B_0 can be viewed as a subalgebra of the Banach algebra $\mathscr{B}(Y)$ of all bounded linear maps on Y (e.g., we can take Y to be the completion of B_0). Then the pair (F, G) defines a system

via the first-order difference equation

$$x_{k+1} = Fx_k + Gu_k, \qquad k \in \mathbb{Z}. \tag{1.9}$$

In (1.9), the state x_k at time k is an element of Y^n, the Banach space of n-element column vectors over Y, and the input or control u_k at time k is an element of Y^m. The terms Fx_k and Gu_k in (1.9) are computed via the usual action of a matrix on a column vector. Linear infinite-dimensional discrete-time systems defined by a difference equation in a Hilbert or Banach space have been studied by Lee, Chow, and Barr [25], Zabczyk [32], [33], Helton [15], [16], [17], Fuhrmann [9], [10], [11], Przyluski [27], [28], [29], [30], and others. In contrast to this past work, by exploiting the fact that B_0 is a commutative algebra we will be able to utilize Gelfand-transform techniques in the study of system behavior and in the study of control.

The primary purpose of the present paper is to study the problem of *feedback stabilization* for linear systems given by a pair (F, G) of matrices with entries in a commutative normed algebra B_0 with identity 1. Our specific objective is to derive necessary and sufficient conditions for the existence of a $m \times n$ *feedback matrix* L over B_0 such that the closed-loop system $(F-GL, G)$ is stable. When (F, G) is interpreted to be a continuous-time system (e.g., given by the state equation (1.1) or (1.7)), stability of the closed-loop system $(F-GL, G)$ means that

$$\|e^{(F-GL)t}\| \to 0 \quad \text{as } t \to \infty. \tag{1.10}$$

When (F, G) is interpreted to be a discrete-time system (e.g., given by the state equation (1.3) or (1.8)), stability of $(F-GL, G)$ means that

$$\|(F-GL)^k\| \to 0 \quad \text{as } k \to \infty \quad (k=0, 1, 2, \cdots). \tag{1.11}$$

The norms in (1.10) and (1.11) are the induced matrix norms (defined later).

The solution to the above-defined stabilization problem can be directly applied to the stabilization of spatially-distributed systems and systems whose coefficients are functions of parameters. The application of this algebra framework to the stabilization of systems with parameters was first considered by Byrnes [3], [4], [5]. To make this precise, suppose that $B_0 = \mathcal{C}(\Omega, \mathbb{R})$ and we interpret a system (F, G) over B_0 as a parameter-dependent discrete-time system given by the state equation (1.8). If we can find a matrix L over B_0 such that $F-GL$ is stable in the sense defined above, then with state feedback $u(k, \omega) = -L(\omega)x(k, \omega)$, the closed-loop system

$$x(k+1, \omega) = (F(\omega) - G(\omega)L(\omega))x(k, \omega), \qquad \omega \in \Omega,$$

will be asymptotically stable for every $\omega \in \Omega$. Hence, via this approach we can consider designing stabilizing compensators without knowing a priori (before the system is in operation) the specific values of the parameters. This control structure can be implemented by estimating the system parameters on-line.

In the application to spatially-distributed systems with $B_0 = l^1(\mathbb{Z}, \mathbb{R})$, the existence of a feedback matrix L over B_0 such that $(F-GL, G)$ is stable implies that with distributed state feedback

$$u(t, r) = -\sum_{j=-\infty}^{\infty} L(r-j)x(t, j), \qquad t \in \mathbb{R} \text{ or } t \in \mathbb{Z},$$

the resulting closed-loop system is stable uniformly across the distributed structure; that is, for any initial state $x(0, r)$ such that $\sup_r \|x(0, r)\| < \infty$, the free (unforced) response $x(t, r)$ of the closed-loop system converges to zero uniformly in the spatial variable r.

The property of stabilizability is of fundamental importance in a wide range of control problems. In fact, certain types of regulation problems, such as tracking and disturbance rejection, can be reduced to the problem of stabilizing an augmented system, a technique that is often employed in the control of finite-dimensional systems. This point is illustrated in § 5 where we apply our stabilizability results to a seismic-cable positioning problem.

In the special case when B_0 is equal to the reals $\mathbb{R}$, so that F and G are matrices over $\mathbb{R}$, it is well known that necessary and sufficient conditions for stabilizability can be expressed in terms of the solution sequence to an associated Riccati difference equation or in terms of a solution to an associated algebraic Riccati equation (see the textbook by Kwakernaak and Sivan [24]). The Riccati-equation approach to stabilizability has been extended by Zabczyk [32] to linear infinite-dimensional discrete-time systems given by bounded linear operators on Hilbert spaces. In the first part of this paper, we apply the Riccati-operator approach to the problem of stabilizing a discrete-time system (F, G) defined over an arbitrary commutative normed algebra B_0. Initial results on this problem were obtained by Kamen [19] and Green and Kamen [12], [13]. Byrnes [3], [4], [5] has developed a Riccati-operator approach to the stabilizability of linear continuous-time systems defined over a Banach or Fréchet algebra. One of the key results in the work of Byrnes is the conclusion that "locally" controllable systems are "globally" stabilizable. There remains the important question of when local stabilizability (which is much weaker than local controllability) is equivalent to global stabilizability. We answer this question below for discrete-time systems, leaving the continuous-time case for future work.

Thus the plan for this paper is as follows. Section 2 consists of preliminaries. In § 3, we associate to each system (F, G) over B_0 a transformed system $(\hat{F}, \hat{G})$; here $\hat{}$ denotes the Gelfand transform, which generalizes to arbitrary B_0 the familiar Fourier transform. Using Zabczyk's results, we give necessary and sufficient conditions for stabilizability of (F, G) in terms of the stabilizability of $(\hat{F}, \hat{G})$. Indeed, we show that if the image of the completion of B_0 under the Gelfand transform is closed under complex conjugation, then (F, G) is stabilizable with respect to B_0 if and only if $(\hat{F}, \hat{G})$ is stabilizable with respect to $\mathscr{C}(X)$, where X is the Gelfand carrier space of B_0. In § 4, we study the notion of local stabilizability, which is characterized by a local rank criterion. It is shown that local stabilizability is always equivalent to (global) stabilizability of the Gelfand-transformed system. An example is given to illustrate how the local rank criterion for stabilizability can be used as a test in the case of linear systems whose coefficients depend on parameters. Section 5 completes our discussion of the seismic cable, and § 6 consists of concluding remarks.

We shall find it desirable to consider complex algebras as well as real algebras, both as a technical convenience and for the sake of wider applicability. To avoid repetition, we shall whenever possible assume that B_0 is an algebra over $\mathbb{K}$, where $\mathbb{K}$ denotes either the field $\mathbb{R}$ of real numbers or the field $\mathbb{C}$ of complex numbers.

2. Preliminaries. Given a positive integer n and a normed $\mathbb{K}$-algebra A with norm $\|\cdot\|$, we shall let A^n denote the normed $\mathbb{K}$-linear space consisting of all n-element column vectors with entries in A and with the norm

$$\|x\|_p = \left[\sum_{i=1}^{n} \|x_i\|^p \right]^{1/p},$$

where x_i is the ith component of the vector $x \in A^n$ and p is a fixed real number with $1 \leqq p < \infty$. We shall let $M_n(A)$ denote the normed $\mathbb{K}$-algebra consisting of all $n \times n$

matrices over A with the norm

$$\|P\| = \sup\{\|Px\|_p : x \in A^n, \|x\|_p = 1\}, \tag{2.1}$$

where Px is the usual action of a matrix on a column vector. It is not difficult to check that a sequence $\{P_k\}$ in $M_n(A)$ converges to $P \in M_n(A)$ if and only if for each i and each j, the (i, j)th entry of P_k converges to the (i, j)th entry of P.

Now as in the Introduction, let B_0 denote a fixed commutative normed $\mathbb{K}$-algebra with identity 1. The completion of B_0 will be denoted by B. If $\mathbb{K} = \mathbb{C}$, we put $B_{\mathbb{C}} = B$. If $\mathbb{K} = \mathbb{R}$, we let $B_{\mathbb{C}}$ denote any complexification of B (as in [1] or [31]). Up to renorming with an equivalent norm (which has no effect on stabilizability), we may identify any two complexifications of B. In particular, we may identify $M_n(B_{\mathbb{C}})$ and $(M_n(B))_{\mathbb{C}}$.

Let X denote the carrier space of $B_{\mathbb{C}}$; that is, X is the set of all nonzero algebra homomorphisms from $B_{\mathbb{C}}$ into $\mathbb{C}$ with the weak* topology. As is well known [31, pp. 110–114], X is a compact Hausdorff space. Let $\mathscr{C}(X)$ denote the commutative C^*-algebra consisting of all continuous complex-valued functions defined on X. The involution $f \to f^*$ on $\mathscr{C}(X)$ is defined by $f^*(x) = \overline{f(x)}$, where "bar" denotes the complex conjugate. The *Gelfand transform* of an element $b \in B$ (or $b \in B_{\mathbb{C}}$) is the element $\hat{b}$ of $\mathscr{C}(X)$ defined by $\hat{b}(x) = x(b)$, $x \in X$. The *Gelfand transformation* from $B_{\mathbb{C}}$ into $\mathscr{C}(X)$ is the norm decreasing (and hence continuous) algebra homomorphism defined by $b \to \hat{b}$.

The Gelfand transform $\hat{T}$ of a $n \times m$ matrix $T = (t_{ij})$ with entries $t_{ij} \in B$ (or $B_{\mathbb{C}}$) is defined componentwise; i.e., $\hat{T} = (\hat{t}_{ij})$. For each $x \in X$, we shall let $\hat{T}(x)$ denote the $n \times m$ matrix over $\mathbb{C}$ given by $\hat{T}(x) = (\hat{t}_{ij}(x))$. Given a $n \times n$ matrix $T \in M_n(B)$, the *spectrum* Sp T of T is defined by

$$\text{Sp } T = \{\lambda \in \mathbb{C} : \lambda I - T \text{ is not invertible in } M_n(B_{\mathbb{C}})\}.$$

Letting $\lambda_i(\hat{T}(x))$, $i = 1, 2, \cdots, n$, denote the eigenvalues of $\hat{T}(x)$, $x \in X$, by [20, Prop. 1, p. 589] we have the following characterization of Sp T:

$$\text{Sp } T = \{\lambda_i(\hat{T}(x)) : i = 1, 2, \cdots, n, x \in X\}. \tag{2.2}$$

The *spectral radius* $\rho(T)$ of $T \in M_n(B)$ is defined by $\rho(T) = \sup\{|\lambda| : \lambda \in \text{Sp } T\}$.

As defined in the Introduction, a system over the $\mathbb{K}$-algebra B_0 is a pair (F, G) consisting of a $n \times n$ matrix F over B_0 and a $n \times m$ matrix G over B_0. The *Gelfand transform of the system* (F, G) is defined to be the system $(\hat{F}, \hat{G})$, where $\hat{F}$ (resp. $\hat{G}$) is the Gelfand transform of the matrix F (resp. G). By definition, the Gelfand transform $(\hat{F}, \hat{G})$ is a system over the commutative C^*-algebra $\mathscr{C}(X)$.

We now define a notion of stability which was studied in [20], [21]. The system (F, G), or the matrix F, is said to be *uniformly asymptotically stable* (u.a.s.) if $F^k \to 0$ in $M_n(B)$ as $k \to \infty$. The system $(\hat{F}, \hat{G})$, or the matrix $\hat{F}$, is u.a.s. if $(\hat{F})^k \to 0$ in $M_n(\mathscr{C}(X))$ as $k \to \infty$. Here the term "uniform" refers to the fact that we are considering asymptotic stability with a uniformity condition on the initial state. More precisely, u.a.s. of (F, G) is equivalent to

$$\sup_{\|x\|_p \leq 1, x \in B^n} \|F^k x\|_p \to 0 \quad \text{as } k \to \infty. \tag{2.3}$$

Viewing $F \in M_n(B)$ as a bounded linear operator on B^n, we can show (see [20, p. 587]) that u.a.s. is equivalent to asymptotic stability defined by $\|F^k x\|_p \to 0$ for every x in B^n. However, for arbitrary bounded linear operators on B^n, it is not true in general that u.a.s. and asymptotic stability are equivalent. For a counterexample, see [20, p. 600], where a sequence $\{F_k\}$ satisfies $F_k x \to 0$ for all x, yet $\|F_k\| \not\to 0$.

As shown in [20, Thms. 1 & 2, pp. 588–589], u.a.s. of (F, G) is equivalent to each of the following conditions:

(2.4) (a) $\rho(F)<1$,

(2.5) (b) there is a positive integer q such that $\|F^q\|<1$,

(2.6) (c) $\sum_{k=0}^{\infty} \|F^k\|<\infty$,

(2.7) (d) $|\lambda_i(\hat{F}(x))|<1, \quad i=1,2,\cdots,n, \quad$ all $x\in X$.

Using the equivalence between u.a.s. and condition (2.7), we get the following result.

PROPOSITION 1. *The system* (F, G) *is* u.a.s. *if and only if the Gelfand transform* $(\hat{F}, \hat{G})$ *is* u.a.s.

It should be mentioned that u.a.s. of (F, G) is equivalent to other notions of stability which have been studied in the literature. One such notion is l^p-stability (Przyluski [30]) defined by

$$\left[\sum_{k=0}^{\infty} \|F^k x\|_p^p\right]^{1/p} <\infty \qquad \text{for all } x\in B^n. \tag{2.8}$$

The equivalence between the condition $\rho(F)<1$ and (2.8) follows directly from the work of Zabczyk [32, § 5, pp. 727–728]. (Zabczyk considers stability of a bounded linear operator on an abstract Banach space.) Another notion of stability which is equivalent to u.a.s. is power stability as defined in the work of Przyluski [30].

Now let A be a commutative normed $\mathbb{K}$-algebra containing the $\mathbb{K}$-algebra B_0. The system (F, G) defined over B_0 is said to be *stabilizable with respect to* A if there exists a $m\times n$ matrix L over A such that the closed-loop system $(F-GL, G)$ is u.a.s. as a system over A. Since $(F-GL)\hat{}=\hat{F}-\hat{G}\hat{L}$, Proposition 1 implies that (F, G) is stabilizable with respect to A if and only if $(\hat{F}, \hat{G})$ is stabilizable with respect to $\hat{A}$, where $\hat{A}$ is the image of A under the Gelfand transformation. Although we are interested in stabilizability of (F, G) with respect to B_0, we will see that it can be helpful to consider first stabilizability of (F, G) with respect to an algebra A containing B_0. An example of an algebra A for which stabilizability with respect to A implies stabilizability with respect to B_0 is given in the following result.

PROPOSITION 2. *Suppose that* B_0 *is a dense subalgebra of* A. *Then a system* (F, G) *over* B_0 *is stabilizable with respect to* B_0 *if and only if it is stabilizable with respect to* A.

Proof. Suppose there exists an L over A such that $F-GL$ is u.a.s. Then by (2.5), there exists an integer q such that $\|(F-GL)^q\|<1$, and since B_0 is dense in A, we can find a matrix $\tilde{L}$ over B_0 such that $\|(F-G\tilde{L})^q\|<1$. Again by (2.5), $\rho(F-G\tilde{L})<1$, and thus (F, G) is stabilizable with respect to B_0. (The proposition also follows from the continuity properties of the spectrum.)

3. Stabilizability and the Riccati operator. In this section we shall study the stabilizability of a system (F, G) over B_0 in terms of the Riccati operator. We begin by considering commutative *-algebras and the notion of positivity.

Let A be a fixed commutative Banach $\mathbb{K}$-algebra with a continuous involution $a\to a^*$ and with identity 1. We also require that the involution be hermitian; that is, every hermitian element $(a=a^*)$ in A has a real spectrum. (See [31, p. 184].) For example, A could be the C^*-algebra $\mathcal{C}(X)$, where X is a compact Hausdorff space. The involution on A can be extended to the algebra $M_n(A)$ of $n\times n$ matrices over A by defining $P^*=(p_{ij})^*=(p_{ji}^*)$. With this involution and the matrix norm (2.1), $M_n(A)$

is a Banach*-algebra (over $\mathbb{K}$) whose involution is continuous and hermitian. In addition, if A is a C^*-algebra, then $M_n(A)$ is a C^*-algebra under a norm which is equivalent to the matrix norm (2.1).

A hermitian element $P \in M_n(A)$ is said to be *positive semidefinite*, denoted by $P \geqq 0$, if $\lambda \geqq 0$ for every $\lambda \in \operatorname{Sp} P$. If $P \in M_n(A)$ is hermitian and if $\lambda > 0$ for every $\lambda \in \operatorname{Sp} P$, P is said to be *positive definite*, and we write $P > 0$. Note that by (2.2), a hermitian element $P \in M_n(A)$ is positive semidefinite (resp. definite) if and only if

$$\lambda_i(\hat{P}(x)) \geqq 0 \text{ (resp. } \lambda_i(\hat{P}(x)) > 0), \quad i = 1, 2, \cdots, n, \quad \text{all } x \in X, \tag{3.1}$$

where X is the carrier space of A_C and $\hat{P}$ is the Gelfand transform of P. If $P > 0$, P has a positive definite inverse P^{-1} belonging to $M_n(A)$. Using (3.1), one can show that if $P \geqq 0$ and $R > 0$, then $T^*PT + R > 0$ for any $T \in M_n(A)$.

Let V denote the subset of $M_n(A)$ consisting of all hermitian positive semidefinite elements of $M_n(A)$. Given hermitian positive definite matrices $Q \in M_n(A)$ and $R \in M_m(A)$ and a pair (D, E) of $n \times n$, $n \times m$ matrices over A, the *Riccati operator on V associated with (Q, R) and (D, E)* is defined by

$$\mathcal{R}(P) = Q + D^*PD - D^*PE(E^*PE + R)^{-1}E^*PD, \qquad P \in V,$$

where $D^* = (d_{ij})^* = (d_{ji}^*)$. Since $P \geqq 0$ and $R > 0$, $E^*PE + R > 0$, and thus $E^*PE + R$ has a positive definite inverse in $M_n(A)$. Hence $\mathcal{R}(P)$ is an element of $M_n(A)$ for every $P \in V$. Now if we define

$$L = (E^*PE + R)^{-1}E^*PD,$$

it is easily checked that

$$\mathcal{R}(P) = Q + L^*RL + (D - EL)^*P(D - EL). \tag{3.2}$$

Since $Q > 0$, it follows from (3.2) that $\mathcal{R}(P)$ is a hermitian positive definite element of $M_n(A)$ for every $P \in V$. In fact, $\mathcal{R}(P) \geqq Q$; that is, $\mathcal{R}(P) - Q \geqq 0$.

Using a matrix-inversion identity, we can also rewrite $\mathcal{R}(P)$ in the form

$$\mathcal{R}(P) = Q + D^*P[I + E(R^{-1})E^*P]^{-1}D. \tag{3.3}$$

The expression (3.3) for the Riccati operator is the one utilized by Zabczyk in his work [32].

The *Riccati difference equation* (RDE) associated with (Q, R) and (D, E) is defined by

$$P_{j+1} = \mathcal{R}(P_j), \qquad j = 0, 1, 2, \cdots,$$
$$P_0 = 0.$$

Note that since $\mathcal{R}(P) > 0$ for all $P \in V$, the solution sequence $\{P_j\}$ to the RDE is a sequence of positive definite matrices belonging to $M_n(A)$.

The *algebraic Riccati equation* (ARE) associated with (Q, R) and (D, E) is defined by

$$P = \mathcal{R}(P).$$

Now given a system (F, G) over B_0 as defined previously, we shall first consider stabilizability of the Gelfand transform $(\hat{F}, \hat{G})$ defined over the commutative C^*-algebra $\mathcal{C}(X)$. As in the previous section, X is the carrier space of B_C, where B is the completion of B_0. Choose $Q \in M_n(B)$ and $R \in M_m(B)$ so that $\hat{Q}$ is a hermitian positive definite element of $M_n(\mathcal{C}(X))$ and $\hat{R}$ is a hermitian positive definite element

of $M_m(\mathscr{C}(X))$ (for example, we could take Q and R to be the identity matrices). Let $S_{j+1}=\mathscr{R}(S_j)$ and $S=\mathscr{R}(S)$ denote the *RDE* and *ARE* associated with $(\hat{Q},\hat{R})$ and $(\hat{F},\hat{G})$. We then have the following result which follows from the work of Zabczyk [32].

THEOREM 1. *The following conditions are equivalent.*

(a) *The Gelfand transform* $(\hat{F},\hat{G})$ *is stabilizable with respect to* $\mathscr{C}(X)$.

(b) *The* ARE $S=\mathscr{R}(S)$ *has a hermitian positive semidefinite solution* $S\in M_n(\mathscr{C}(X))$.

(c) *The solution sequence* $\{S_j\}$ *to the* RDE $S_{j+1}=\mathscr{R}(S_j)$ *converges in norm in* $M_n(\mathscr{C}(X))$ *to a hermitian positive semidefinite* S, *and* S *is a solution to the* ARE.

Further, if either (a) *or* (b) *holds, a stabilizing feedback over* $\mathscr{C}(X)$ *for* $(\hat{F},\hat{G})$ is

$$L=(\hat{G}^*S\hat{G}+\hat{R})^{-1}\hat{G}^*S\hat{F}. \tag{3.4}$$

Proof. Since $\mathscr{C}(X)$ is a C^*-algebra, by the Gelfand–Naimark theorem [1, Thm. 10, p. 209], there exists a complex Hilbert space H and an isometric *-isomorphism of $\mathscr{C}(X)$ onto a C^*-subalgebra of the algebra $\mathscr{B}(H)$ of bounded linear operators from H into H. If we take the space H^n of n-element column vectors as the state space and the space H^m of m-element column vectors as the input space, the Gelfand transform $(\hat{F},\hat{G})$ defines a linear infinite-dimensional discrete-time system given by $x_{k+1}=\hat{F}x_k+\hat{G}u_k$, where $x_k\in H^n$, $u_k\in H^m$, and where $\hat{F}$ (resp. $\hat{G}$) is viewed as a bounded linear operator from H^n into H^n (resp. from H^m into H^n). By Theorem 6.2 in Zabczyk [32, p. 729], there exists a bounded linear operator $L: H^n\to H^m$ such that $\rho(\hat{F}-\hat{G}L)<1$ if and only if the associated algebraic Riccati equation has a positive semidefinite solution. Since $\hat{F}$ and $\hat{G}$ are over $\mathscr{C}(X)$, using norm convergence of the solution sequence to the Riccati difference equation as given in [32], we have that the existence of a positive semidefinite solution to the algebraic Riccati equation is equivalent to the existence of a stabilizing operator over $\mathscr{C}(X)$. So the proof of the theorem follows by application of the results of Zabczyk.

COROLLARY. *Let* W *be any real (or complex)***-subalgebra of* $\mathscr{C}(X)$ *such that* W *contains the image* $\hat{B}_0$ *of* B_0 *under the Gelfand transformation. Then the following conditions are equivalent.*

(a) $(\hat{F},\hat{G})$ *is stabilizable with respect to* W.

(b) $(\hat{F},\hat{G})$ *is stabilizable with respect to* $\mathscr{C}(X)$.

(c) $(\hat{F},\hat{G})$ *is stabilizable with respect to a* C^**-algebra containing* $\mathscr{C}(X)$ *as a* C^**-subalgebra (i.e. as a norm closed complex* **-subalgebra).*

Proof. The implications $(a)\Rightarrow(b)\Rightarrow(c)$ are obvious, since every complex algebra is also a real algebra. Suppose then that $\mathscr{C}(X)$ is a C^*-subalgebra of a C^*-algebra A and that $(\hat{F},\hat{G})$ is stabilizable with respect to A. The closure $\bar{W}$ of W in $\mathscr{C}(X)$ is a closed real *-subalgebra of $\mathscr{C}(X)$ containing $\hat{B}$. (The Gelfand transform $\hat{}: B\to\mathscr{C}(X)$ is continuous, and B_0 is dense in B.) Since $\hat{F}$, $\hat{G}$, $\hat{R}$, and $\hat{Q}$ are matrices over $\hat{B}$, the solution sequence $\{S_j\}$ to the RDE $S_{j+1}=\mathscr{R}(S_j)$ is over $\bar{W}$. Again by Zabczyk's results [32], $\{S_j\}$ must converge in norm in $M_n(A)$, and thus $\{S_j\}$ must converge in norm to a positive semidefinite element S of $M_n(\bar{W})$. In particular, the stabilizing feedback (3.4) has entries from $\bar{W}$. By Proposition 2, $(\hat{F},\hat{G})$ is stabilizable with respect to W, and we have $(c)\Rightarrow(a)$.

Let us now consider stabilizability of (F, G) with respect to B_0. First, it follows from Proposition 1 that stabilizability of (F, G) with respect to B_0 implies that the Gelfand transform $(\hat{F},\hat{G})$ is stabilizable with respect to $\hat{B}$. Hence the conditions in Theorem 1 are necessary conditions for stabilizability of (F, G) with respect to B_0. However, these conditions are *not* in general sufficient for stabilizability with respect to B_0. A counterexample is given in the next section. For a large class of algebras, it

turns out that the conditions in Theorem 1 are necessary and sufficient for stabilizability. In particular, we have the following result.

THEOREM 2. *Let * denote the involution on $\mathscr{C}(X)$. Suppose that either $\hat{B}_0$ or $\hat{B}$ is a *-subalgebra of $\mathscr{C}(X)$. Then (F, G) is stabilizable with respect to B_0 if and only if the (equivalent) conditions in Theorem 1 are satisfied.*

Proof. Apply the corollary to Theorem 1 with $W = \hat{B}_0$ or with $W = \hat{B}$. In the latter case, apply Proposition 2.

Now suppose that the completion B of B_0 admits a hermitian involution. It follows easily that $a + ib \to a^* - ib^*$ is a hermitian involution on $B_{\mathbb{C}}$, and hence (see [1, p. 188]) that the restriction of the Gelfand transformation to B is a *-homomorphism from B into $\mathscr{C}(X)$. Thus $\hat{B}$ is a *-*subalgebra* of $\mathscr{C}(X)$. In particular, the hypothesis of Theorem 2 is satisfied, and we have the following result.

COROLLARY. *If the completion B admits a hermitian involution, then (F, G) is stabilizable with respect to B_0 if and only if the conditions in Theorem 1 are satisfied.*

Both of the examples $B_0 = B = l^1(\mathbb{Z}, \mathbb{R})$ and $B_0 = B = \mathscr{C}(\Omega, \mathbb{R})$ considered in the Introduction have continuous hermitian involutions. In particular, for $l^1(\mathbb{Z}, \mathbb{R})$ the involution is given by $\alpha(k)^* = \alpha(-k)$, $k \in \mathbb{Z}$, and for $\mathscr{C}(\Omega, \mathbb{R})$ the involution is the identity map. So by the corollary to Theorem 2, stabilizability of a system over either of these algebras is equivalent to the conditions in Theorem 1.

Let us now assume that the completion B admits a continuous hermitian involution $b \to b^*$. Then we can consider the positivity of matrices over B, as we did in the first part of this section. Given hermitian positive definite matrices $Q \in M_n(B)$ and $R \in M_m(B)$, let $P_{j+1} = \mathscr{R}(P_j)$ denote the RDE associated with (Q, R) and the system (F, G) defined over B_0. As noted previously, the solution sequence $\{P_j\}$ to the RDE is a sequence of positive definite matrices belonging to $M_n(B)$. By the Corollary to Theorem 2, the system (F, G) is stabilizable with respect to B_0 if and only if the Gelfand transform sequence $\{\hat{P}_j\}$ converges in norm in $M_n(\mathscr{C}(X))$ to some positive semidefinite $T \in M_n(\mathscr{C}(X))$. If $\{\hat{P}_j\}$ does converge to T in norm, the matrix $L = (\hat{G}^* T \hat{G} + \hat{R})^{-1} \hat{G}^* T \hat{F}$ is a stabilizing matrix for the Gelfand transform $(\hat{F}, \hat{G})$. Using the same type of argument as in the proof of Proposition 2, we can show that there exists a positive integer q such that

$$\hat{L}_q = (\hat{G}^* \hat{P}_q \hat{G} + \hat{R})^{-1} \hat{G}^* \hat{P}_q \hat{F}$$

is also a stabilizing matrix for $(\hat{F}, \hat{G})$. Since $*: \mathrm{B} \to \mathrm{B}$ is hermitian, it is easy to check that the Gelfand transform on $M_n(B)$ is a *-homomorphism. Thus $\hat{L}_q$ is the Gelfand transform of $L_q = (G^* P_q G + R)^{-1} G^* P_q F$, and thus by Proposition 1, the matrix L_q stabilizes (F, G). Hence it is possible to compute a stabilizing feedback over B without having to compute the limit of the sequence $\{\hat{P}_j\}$.

To illustrate some of the interesting aspects of the stability criteria given above, let us consider the case $n = m = 1$; that is $F = f \in B_0$ and $G = g \in B_0$. We also assume that $B_0 = B$ and that B admits a continuous hermitian involution. Choosing $Q = R = 1$, the ARE associated with (Q, R) and (f, g) is given by

$$p = \mathscr{R}(p) = 1 + f^* p (1 + g(g^*) p)^{-1} f. \tag{3.5}$$

Rewriting (3.5) using commutativity of B, we have

$$(gg^*)(p^2) + (1 - f^* f) p - 1 = 0. \tag{3.6}$$

Taking the Gelfand transform of both sides of (3.6), we have

$$|\hat{g}(x)|^2 \hat{p}(x)^2 + (1 - |\hat{f}(x)|^2) \hat{p}(x) - 1 = 0, \qquad x \in X. \tag{3.7}$$

Let $K=\{x\in X:\hat{g}(x)=0\}=\{x\in X:(gg^*)\hat{}(x)=0\}$. If $\hat{p}\in\mathscr{C}(X)$ is a nonnegative solution to (3.7), then we must have

$$\hat{p}(x)=\frac{-(1-|\hat{f}(x)|^2)+\sqrt{(1-|\hat{f}(x)|^2)^2+4|\hat{g}(x)|^2}}{2|\hat{g}(x)|^2} \tag{3.8a}$$

if $x\notin K$ and

$$\hat{p}(x)=(1-|\hat{f}(x)|^2)^{-1} \tag{3.8b}$$

if $x\in K$. It follows immediately that (3.7) has a positive semidefinite solution $\hat{p}$ in $\mathscr{C}(X)$ only if (f, g) satisfies

$$|\hat{f}(x)|<1 \quad \text{for all } x\in X \text{ such that } \hat{g}(x)=0. \tag{3.9}$$

Conversely, suppose that (3.9) holds and put $t=t(x)=|\hat{g}(x)|(1-|\hat{f}(x)|^2)^{-1}$ whenever $|\hat{f}(x)|<1$. Clearly (3.8a, b) gives an everywhere positive solution $\hat{p}$ to (3.7), and moreover

$$\hat{p}=\frac{1}{1-|\hat{f}(x)|^2}\frac{(1+4t^2)^{1/2}-1}{2t^2}$$

on an open subset of X which contains K. Since $t(x)\to 0$ as x approaches the set K, an application of L'Hospital's rule shows that $\hat{p}$ is continuous on X. We have thus shown that the ARE (3.7) has a positive semidefinite solution (given by (3.8a) and (3.8b)) in $\mathscr{C}(X)$ if and only if condition (3.9) is satisfied, in which case the solution is unique and positive definite. By the Corollary to Theorem 2, condition (3.9) is also equivalent to stabilizability of (f, g) with respect to B. Note that if $(gg^*)\hat{}$ is never zero on X, then the element $b=(1-f^*f)^2+4gg^*$ of B can be shown to have a positive definite square root β in B. Thus if $\hat{g}$ never vanishes on X, then condition (3.9) is vacuously satisfied and the ARE (3.5) has the positive definite solution

$$p=(-1+f^*f+\beta)(2gg^*)^{-1}$$

in B. Of course, (f, g) must then be stabilizable with respect to B, and indeed $L=g^{-1}f\in B$ is a stabilizing feedback.

It is worth noting that the arguments above go through even when $m>1$, provided that $n=1$ and that we replace $|\hat{g}(x)|^2$ by $(gg^*)\hat{}(x)$. In the case when $n>1$ and B admits a hermitian involution, it is still true (by the Corollary to Theorem 2) that stabilizability with respect to B is equivalent to the existence of a positive semidefinite solution in $M_n(\mathscr{C}(X))$ to the transformed ARE. However, even when GG^* is invertible, it remains unclear whether this is also equivalent to the existence of a positive semidefinite solution in $M_n(B)$ to the ARE (assuming that B is not a C^*-algebra). It is thus fortunate that stabilizability with respect to B can be reduced to stabilizability with respect to $\mathscr{C}(X)$, especially since this latter type of stabilizability can be checked pointwise on X. (See Theorem 4 in the next section.) The fact that one can more easily demonstrate the existence of a solution to the ARE in the Gelfand transform domain is of course a reflection of the very special tractability of $\mathscr{C}(X)$ (or more generally of any C^*-algebra) in comparison to the much broader class of Banach algebras with continuous hermitian involutions.

4. Local stabilizability. The conditions for stabilizability derived in the previous section are all expressed in terms of a solution to an associated Riccati equation, and thus it is necessary to solve the Riccati equation in order to test for stabilizability. For finite-dimensional systems given by a pair (F, G) of matrices over the reals $\mathbb{R}$, it is

well known that one can test for stabilizability without having to solve the associated Riccati equation. In particular, there is the Hautus stabilizability criterion [14] given by

$$\text{rank}\,[zI - F \quad G] = n, \qquad \text{all } z \in \mathbb{C} : |z| \geqq 1. \tag{4.1}$$

In the first part of this section we show that stabilizability of the Gelfand transform $(\hat{F}, \hat{G})$ of a system (F, G) over B_0 can be checked by applying the Hautus criterion "point-by-point". We begin with the notion of local stabilizability. As in the previous sections, B_0 is a commutative normed $\mathbb{K}$-algebra with identity. We do not require that the completion B admit an involution until later.

DEFINITION. A system (F, G) over B_0 is said to be *locally stabilizable* if for each $x \in X$, the system $(\hat{F}(x), \hat{G}(x))$ defined over $\mathbb{C}$ is stabilizable with respect to $\mathbb{C}$; that is, if for each $x \in X$, there is a matrix L_x over $\mathbb{C}$ such that $\hat{F}(x) - \hat{G}(x)L_x$ has spectral radius strictly less than one.

It is clear that any system (F, G) that is stabilizable with respect to B_0 (or B) is locally stabilizable. By applying point-by-point the Hautus criterion (4.1), generalized to finite-dimensional systems over $\mathbb{C}$, we have that local stability of (F, G) is equivalent to

$$\text{rank}\,[zI - \hat{F}(x) \quad \hat{G}(x)] = n, \qquad \text{all } |z| \geqq 1 \text{ and all } x \in X. \tag{4.2}$$

It is interesting to note that when $n = m = 1$, local stabilizability is equivalent to requiring that $|\hat{F}(x)| < 1$ for all $x \in X$ such that $\hat{G}(x) = 0$, which is the same as condition (3.9) derived from the Riccati-equation approach.

THEOREM 3. *The Gelfand transform $(\hat{F}, \hat{G})$ is stabilizable with respect to $\mathscr{C}(X)$ if and only if (F, G) is locally stabilizable.*

Proof. The only difficult part of the proof is showing that local stabilizability implies that the Gelfand transform is stabilizable with respect to $\mathscr{C}(X)$. To prove this, we will first show that local stabilizability implies that $(\hat{F}, \hat{G})$ is stabilizable with respect to $l^\infty(X)$, where $l^\infty(X)$ is the commutative C^*-algebra consisting of *all* bounded functions from X into $\mathbb{C}$ with the sup norm. For each $x \in X$, choose L_x, a matrix over $\mathbb{C}$, such that $\rho(\hat{F}(x) - \hat{G}(x)L_x) < 1$. Choose a positive integer $k_x \geqq 1$ such that

$$\|(\hat{F}(x) - \hat{G}(x)L_x)^{k_x}\| < 1.$$

Let $\varepsilon_x > 0$ be chosen so that

$$\|(\hat{F}(x) - \hat{G}(x)L_x)^{k_x}\| < 1 - \varepsilon_x.$$

Then $\{O_x : x \in X\}$, where

$$O_x = \{y \in X : \|(\hat{F}(y) - \hat{G}(y)L_x)^{k_x}\| < 1 - \varepsilon_x\},$$

is an open cover for X. Since X is compact, there exist $x_1, x_2, \cdots, x_l$ such that $X \subseteq \bigcup_{i=1}^{l} O_{x_i}$. Put $k_{x_i} = k_i$, $O_{x_i} = O_i$, and $L_{x_i} = L_i$ where $i = 1, 2, \cdots, l$. Define L as follows:

$$L = L_1 \text{ on } O_1, \quad L = L_2 \text{ on } O_2 \backslash O_1, \quad L = L_3 \text{ on } O_3 \backslash (O_1 \cup O_2), \quad \text{etc.}$$

Since each entry of L is piecewise constant on X, clearly each entry of L lies in $l^\infty(X)$. It remains only to show that L is a stabilizing feedback for (F, G) with respect to $l^\infty(X)$. To do this, it suffices to show that $(\hat{F} - \hat{G}L)^k \to 0$ in $M_n(l^\infty(X))$. Let $\varepsilon = \min(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_l)$, and let

$$S = \sup_{\substack{1 \leqq i \leqq l \\ 0 \leqq j \leqq k_i - 1}} \left\{ \sup_{x \in X} \|\hat{F}(x) - \hat{G}(x)L_i\|^j \right\}.$$

Since X is compact and $x \to \|\hat{F}(x) - \hat{G}(x)L_i\|^j$ is continuous for each fixed i, j, we have $S < \infty$. Let $x \in X$. Then for some i with $1 \leqq i \leqq l$, we have $x \in O_i$ *and* $L(x) = L_i$. Thus for all q and $j = 0, 1, 2, \cdots, k_i - 1$, we have

$$\|(\hat{F}(x) - \hat{G}(x)L(x))^{qk_i+j}\| \leqq \|(\hat{F}(x) - \hat{G}(x)L_i)^{k_i}\|^q \|\hat{F}(x) - \hat{G}(x)L_i\|^j \leqq (1-\varepsilon)^q S.$$

It follows that $\sup_{x \in X} \|(\hat{F}(x) - \hat{G}(x)L(x))^k\| \to 0$ as $k \to \infty$, since $(1-\varepsilon)^q S$ is independent of x. But this last supremum is the norm of $(\hat{F} - \hat{G}L)^k$ in the C^*-algebra $M_n(l^\infty(X))$. Hence $(\hat{F}, \hat{G})$ is stabilizable with respect to $l^\infty(X)$, and by the corollary to Theorem 1, $(\hat{F}, \hat{G})$ is stabilizable with respect to $\mathscr{C}(X)$.

Combining Theorems 2 and 3, we get the following central result.

THEOREM 4. *If $\hat{B}_0$ or $\hat{B}$ is a *-subalgebra of $\mathscr{C}(X)$ (e.g. the latter is true if B admits a hermitian involution), then (F, G) is stabilizable with respect to B_0 if and only if (F, G) is locally stabilizable.*

The following example shows that if $\hat{B}$ is not *-closed, the equivalence stated in Theorem 4 can fail.

Example 1. Let $B_0 = B$, where B is the disc algebra; that is, B is the algebra of all continuous complex-valued functions defined on the closed unit disc $\bar{\Delta}$ in $\mathbb{C}$ which are analytic on the open unit disc Δ. In this case, we may identify B with $B_{\mathbb{C}}$, X with $\bar{\Delta}$, and the Gelfand transform with the identity map. Now consider the system (f, g), where $f(x) = x$ and $g(x) = x^2$, $x \in \bar{\Delta}$. Since the only zero of $z - f(x) = z - x$ is at $z = x$, and $g(x) = 0$ if and only if $x = 0$, the rank criterion (4.2) is satisfied, and thus (f, g) is locally stabilizable. By Theorem 3, the Gelfand transform (f, g) is stabilizable with respect to $\mathscr{C}(\bar{\Delta})$. Now suppose that (f, g) is stabilizable with respect to B, so there exists an $L \in B$ such that $\rho(f - gL) < 1$. By the maximum-modulus theorem, we have (letting T = unit circle in $\mathbb{C}$)

$$1 > \sup_{x \in \bar{\Delta}} |x - x^2 L(x)| = \sup_{x \in T} |x - x^2 L(x)| = \sup_{x \in T} |1 - xL(x)|$$

$$= \sup_{x \in \bar{\Delta}} |1 - xL(x)| \geqq |1 - (0)L(0)| = 1.$$

We have a contradiction, and thus there is no stabilizing feedback belonging to B.

This last example shows that local stabilizability is not in general equivalent to stabilizability with respect to B. It also shows that stabilizability of $(\hat{F}, \hat{G})$ with respect to $\mathscr{C}(X)$ is not in general equivalent to stabilizability of (F, G) with respect to B. Of course, these equivalences fail in this example only because the disc algebra is not a *-subalgebra of $\mathscr{C}(\bar{\Delta})$. Indeed, if $b(x) = x$, then b lies in the disc algebra and $b^*(x) = \overline{b(x)} = \bar{x}$ for $x \in \bar{\Delta}$. Since b^* is continuous but not analytic, $B = \hat{B}$ is not closed with respect to the involution on $\mathscr{C}(X)$.

We shall conclude this section with an example illustrating the use of the local rank criterion (4.2) as a stabilizability test for systems depending on parameters.

Example 2. Let $B_0 = B = \mathscr{C}(\Omega, \mathbb{R})$, where Ω is a compact subset of $\mathbb{R}^2$. In this case, the complexification $B_{\mathbb{C}}$ can be identified with the algebra $\mathscr{C}(\Omega, \mathbb{C})$ of complex-valued continuous functions on Ω, and the carrier space X of $B_{\mathbb{C}}$ can be identified with Ω. Thus the Gelfand transformation is the identity map on $\mathscr{C}(\Omega, \mathbb{C})$. Now consider the pair (F, G) over $\mathscr{C}(\Omega, \mathbb{R})$, where

$$F(w_1, w_2) = \begin{bmatrix} w_1 & 1 \\ w_2 & 1 \end{bmatrix} \quad \text{and} \quad G(w_1, w_2) = \begin{bmatrix} 1 \\ w_2 \end{bmatrix}.$$

As noted in the Introduction, the pair (F, G) can be interpreted as a linear discrete-time

system with parameters given by the dynamical equation

$$x(k+1) = F(w_1, w_2)x(k) + G(w_1, w_2)u(k), \tag{4.3}$$

or in component form,

$$x_1(k+1) = w_1 x_1(k) + x_2(k) + u(k),$$
$$x_2(k+1) = w_2 x_1(k) + x_2(k) - w_2 u(k).$$

Here w_1 and w_2 are parameters with (w_1, w_2) taking values from Ω. We would like to know if there is a stabilizing feedback $L(w_1, w_2)$ over $\mathscr{C}(\Omega, \mathbb{R})$ for the system (4.3). Since the identity map on $\mathscr{C}(\Omega, \mathbb{R})$ is a continuous hermitian involution, by Theorem 4 such a feedback exists if and only if

$$\operatorname{rank}[zI - F(w_1, w_2) \quad G(w_1, w_2)] = 2, \quad |z| \geqq 1, \quad (w_1, w_2) \in \Omega. \tag{4.4}$$

We have

$$[zI - F(w_1, w_2) \quad G(w_1, w_2)] = \begin{bmatrix} z - w_1 & -1 & 1 \\ -w_2 & z-1 & -w_2 \end{bmatrix}.$$

The common zeros of the three 2×2 minors of $[zI - F \;\; G]$ are $z = 1$ when $w_2 = 0$ and $z = w + 1$ when $w_1 = w_2 = w$. Thus the rank condition (4.4) is satisfied if and only if Ω does not intersect the set

$$\{(w_1, 0): w_1 \in \mathbb{R}\} \cup \{(w, w): w \in \mathbb{R}, |w+1| \geqq 1\}. \tag{4.5}$$

Therefore the system (F, G) defined over $\mathscr{C}(\Omega, \mathbb{R})$ is stabilizable with respect to $\mathscr{C}(\Omega, \mathbb{R})$ if and only if Ω does not intersect the set given by (4.5).

5. Application to the positioning of a seismic cable. Let us again consider the seismic cable which is given by the discrete-time approximation

$$x(kT + T, r) = e^{F(0)T}x(kT, r) + C_T x(kT, r-1) + C_T x(kT, r+1) + D_T u(kT, r), \tag{5.1a}$$

$$q(kT, r) = Hx(kT, r), \tag{5.1b}$$

where $H = [1 \quad 0]$, C_T and D_T are defined by (1.6), and as in the Introduction, $q(kT, r)$ is the position of the rth cable segment at time kT, $k \in \mathbb{Z}$.

Let $l^\infty(\mathbb{Z}, \mathbb{R})$ denote the Banach space of bounded real-valued functions defined on $\mathbb{Z}$ with the sup norm. We would like to know if there exists a feedback controller which brings the cable to a specified position $q_0 \in l^\infty(\mathbb{Z}, \mathbb{R})$, starting from any initial position $q(0, \cdot) \in l^\infty(\mathbb{Z}, \mathbb{R})$. In other words, we want

$$q(kT, r) \to q_0(r) \quad \text{in } l^\infty(\mathbb{Z}, \mathbb{R}) \quad \text{as } k \to \infty \tag{5.2}$$

for *any* $q_0 \in l^\infty(\mathbb{Z}, \mathbb{R})$ and *any* $q(0, \cdot) \in l^\infty(\mathbb{Z}, \mathbb{R})$. This problem can be reduced to the stabilization of an augmented system defined as follows. First, let $e(kT, r)$ denote the error between the actual position $x(kT, r)$ and the desired position $q_0(r)$ of the cable; i.e.,

$$e(kT, r) = q(kT, r) - q_0(r), \qquad r \in \mathbb{Z}.$$

Consider the error-driven system given by the scalar state equation

$$v(kT + T, r) = v(kT, r) + e(kT, r) = v(kT, r) + Hx(kT, r) - q_0(r). \tag{5.3}$$

Now combine the cable system (5.1a) with the error-driven system (5.3), which yields

the augmented system given by

$$\left[\begin{array}{c} x(kT+T,r) \\ \hline v(kT+T,r) \end{array}\right] = \left[\begin{array}{c|c} e^{F(0)T} & \begin{matrix}0\\0\end{matrix} \\ \hline H & 1 \end{array}\right] \left[\begin{array}{c} x(kT,r) \\ \hline v(kT,r) \end{array}\right] + \left[\begin{array}{cc|c} C_T & & 0 \\ & & 0 \\ \hline 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} x(kT,r-1) \\ \hline v(kT,r-1) \end{array}\right]$$

(5.4)

$$+ \left[\begin{array}{cc|c} C_T & & 0 \\ & & 0 \\ \hline 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} x(kT,r+1) \\ \hline v(kT,r+1) \end{array}\right] + \left[\begin{array}{c} D_T \\ \hline 0 \end{array}\right] u(kT,r) - \left[\begin{array}{c} 0 \\ 0 \\ \hline q_0(r) \end{array}\right].$$

Let F_a and G_a denote the coefficient matrices of the augmented system given by

$$F_a(0) = \left[\begin{array}{c|c} e^{F(0)T} & \begin{matrix}0\\0\end{matrix} \\ \hline H & 1 \end{array}\right], \qquad F_a(-1) = F_a(1) = \left[\begin{array}{cc|c} C_T & & 0 \\ & & 0 \\ \hline 0 & 0 & 0 \end{array}\right], \qquad F_a(r) = 0, \quad |r| \geqq 2,$$

$$G_a(0) = \left[\begin{array}{c} D_T \\ \hline 0 \end{array}\right], \qquad G_a(r) = 0, \quad |r| \geqq 1.$$

Clearly, F_a and G_a are defined over $l^1(\mathbb{Z}, \mathbb{R})$, so the augmented system is a discrete-time system over $B_0 = l^1(\mathbb{Z}, \mathbb{R})$.

Now suppose that the augmented system is stabilizable; i.e., there exists a three-element row vector L over $l^1(\mathbb{Z}, \mathbb{R})$ such that $F_a - G_a L$ is u.a.s. Partitioning L into the form $L = [L_1 \quad L_2]$, where L_1 is a two-element row vector vector over $l^1(\mathbb{Z}, \mathbb{R})$ and $L_2 \in l^1(\mathbb{Z}, \mathbb{R})$, consider the feedback control law

$$u(kT, r) = -\sum_{j=-\infty}^{\infty} L_1(r-j)x(kT, j) - \sum_{j=-\infty}^{\infty} L_2(r-j)v(kT, j). \tag{5.5}$$

We claim that our "tracking" objective (5.2) is satisfied with the feedback control law (5.5). In other words, the feedback controller defined by (5.3) and (5.5) solves the cable positioning problem posed above. This result is a generalization of the well-known result in the finite-dimensional case that tracking of a step function can be achieved with integral control. We omit the details.

By the above analysis, the existence of a solution to the cable positioning problem reduces to determining whether or not the augmented system (F_a, G_a) is stabilizable. Since $l^1(\mathbb{Z}, \mathbb{R})$ is a hermitian algebra, we can test for this by using the local rank criterion (4.2). In this application, the carrier space of the complexification of the algebra $l^1(\mathbb{Z}, \mathbb{R})$ can be identified with the interval $[0, 2\pi]$ and the Gelfand transform can be identified with the Fourier transform. Taking the Fourier transform of F_a and G_a, we get

$$\hat{F}_a(\omega) = \left[\begin{array}{c|c} e^{F(0)T} + (2\cos\omega)C_T & \begin{matrix}0\\0\end{matrix} \\ \hline H & 1 \end{array}\right], \qquad \hat{G}_a(\omega) = \left[\begin{array}{c} D_T \\ \hline 0 \end{array}\right].$$

Thus the augmented system (F_a, G_a) is stabilizable if and only if

$$\operatorname{rank}[zI_3 - \hat{F}_a(\omega) \quad \hat{G}_a(\omega)] = 3, \quad |z| \geqq 1, \quad \omega \in [0, 2\pi], \tag{5.6}$$

where I_3 is the 3×3 identity matrix. It is easy to see that (5.6) is equivalent to

$$\text{(5.7a)} \qquad \operatorname{rank}[zI_2 - e^{F(0)T} - (2\cos\omega)C_T \quad D_T] = 2, \quad |z| \geqq 1, \quad \omega \in [0, 2\pi]$$

and

$$\text{(5.7b)} \qquad \operatorname{rank}\left[\begin{array}{c|c|c} I_2 - e^{F(0)T} - (2\cos\omega)C_T & 0 & D_T \\ & 0 & \\ \hline -H & 0 & 0 \end{array}\right] = 3, \qquad \omega \in [0, 2\pi].$$

The criterion (5.7a) is equivalent to stabilizability of the given cable system (5.1a), while (5.7b) can be shown to be equivalent to the requirement that the transfer function for the cable does not have any zeros at $z = 1$. Hence, the cable positioning problem has a solution if the cable system (5.1a) is stabilizable and if the rank condition (5.7b) (which is a zero criterion) is satisfied.

As a particular example, if we take $M = 1$, $T = 1$, $c = 0.1$, and $k = 0.2525$, we have

$$[zI_2 - e^{F(0)T} - (2\cos\omega)C_T \quad D_T] = \begin{bmatrix} z - .8802 - .0708\cos\omega & -.9119 & .2804 \\ .2303 - .2421\cos\omega & z - .7900 & .9588 \end{bmatrix}.$$

The determinant of the 2×2 matrix formed from the first and third columns of $[zI_2 - e^{F(0)T} - (2\cos\omega)C_T \quad D_T]$ is equal to $z - .9475$. Since $.9475 < 1$, (5.7a) is satisfied and the cable is stabilizable. Now the matrix in (5.7b) is given by

$$\begin{bmatrix} .1198 - .0708\cos\omega & -.9119 & 0 & .2804 \\ .2302 - .2421\cos\omega & .2100 & 0 & .9588 \\ -1 & 0 & 0 & 0 \end{bmatrix},$$

and this clearly has rank 3 for all $\omega \in [0, 2\pi]$. Therefore, for the particular values of the system parameters selected above, the cable positioning problem has a solution. The stabilizing gain vector $L = [L_1 \quad L_2]$ could be computed using the Riccati difference equation as discussed in § 3. We shall not pursue this here.

6. Discussion. We have seen that the use of the Gelfand transform, together with a Riccati equation, yields a useful "local" criterion for stabilizability of a system defined over B_0, and we applied this criterion explicitly to the stabilization of several systems, including the long seismic cable. Some reflection on extensions of these techniques, and on other recent approaches to similar problems, is in order. As we saw in § 4, local stabilizability and stabilizability with respect to B_0 are not equivalent in general unless the norm closure B of B_0 admits a hermitian involution. We also saw that the existence of a positive semidefinite solution $S \in M_n(\mathscr{C}(X))$ to the "transformed" Riccati equation is not in general equivalent to stabilizability with respect to B_0 unless B admits a hermitian involution. Thus it appears that hermitian *-algebras are the most general framework (in the commutative case) in which the methods of this paper can be expected to work without exception. We should note that after the first version of this paper was written, further results on the stabilization of linear systems over $\mathscr{C}(\Omega, \mathbb{K})$, where Ω is an arbitrary subset of $\mathbb{K}^N$, were obtained by Kamen and Khargonekar [22].

At present, it is unclear whether topologies on commutative rings without an involution can be used to study systematically stabilization by state feedback. However, for linear systems over commutative rings without an involution, there is a theory of stabilization by *dynamic* output feedback based on a polynomial-matrix representation of the system. For results on this, see Emre [8], Khargonekar and Sontag [23] and the references in these papers.

It is possible to give a stabilization theory similar to that presented above for systems (F, G) defined over a noncommutative C^*-algebra B. In particular, for such systems there is a natural notion of local stabilizability based on the primitive ideal space of B. This notion of stabilizability can be shown to be equivalent to stabilizability over B. Results on the noncommutative case will be available in a separate paper.

Acknowledgments. The authors are indebted to the reviewers for several constructive comments, and in particular, for pointing out the related work of Zabczyk. The first-named author is also grateful to A. D. Andrew for several helpful conversations.

REFERENCES

[1] F. F. BONSAL AND J. DUNCAN, *Complete Normed Algebras*, Ergebnisse der Mathematik und ihrer Grenzgebiete 80, Springer-Verlag, New York, 1973.

[2] R. W. BROCKETT, *Finite-Dimensional Linear Systems*, John Wiley, New York, 1970.

[3] C. I. BYRNES, *On the stabilizability of linear control systems depending on parameters*, in Proc. 18th IEEE Conference on Decision and Control, Ft. Lauderdale, FL, 1979, pp. 233–236.

[4] ——— *Realization theory and quadratic optimal controllers for systems defined over Banach and Fréchet algebras*, in Proc. 19th IEEE Conference on Decision and Control, Albuquerque, NM, 1980, pp. 247–251.

[5] ———, *Algebraic and geometric aspects of the analysis of feedback systems*, in Geometrical Methods for the Theory of Linear Systems, C. I. Byrnes and C. Martin, eds., D. Reidel, Dordrecht, 1980, pp. 85–124.

[6] K. C. CHU, *Optimal decentralized regulation for a string of coupled systems*, IEEE Trans. Automatic Control, AC-19 (1974), pp. 243-246.

[7] M. L. EL-SAYED AND P. S. KRISHNAPRASAD, *Homogeneous interconnected systems: An example*, IEEE Trans. Automatic Control, AC-26 (1981), pp. 894–901.

[8] E. EMRE, *On necessary and sufficient conditions for regulation of linear systems over rings*, this Journal, 20 (1982), pp. 155–160.

[9] P. A. FUHRMANN, *On realization of linear systems and applications to some questions of stability*, Math. Systems Theory, 8 (1974), pp. 132–141.

[10]———, *Realization theory in Hilbert space for a class of transfer functions*, J. Functional Analysis, 18 (1975), pp. 338–349.

[11] ——, *Exact controllability and observability and realization theory in Hilbert space*, J. Math. Anal. Appl., 53 (1976), pp. 377–392.

[12] W. L. GREEN AND E. W. KAMEN, *Stabilizability of linear discrete-time systems defined over a commutative normed algebra*, in Proc. 19th IEEE Conference on Decision and Control, Albuquerque, NM, 1980, pp. 264–268.

[13] ———, *Necessary and sufficient conditions for stabilizability of linear discrete-time systems defined over a subalgebra of a C^*-algebra*, in Proc. International Symposium on Mathematical Theory of Networks and Systems, Santa Monica, CA, 1981, pp. 92–97.

[14] M. L. J. HAUTUS, *Stabilization controllability and observability of linear autonomous systems*, Ned. Akad. Wetenschappen, Proc., Ser. A, 73 (1970), pp. 448–455.

[15] J. W. HELTON, *Discrete time systems, operator models and scattering theory*, J. Funct. Anal., 16 (1974), pp. 15–38.

[16] ———, *A spectral factorization approach to the distributed stable regulator problem; the algebraic Riccati equation*, this Journal, 14 (1976), pp. 639–661.

[17] ———, *Systems with infinite dimensional state spaces: The Hilbert space approach*, Proc. IEEE, 60 (1976), pp. 145–160.

[18] E. W. KAMEN, *Asymptotic stability of linear shift-invariant two-dimensional digital filters*, IEEE Trans. Circuits and Systems, CAS-27 (1980), pp. 1234–1240.

[19] ———, *Linear discrete-time systems over a commutative Banach algebra with applications to two-dimensional systems*, in Algebraic and Geometric Methods in Linear Systems Theory, Lectures in Applied Mathematics 18, C. I. Byrnes and C. Martin, eds., American Mathematical Society, Providence, RI, 1980, pp. 225–237.

[20] E. W. KAMEN AND W. L. GREEN, *Asymptotic stability of linear difference equations defined over a commutative Banach algebra*, J. Math. Anal. Appl., 75 (1980), pp. 584–601.

[21] E. W. KAMEN AND W. L. GREEN, *Addendum to 'Asymptotic stability of linear difference equations defined over a commutative Banach algebra'*, J. Math. Anal. Appl., 84 (1981), pp. 295–298.

[22] E. W. KAMEN AND P. P. KHARGONEKAR, *On the control of linear systems whose coefficients are functions of parameters*, IEEE Trans. Automatic Control, AC-29 (1984), pp. 25–33.

[23] P. P. KHARGONEKAR AND E. D. SONTAG, *On the relation between stable matrix fraction factorizations and regulable realizations of linear systems over rings*, IEEE Trans. Automatic Control, AC-27 (1982), pp. 627–638.

[24] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*, John Wiley, New York, 1972.

[25] K. Y. LEE, S. CHOW AND R. BARR, *On the control of discrete-time distributed parameter systems*, this Journal, 10 (1972), pp. 361–376.

[26] S. M. MELZER AND B. C. KUO, *Optimal regulation of systems described by a countably infinite number of objects*, Automatica, 7 (1971), pp. 359–366.

[27] K. M. PRZYLUSKI, *Infinite dimensional discrete-time systems as models for linear systems with time delay*, in Proc. 2nd IFAC Symposium on Distributed Parameter Systems, Warwick, England, 1977.

[28] ———, *Arbitrary stabilizability of infinite-dimensional discrete-time systems with applications to linear hereditary control systems*, Preprint No. 212, Institute of Mathematics, Polish Academy of Sciences, Warsaw, 1980.

[29] ———, *The Lyapunov equation and the problem of stability for linear bounded discrete-time systems in Hilbert space*, Appl. Math. Optim., 6 (1980), pp. 97–112.

[30] ———, *Stability of linear infinite dimensional systems revisited*, preprint.

[31] C. E. RICKART, *General Theory of Banach Algebras*, Van Nostrand, New York, 1960.

[32] J. ZABCZYK, *Remarks on the control of discrete-time distributed parameter systems*, this Journal, 12 (1974), pp. 721–735.

[33] ———, *Stability properties of the discrete Riccati operator equation*, Kybernetica, 13 (1977), pp. 1–10.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

MEAN SQUARE STABILITY FOR DISCRETE BOUNDED LINEAR SYSTEMS IN HILBERT SPACE*

C. S. KUBRUSLY†

Abstract. The asymptotic behaviour for infinite-dimensional discrete linear systems driven by white noise is considered in this paper. Both the evolution and convergence of the state correlation operators sequence are analysed. Mean square stability conditions are investigated, including a comparison with the deterministic stability problem. The particular case of compact operators is considered in some detail.

Key words. asymptotic stability, infinite-dimensional systems, linear dynamical systems, discrete-time systems, stochastic systems

1. Introduction. Conditions for asymptotic stability of finite-dimensional discrete linear system operating either in a deterministic or stochastic environment are by now well established (cf. [10], [9]). On the other hand the same problem in an infinite-dimensional setting, which is endowed with a much richer structure, still presents some unsolved questions.

As far as the asymptotic stability problem for infinite-dimensional discrete linear deterministic systems is concerned, there is available in the current literature a fairly complete collection of results (cf. § 3). This does not seem to be the case for discrete stochastic systems, although some few results have already been investigated by using different approaches and under different motivations. For instance, the convergence analysis of stochastic approximation algorithms in Hilbert space considered in [13] and [8] actually gives asymptotic stability conditions for infinite-dimensional dynamical systems. Questions related to optimal stochastic control problems have also motivated some partial results in this direction (cf. [6], [16] and [17]).

In this paper we consider the mean square stability problem, by analysing both the evolution and asymptotic behaviour of state correlation operators, for discrete linear systems in Hilbert space. The paper is organized as follows. Notational preliminaries and basic concepts, which will be needed along the text, are considered in § 2. These comprise bounded linear transformations, positive and nuclear operators, correlation operators, and approximate controllability. A brief review on asymptotic stability for deterministic discrete systems is presented in § 3, including the auxiliary results which will be used in the sequel. The central theme of the paper appears in § 4. There it is analysed the evolution and convergence of the state correlation sequence $\{Q_i; i \geqq 0\}$ for discrete linear systems driven by white noise. The main results (cf. Lemma 2, Theorems 1, 2 and Corollary 1) deal with the relationship between convergence of $\{Q_i; i \geqq 0\}$ and the spectral radius $r_\sigma(A)$ of the system operator A. It is shown that $r_\sigma(A) < 1$ (i.e. uniform asymptotic stability for the free system) is sufficient to ensure uniform convergence of $\{Q_i; i \geqq 0\}$ to a correlation operator (i.e. mean square stability for the disturbed system). Necessary and sufficient conditions for uniform convergence of $\{Q_i; i \geqq 0\}$ to a positive correlation operator are also given, for the case of a compact system operator A.

* Received by the editors February 3, 1983, and in revised form February 3, 1984. This research was partially supported by CNPq (Brazilian Research Council) under grant 401902/81.

† Department of Research and Development, Scientific Computation Laboratory—LCC/CNPq, R. Lauro Müller 455, Rio de Janeiro, RJ, 22290, Brazil and Department of Electrical Engineering, Catholic University—PUC/RJ, R. Marquês de S. Vicente 209, Rio de Janeiro, RJ, 22453, Brazil.

2. Notational and conceptual preliminaries. In this section we pose the notation and some basic concepts which will be used in the sequel. Throughout this paper we assume that U and H are separable nontrivial Hilbert spaces. $\langle\ ;\ \rangle$ and $\|\ \ \|$ will stand for inner product and norm, respectively.

Bounded linear transformations. Let X and Y be Banach spaces. $\mathcal{B}[X, Y]$ will denote the Banach space of all bounded linear transformations of X into Y. For notational simplicity we write $\mathcal{B}[X]$ for $\mathcal{B}[X, X]$. $\mathcal{N}(T)$ and $\mathcal{R}(T)$ will stand for the null space and range space of $T \in \mathcal{B}[X, Y]$, respectively. The spectrum of $T \in \mathcal{B}[X]$ will be denoted by $\sigma(T)$. $P\sigma(T) \subset \sigma(T)$ will denote the point spectrum (i.e. the set of all eigenvalues) of $T \in \mathcal{B}[X]$. $r_\sigma(T) = \sup\{|\lambda|: \lambda \in \sigma(T)\}$ is the spectral radius of $T \in \mathcal{B}[X]$. $T^* \in \mathcal{B}[H, U]$ is the adjoint of $T \in \mathcal{B}[U, H]$. We shall write $T_i \to^w T$, $T_i \to^s T$, or $T_i \to^u T$ if a sequence $\{T_i; i \geqq 0\}$ of operators in $\mathcal{B}[H]$ converges weakly, strongly, or uniformly to $T \in \mathcal{B}[H]$ as $i \to \infty$, respectively.

Positive and nuclear operators. A self-adjoint operator $T = T^* \in \mathcal{B}[H]$ will be called nonnegative ($T \geqq 0$), positive ($T > 0$), or strictly positive ($T > 0$) according to the following standard definitions:

$$T \geqq 0 \quad \Leftrightarrow \quad \langle Tx; x\rangle \geqq 0 \quad \forall x \in H,$$

$$T > 0 \quad \Leftrightarrow \quad \langle Tx; x\rangle > 0 \quad \forall x \neq 0 \in H,$$

$$T > 0 \quad \Leftrightarrow \quad \exists \gamma > 0 \text{ such that } \langle Tx; x\rangle \geqq \gamma\|x\|^2 \quad \forall x \in H.$$

If $T \geqq 0 \in \mathcal{B}[H]$ ($T > 0$, $T > 0$), then there exists a unique $T^{1/2} \geqq 0 \in \mathcal{B}[H]$ ($T^{1/2} > 0$, $T^{1/2} > 0$) such that $(T^{1/2})^2 = T$. For $T \geqq 0 \in \mathcal{B}[H]$ we define the trace of T as usual:

$$\operatorname{tr}(T) \overset{\text{def.}}{=} \sum_k \langle Te_k; e_k\rangle = \sum_k \lambda_k,$$

where $\{e_k; k \geqq 1\}$ is any orthonormal basis for H, and $\{\lambda_k \geqq 0; k \geqq 1\}$ is the set of all $\lambda \in P\sigma(T)$, each of them counted according to its multiplicity. $T \geqq 0 \in \mathcal{B}[H]$ is nuclear (or trace-class) if $\operatorname{tr}(T) < \infty$. Let $\mathcal{B}_1[H]$ denote the class of all nuclear operators on H, and recall that $\mathcal{B}_1[H] \subset \mathcal{B}_\infty[H] \subset \mathcal{B}[H]$, where $\mathcal{B}_\infty[X]$ denotes the class of all compact linear operators on a Banach space X. The following well-known result will be needed in the sequel.

Remark 1. If $T \in \mathcal{B}[H]$ has a bounded inverse (in particular, is strictly positive) and it is compact (in particular, nuclear), then H is necessarily finite-dimensional.

Correlation operators. For arbitrary $x, y \in H$ define the operator $x \circ y \in \mathcal{B}[H]$ as follows [3]:

$$(x \circ y)z = x\langle z; y\rangle,$$

for every $z \in H$. Now let u and v be H-valued second order random variables,[1] and define the following sesquilinear form:

$$E\{\langle (u \circ v)x; y\rangle\} \overset{\text{def.}}{=} E\{\langle x; v\rangle\langle u; y\rangle\}$$

[1] Let $(\Omega, \mathcal{A}, p)$ be a probability space where $\mathcal{A}$ is a σ-algebra of subsets of a nonempty basic set Ω, and p is a probability measure defined on $\mathcal{A}$. An H-valued second order random variable is a p-measurable map $u: \Omega \to H$ such that

$$E\{\|u\|^2\} = \int_\Omega \|u(\omega)\|^2\, dp < \infty$$

(i.e. $u \in L_2(\Omega, p; H)$). Here E denotes the expectation operator. An H-valued second order random sequence $\{u_i; i \geqq 0\}$ is a family of H-valued second order random variables. For an introduction to the theory of H-valued random variables see, for instance, [1].

on $H \times H$, which is bounded. The symbol E on the right hand side denotes expectation in the usual way. Then (cf. [14, p. 120]) there exists a unique operator in $\mathcal{B}[H]$, say $E\{u \circ v\}$, defined by

$$\langle E\{u \circ v\}x; y\rangle = E\{\langle (u \circ v)x; y\rangle\}$$

for every $x, y \in H$. We call $E\{u \circ v\} \in \mathcal{B}[H]$ the correlation of u and v.

Remark 2. The following auxiliary results are readily verified.

$$E\{u \circ v\} = E\{v \circ u\}^*,$$

$$E\{(u+v) \circ (u+v)\} = E\{u \circ u\} + E\{v \circ v\} + E\{u \circ v\} + E\{v \circ u\},$$

$$E\{Au \circ Bv\} = A\, E\{u \circ v\} B^* \quad \forall\, A, B \in \mathcal{B}[H].$$

Moreover, the correlation of u is self-adjoint nonnegative and nuclear; that is

$$0 \leqq E\{u \circ u\} = E\{u \circ u\}^* \in \mathcal{B}_1[H],$$

since

$$\langle E\{u \circ u\}x; x\rangle = E\{|\langle x; u\rangle|^2\} \quad \forall x \in H,$$

$$\operatorname{tr}(E\{u \circ u\}) = E\{\|u\|^2\}.$$

H-valued second order random variables u and v are said to be uncorrelated if $E\{u \circ v\} = E\{u\} \circ E\{v\}$. An H-valued second order random sequence $\{u_i; i \geqq 0\}$ is wide sense stationary if $E\{u_i \circ u_j\}$ depends only on the difference $i-j$ for all $i, j \geqq 0$. It is a white noise if $E\{u_i \circ u_j\} = 0$ for all $i \neq j$.

Approximate controllability. A pair of operators $A \in \mathcal{B}[H]$ and $B \in \mathcal{B}[U, H]$ is approximate controllable [2], briefly (A, B) is A-C (also called weakly reachable [4]), if

$$\bigcap_{j=0}^{\infty} \mathcal{N}(B^* A^{*j}) = \{0\}.$$

We shall be particularly interested in the approximate controllability for the pair $(A, BR^{1/2})$, for some $R = R^* \geqq 0 \in \mathcal{B}[U]$[2]. Notice that

$$(A, BR^{1/2}) \text{ is A-C} \quad \Rightarrow \quad (A, B) \text{ is A-C},$$

since $\mathcal{N}(B^* A^{*j}) \subset \mathcal{N}(R^{1/2} B^* A^{*j})$, and

$$R > 0 \text{ and } (A, B) \text{ is A-C} \quad \Rightarrow \quad (A, BR^{1/2}) \text{ is A-C},$$

since $R > 0 \Rightarrow \mathcal{N}(R^{1/2}) = \{0\} \Rightarrow \mathcal{N}(R^{1/2} B^* A^{*j}) \subset \mathcal{N}(B^* A^{*j})$. Also notice that the reverses of the above statements are not generally true.

3. Deterministic asymptotic stability. Asymptotic stability for discrete deterministic infinite-dimensional linear systems has been investigated by several authors (e.g. see [5], [15], [7], [12]). In this section we present some basic concepts and auxiliary results which will be used in § 4.

DEFINITION 1. Let X be a Banach space, $A \in \mathcal{B}[X]$, and define an X-valued sequence $\{x_i; i \geqq 0\}$ as follows:

$$x_{i+1} = Ax_i, \qquad x_0 \in X. \tag{1}$$

[2] If R is thought of as a correlation operator for an input disturbance sequence, then approximate controllability for the pair $(A, BR^{1/2})$ is sometimes termed stochastic approximate controllability.

The free linear system given in (1) (or equivalently, the operator $A \in \mathcal{B}[X]$) is:

(a) *uniformly asymptotically stable* if $A^i \to^u 0$. That is,

$$\|A^i\| \to 0 \quad \text{as } i \to \infty.$$

(b) *strongly asymptotically stable* if $A^i \to^s 0$. That is,

$$\|A^i x\| \to 0 \quad \text{as } i \to \infty \quad \forall x \in X.$$

Remark 3. By the Banach–Steinhaus theorem [14] it is immediate to verify that

$$\sup_i \|A^i x\| < \infty \quad \forall x \in X \quad \Rightarrow \quad r_\sigma(A) \leqq 1,$$

since $r_\sigma(A)^j = r_\sigma(A^j) \leqq \|A^j\| \leqq \sup_i \|A^i\| < \infty$, $\forall j \geqq 0$ (the reverse is clearly not true, for take any operator $A \in \mathcal{B}[\mathbb{R}^2]$ such that $r_\sigma(A) = 1$ and $\|A^i\| \to \infty$ as $i \to \infty$). Moreover it is also readily verified by contradiction that

$$A^i \overset{s}{\to} 0 \quad \Rightarrow \quad P\sigma(A) \subset \{\lambda \in \mathbb{C}: |\lambda| < 1\}.$$

However even the combined reverse is not true, that is

$$r_\sigma(A) \leqq 1 \text{ and } P\sigma(A) \subset \{\lambda \in \mathbb{C}: |\lambda| < 1\} \quad \not\Rightarrow \quad A^i \overset{s}{\to} 0,$$

since by setting $X = l_2$ and letting $A \in \mathcal{B}[l_2]$ be the right shift operator (i.e. $A(\xi_1, \xi_2, \cdots) = (0, \xi_1, \xi_2, \cdots)$ for all $x = (\xi_1, \xi_2, \cdots) \in l_2$), it follows [11] that $r_\sigma(A) = 1$, $P\sigma(A) = \varnothing$, but $\|A^i x\| = \|x\|$ $\forall i \geqq 0$, for an arbitrary $x \in l_2$.

On the other hand there are several equivalent ways of stating uniform asymptotic stability.

LEMMA 1. *Let X be a complex*[3] *Banach space and $A \in \mathcal{B}[X]$. The following properties are equivalent*:

(a) $\|A^i\| \to 0$ as $i \to \infty$.

(b) $r_\sigma(A) < 1$.

(c) *There exist real constants $\gamma \geqq 1$ and $\rho \in (0, 1)$, such that*

$$\|A^i\| \leqq \gamma \rho^i \quad \forall\, i \geqq 0.$$

(d) $\|A^{i_0}\| < 1$ *for some* $i_0 \geqq 0$.

(e) $\sum_{i=0}^{\infty} \|A^i\|^k < \infty$ *for any* $k > 0$.

(f) $\sum_{i=0}^{\infty} \|A^i\|^{k_0} < \infty$ *for some* $k_0 > 0$.

(g) $\sum_{i=0}^{\infty} \|A^i x\|^k < \infty$, $\forall x \in X$, *for any* $k \geqq 1$.

(h) $\sum_{i=0}^{\infty} \|A^i x\|^{k_0} < \infty$, $\forall \in X$, *for some* $k_0 \geqq 1$.

Proof. It is trivially verified that (c)$\Rightarrow$(e)$\Rightarrow$(f)$\Rightarrow$(a). Since $r_\sigma(A)^i = r_\sigma(A^i) \leqq \|A^i\|$, $\forall i \geqq 0$, one gets (a)$\Rightarrow$(b). By the well-known Gelfand formula, $\|A^i\|^{1/i} \to r_\sigma(A)$ as $i \to \infty$, and by the radical test for infinite series, it follows that (b) $\Rightarrow \|A^i\| < \rho^i$, $\forall i \geqq i_0$,

[3] For a real Banach space X the lemma still holds if $r_\sigma(A)$ is changed to $r_\sigma(A^+)$, where $A^+ \in \mathcal{B}[X^+]$ is defined by $A^+(x + \sqrt{-1}y) = Ax + \sqrt{-1}Ay$, for all $x, y \in X$, with the complex Banach space X^+ denoting the complexification of X (cf. [14]). Notice that $\|A^{+i}\|_{\mathcal{B}[X^+]} = \|A^i\|_{\mathcal{B}[X]}$, $\forall i \geqq 0$.

for some integer $i_0 \geqq 0$ and any $\rho \in (r_\sigma(A), 1)$; which implies (c) with $\gamma = \max\{\|A^j\|: 0 \leqq j \leqq i_0\}\rho^{-i_0} \geqq 1$. Since $\|A^i x\|^k \leqq \|A^i\|^k \|x\|^k$, for all $x \in X$, it is immediate to verify that (e) $\Rightarrow$(g). That (g)$\Rightarrow$(h) is trivial. It has been proved in [15] that (h)$\Rightarrow$(b). Finally it is clear that (c)$\Rightarrow$(d), and (d)$\Rightarrow$(a) since $\|A^{ji_0}\| \leqq \|A^{i_0}\|^j$, $\forall j \geqq 0$. □

Remark 4. Obviously uniform asymptotic stability implies strong asymptotic stability. The fundamental difference between finite- and infinite-dimensional formulations relies upon the reverse of the above statement, which is not generally true for infinite-dimensional spaces. For instance, set $X = l_2$ and let $A \in \mathcal{B}[l_2]$ be the left shift operator (i.e. $A(\xi_1, \xi_2, \cdots) = (\xi_2, \xi_3, \cdots)$ for all $x = (\xi_1, \xi_2, \cdots) \in l_2$). It is easy to show that $\|A^i x\| \to 0$ as $i \to \infty$ for all $x \in l_2$, but $\|A^i\| = 1$ $\forall i \geqq 0$. However, if $A \in \mathcal{B}_\infty[X]$ (in particular, if $\dim(X) < \infty$), then strong and uniform asymptotic stability are equivalent concepts. Indeed, for $A \in \mathcal{B}_\infty[X]$, $\sigma(A) - \{0\} = P\sigma(A) - \{0\}$. Hence, if $A \in \mathcal{B}_\infty[X]$ is strongly asymptotically stable, then the compact set $\sigma(A)$ is contained in the unit open ball, according to Remark 3, and so $r_\sigma(A) < 1$.

4. State correlation evolution and mean square stability. Consider a discrete linear dynamical system evolving in a stochastic environment, and modelled by the following autonomous difference equation.

$$x_{i+1} = Ax_i + Bu_{i+1}, \qquad x_0 = Bu_0, \tag{2}$$

where $A \in \mathcal{B}[H]$ and $B \in \mathcal{B}[U, H]$. Here $\{x_i; i \geqq 0\}$ denotes an H-valued state sequence such that x_0 is an $\mathcal{R}(B) \subset H$-valued second order random variable. The input disturbance sequence $\{u_i; i \geqq 0\}$ is assumed to be an U-valued second order wide sense stationary white noise, with correlation operator

$$R = R^* = E\{u_i \circ u_i\} \geqq 0 \in \mathcal{B}_1[U] \quad \forall i \geqq 0.$$

Now define the following self-adjoint nonnegative operator.

$$Q_i = Q_i^* = \sum_{j=0}^{i} A^j Q_0 A^{*j} \geqq 0 \in \mathcal{B}_1[H], \qquad Q_0 = BRB^*,$$

for every $i \geqq 0$. Notice that Q_i is actually nuclear since R is nuclear, A and B are bounded, and $\mathcal{B}_1[H]$ is a two-sided ideal of $\mathcal{B}[H]$ (cf. [14, p. 173]). On iterating (2) from x_0 onwards, and using Remark 2, it is a simple matter to show that Q_i is the state correlation operator; that is,

$$Q_i = E\{x_i \circ x_i\} \quad \forall i \geqq 0,$$

which has the following further properties.

PROPOSITION 1.

(a) $Q_j = A^{i+1} Q_{j-i-1} A^{*i+1} + Q_i \quad \forall j > i \geqq 0.$

In particular,

$$Q_{i+1} = AQ_iA^* + Q_0 = A^{i+1} Q_0 A^{*i+1} + Q_i \quad \forall i \geqq 0.$$

Therefore, for every $i \geqq 0$,

(b) $Q_i \leqq Q_{i+1},$

thus $\operatorname{tr}(Q_i) \leqq \operatorname{tr}(Q_{i+1})$ *and* $\|Q_i\| \leqq \|Q_{i+1}\|$. *Moreover,*

(c) $Q_i > 0 \quad \Leftrightarrow \quad \bigcap_{j=0}^{i} \mathcal{N}(R^{1/2} B^* A^{*j}) = \{0\}.$

Proof. Let i, j be any integers such that $j > i \geqq 0$. Then

$$Q_j = \sum_{l=0}^{i} A^l Q_0 A^{*l} + \sum_{l=i+1}^{j} A^l Q_0 A^{*l} = Q_i + \sum_{l=0}^{j-i-1} A^{l+i+1} Q_0 A^{*l+i+1},$$

thus following the result in (a). The particular cases are trivially obtained by setting $i = 0$ and $j = i+1$, respectively. The result in (b) is then readily verified since

$$\langle Q_i x; x\rangle = \sum_{j=0}^{i} \|R^{1/2} B^* A^{*j} x\|^2 \quad \forall x \in H.$$

Therefore $\{\mathrm{tr}\,(Q_i)\}$ and $\{\|Q_i\|\}$ are nondecreasing sequences. Since $Q_i \geqq 0$ one gets $Q_i > 0 \Leftrightarrow \{\langle Q_i x; x\rangle = 0 \Rightarrow x = 0\}$. But

$$\langle Q_i x; x\rangle = 0 \quad \Leftrightarrow \quad x \in \bigcap_{j=0}^{i} \mathcal{N}(R^{1/2} B^* A^{*j}),$$

thus following the result in (c). □

We shall be particularly interested in the asymptotic behaviour of the sequence $\{Q_i; i \geqq 0\}$.

LEMMA 2. (a) *If* $Q_i \to^w Q \in \mathcal{B}[H]$, *then* $Q_i \to^s Q$, *and the limit has the following properties*: $0 \leqq Q_i \leqq Q = Q^*$, $\|Q_i\| \nearrow \|Q\|$, *and*

$$Q = A^{i+1} Q A^{*i+1} + Q_i \quad \forall i \geqq 0.$$

Moreover,

$$(A, BR^{1/2}) \text{ is A-C} \quad \Leftrightarrow \quad Q > 0 \quad \Rightarrow \quad P\sigma(A^*) \subset \{\lambda \in \mathbb{C}: |\lambda| < 1\}.$$

(b) *If* $Q_i \to^w Q \in \mathcal{B}_1[H]$, *then* $\mathrm{tr}\,(Q_i) \nearrow \mathrm{tr}\,(Q)$, *and* $Q_i \to^u Q$.

Proof. If $Q_i \to^w Q \in \mathcal{B}[H]$, then by the Banach-Steinhaus theorem $\{Q_i\}$ is uniformly bounded (cf. [14, p. 78]). Therefore since $\{Q_i\}$ is a nondecreasing sequence (according to Proposition 1(b)) of self-adjoint operators, it follows that $Q_i \to^s Q$, and $Q = Q^*$ (cf. [14, p. 79]). Actually $0 \leqq Q_i \leqq Q$ for every $i \geqq 0$, since

$$\langle Q_i x; x\rangle = \sum_{j=0}^{i} \|R^{1/2} B^* A^{*j} x\|^2 \leqq \sum_{j=0}^{\infty} \|R^{1/2} B^* A^{*j} x\|^2 = \langle Qx; x\rangle$$

for all $x \in H$. Thus $\|Q_i\| \leqq \|Q\|$. Hence the nondecreasing sequence $\{\|Q_i\|\}$ converges, and $\|Q\| = \sup_{\|x\|=1} \lim_{i\to\infty} \langle Q_i x; x\rangle \leqq \lim_{i\to\infty} \|Q_i\|$. Then $\|Q_i\| \nearrow \|Q\|$. By Proposition 1(a) it follows that

$$Q_j - (A^{i+1} Q A^{*i+1} + Q_i) = A^{i+1}(Q_{j-i-1} - Q) A^{*i+1}$$

for every $j > i \geqq 0$. Therefore, since $Q_j \to^s Q$,

$$\|[Q_j - (A^{i+1} Q A^{*i+1} + Q_i)]x\| \leqq \|A^{i+1}\| \, \|(Q_{j-i-1} - Q) A^{*i+1} x\| \to 0$$

as $j \to \infty$, for all $x \in H$ and every $i \geqq 0$. Then, by uniqueness of the strong limit, $Q = A^{i+1} Q A^{*i+1} + Q_i$, $\forall i \geqq 0$. Moreover,

$$\langle Qx; x\rangle = 0 \quad \Leftrightarrow \quad x \in \bigcap_{j=0}^{\infty} \mathcal{N}(R^{1/2} B^* A^{*j}).$$

So, recalling that $Q \geqq 0$, one has

$$Q > 0 \quad \Leftrightarrow \quad \{\langle Qx; x\rangle = 0 \Rightarrow x = 0\} \quad \Leftrightarrow \bigcap_{j=0}^{\infty} \mathcal{N}(R^{1/2} B^* A^{*j}) = \{0\}.$$

Finally take any $\lambda \in P\sigma(A^*)$ (if $P\sigma(A^*) = \varnothing$ the result is trivial), and let $x \neq 0$ be an eigenvalue associated to λ. Then

$$\langle (Q - Q_{i-1})x; x\rangle = \langle A^i Q A^{*i} x; x\rangle = |\lambda|^{2i}\langle Qx; x\rangle \quad \forall i \geqq 1.$$

Hence $|\lambda| < 1$ whenever $Q_i \to^w Q > 0$, which completes the proof of part (a). Now assume that $Q \in \mathcal{B}_1[H]$. Then $\operatorname{tr}(Q_i) \leqq \operatorname{tr}(Q)$, since $Q_i \leqq Q$, and the nondecreasing sequence $\{\operatorname{tr}(Q_i)\}$ converges. Thus, for any orthonormal basis $\{e_k\}$ and for every $n \geqq 1$,

$$\operatorname{tr}(Q) = \lim_{i\to\infty} \sum_{k=1}^{n} \langle Q_i e_k; e_k\rangle + \sum_{k=n+1}^{\infty} \langle Q e_k; e_k\rangle$$

$$\leqq \lim_{i\to\infty} \operatorname{tr}(Q_i) + \sum_{k=n+1}^{\infty} \langle Q e_k; e_k\rangle \searrow \lim_{i\to\infty} \operatorname{tr}(Q_i) \quad \text{as } n \to \infty.$$

Then $\operatorname{tr}(Q_i) \nearrow \operatorname{tr}(Q)$. Therefore

$$\|Q - Q_i\| \leqq \operatorname{tr}(Q - Q_i) = \operatorname{tr}(Q) - \operatorname{tr}(Q_i) \to 0 \quad \text{as } i \to \infty. \qquad \square$$

Remark 5. Concerning the final statement of Lemma 2(a) it is worth mentioning that positivity (which is sufficient) is not necessary, but nonnegativity is not sufficient (i.e. $Q \geqq 0 \not\Rightarrow P\sigma(A^*) \subset \{\lambda \in \mathbb{C}: |\lambda| < 1\} \not\Rightarrow Q > 0$). It is also easy to show that $\mathcal{N}(Q) \subset \mathcal{N}(Q_{i+1}) \subset \mathcal{N}(Q_i)$, $\forall i \geqq 0$.

We shall say that the linear system in (2) is mean square stable if the state correlation sequence $\{Q_i; i \geqq 0\}$ converges to a correlation operator Q (i.e. $E\{x_i \circ x_i\} \to E\{x \circ x\}$ as $i \to \infty$ for some second order H-valued random variable x), such that the Lyapunov equation $Q = AQA^* + Q_0$ in Lemma 2(a) has a solution $Q \geqq 0 \in \mathcal{B}_1[H]$. However, by Lemma 2(b), the above convergence has to be uniform. So we define as follows.

DEFINITION 2. The linear system in (2) is *mean square stable* if

$$Q_i \overset{u}{\to} Q \in \mathcal{B}_1[H].$$

We now investigate the connection between mean square and uniform asymptotic stability concepts.

THEOREM 1.

a) $Q_i \overset{u}{\to} Q > 0 \in \mathcal{B}[H] \Rightarrow r_\sigma(A) < 1.$

b) $Q_i \overset{s}{\to} Q > 0 \in \mathcal{B}[H] \Rightarrow r_\sigma(A) \leqq 1.$

Proof. Since $Q > 0$, $\exists Q^{-1} \in \mathcal{B}[H]$. By Lemma 2(a) $Q - Q_i = A^{i+1} Q A^{*i+1}$, $\forall i \geqq 0$. Hence

(a) $$\|A^{i+1}\|^2 = \|A^{i+1}(Q^{1/2})(Q^{1/2})^{-1}\|^2 \leqq \|A^{i+1}(Q^{1/2})\|^2 \|(Q^{1/2})^{-1}\|^2$$
$$= \|A^{i+1} Q A^{*i+1}\| \, \|Q^{-1}\| = \|Q - Q_i\| \, \|Q^{-1}\| \to 0 \quad \text{as } i \to \infty,$$

thus following the desired result by Lemma 1.

(b) $$\|A^{*i+1} x\| = \|(Q^{1/2})^{-1}(Q^{1/2}) A^{*i+1} x\|^2 \leqq \|(Q^{1/2})^{-1}\|^2 \|(Q^{1/2}) A^{*i+1} x\|^2$$
$$= \|Q^{-1}\| \, \langle A^{i+1} Q A^{*i+1} x; x\rangle = \|Q^{-1}\| \, \langle (Q - Q_i)x; x\rangle \to 0 \quad \text{as } i \to \infty,$$

for all $x \in H$, thus following part (b) by Remark 3, since $r_\sigma(A^*) = r_\sigma(A)$. $\square$

THEOREM 2.

$$r_\sigma(A) < 1 \Rightarrow Q_i \overset{u}{\to} Q \in \mathcal{B}_1[H].$$

Proof. Suppose $r_\sigma(A) < 1$. Since

$$\|Q_i\| \leqq \sum_{j=0}^{i} \|A^j Q_0 A^{*j}\| \leqq \|Q_0\| \sum_{j=0}^{i} \|A^j\|^2 \quad \forall i \geqq 0,$$

it follows by Lemma 1 that $\{Q_i\,;\, i \geqq 0\}$ is uniformly bounded. Therefore, using Proposition 1(a), we get for every $j > i \geqq 0$,

$$\|Q_j - Q_i\| \leqq \|Q_{j-i-1}\| \, \|A^{i+1}\|^2 \leqq \sup_k \|Q_k\| \, \|A^{i+1}\|^2 \to 0 \quad \text{as } i \to \infty,$$

by Lemma 1. Then $Q_i \to^u Q \in \mathcal{B}[H]$, since $\{Q_i\}$ is a Cauchy sequence and $\mathcal{B}[H]$ is a Banach space. Finally, since A^j and B are bounded and R is nuclear, it can be shown [14, p. 173] that $\operatorname{tr}(A^j B R B^* A^{*j}) \leqq \|A^j\|^2 \|B\|^2 \operatorname{tr}(R)$. Therefore, using Lemma 1 again,

$$\operatorname{tr}(Q_i) = \sum_{j=0}^{i} \operatorname{tr}(A^j B R B^* A^{*j}) \leqq \operatorname{tr}(R) \, \|B\|^2 \sum_{j=0}^{\infty} \|A^j\|^2.$$

Hence $\{\operatorname{tr}(Q_i)\}$ is a bounded sequence, and so (cf. [14, p. 179]) the uniform limit Q of the nuclear sequence $\{Q_i\}$ must be nuclear. □

Remark 6. We notice from Lemma 2 that $Q_i \to^w Q \Leftrightarrow Q_i \to^s Q$, and $Q_i \to^w Q \in \mathcal{B}_1[H] \Rightarrow Q_i \to^u Q$. However it can be shown that

(a) $Q_i \xrightarrow{s} Q \in \mathcal{B}[H] \quad \not\Rightarrow \quad Q_i \xrightarrow{u} Q \in \mathcal{B}[H]$,

(b) $Q_i \xrightarrow{u} Q \in \mathcal{B}[H] \quad \not\Rightarrow \quad Q \in \mathcal{B}_1[H]$.

Moreover, it can also be verified that both strong convergence and positivity are not sufficient in Theorem 1(a). That is,

(c) $Q_i \xrightarrow{s} Q > 0 \in \mathcal{B}[H] \quad \not\Rightarrow \quad r_\sigma(A) < 1$,

(d) $Q_i \xrightarrow{u} Q > 0 \in \mathcal{B}_1[H] \quad \not\Rightarrow \quad r_\sigma(A) \leqq 1$.

To illustrate the above statements we consider the following examples.

Example 1. First we show that the statements (a) and (c) in Remark 6 hold true. Set $H = l_2$ and $U = \mathbb{R}^1$. Let $A \in \mathcal{B}[l_2]$ be the right shift operator, $A(\xi_1, \xi_2, \cdots) = (0, \xi_1, \xi_2, \cdots)$ for all $x = (\xi_1, \xi_2, \cdots) \in l_2$. Let $B \in \mathcal{B}[\mathbb{R}^1, l_2]$ be given by $Bu = (u, 0, \cdots)$ for all $u \in \mathbb{R}^1$, and set $R = 1$, the identity operator in $\mathbb{R}^1$. It is a simple matter to verify that

$$Q_i = \sum_{j=0}^{i} A^j B B^* A^{*j} = \operatorname{diag}(1, \cdots, 1, 0, \cdots) \geqq 0 \in \mathcal{B}_1[l_2] \quad \forall i \geqq 0,$$

with the nonzero entries at the first $i+1$ positions, such that $\operatorname{tr}(Q_i) = i+1$. Hence

$$Q_i \xrightarrow{s} Q = I > 0 \in \mathcal{B}[l_2],$$

since $\|(I - Q_i)x\|^2 = \sum_{k=i+2}^{\infty} |\xi_k|^2 \to 0$ as $i \to \infty$ for all $x = (\xi_1, \xi_2, \cdots) \in l_2$, although $\{Q_i\}$ does not converge uniformly since $\|I - Q_i\| = 1$, $\forall i \geqq 0$. This supports the statement (a) in Remark 6. However, as it is well known [11], $r_\sigma(A) = 1$, thus confirming the statement (c) in Remark 6.

Example 2. Now we illustrate the statements (b) and (d) in Remark 6. Let $\{\varepsilon_k\,;\, k \geqq 1\}$ be a real positive sequence in l_1, and define a real positive strictly decreasing null sequence $\{\lambda_k\,;\, k \geqq 1\}$ as follows.

$$\lambda_{k+1} = \lambda_k - \varepsilon_k, \qquad \lambda_1 = \sum_{k=1}^{\infty} \varepsilon_k.$$

Set $H = U = l_2$. Let $A \in \mathcal{B}[l_2]$ be a constantly weighted left shift operator, $A(\xi_1, \xi_2, \cdots) = \rho^{1/2}(\xi_2, \xi_3, \cdots)$ for all $x = (\xi_1, \xi_2, \cdots) \in l_2$, such that [11] $r_\sigma(A) = \rho^{1/2} > 0$. Let $B = I \in \mathcal{B}[l_2]$, the identity operator, and $R = \operatorname{diag}(\varepsilon_1, \varepsilon_2, \cdots) > 0 \in \mathcal{B}_1[l_2]$, with $\operatorname{tr}(R) = \lambda_1$. It is readily verified that

$$Q_i = \sum_{j=0}^{i} A^j R A^{*j} = \operatorname{diag}\left(\sum_{j=0}^{i} \rho^j \varepsilon_{j+1}, \sum_{j=0}^{i} \rho^j \varepsilon_{j+2}, \cdots \right) > 0 \in \mathcal{B}_1[l_2],$$

with $\operatorname{tr}(Q_i) = \sum_{j=0}^{i} \rho^j \lambda_{j+1}$, $\forall i \geqq 0$. In particular, with $r_\sigma(A) = \rho = 1$ it follows that

$$Q_i = \operatorname{diag}(\lambda_1 - \lambda_{i+2}, \lambda_2 - \lambda_{i+3}, \cdots) > 0 \in \mathcal{B}_1[l_2],$$

with $\operatorname{tr}(Q_i) = \sum_{k=1}^{i+1} \lambda_k$. Thus

$$Q_i \xrightarrow{u} Q = \operatorname{diag}(\lambda_1, \lambda_2, \cdots) > 0 \in \mathcal{B}[l_2],$$

since $\|Q - Q_i\| = \lambda_{i+2} \to 0$ as $i \to \infty$. However

$$Q \in \mathcal{B}_1[l_2] \quad \Leftrightarrow \quad \operatorname{tr}(Q) = \sum_{k=1}^{\infty} \lambda_k < \infty,$$

which does not necessarily happen. For instance,

$$\varepsilon_k = \frac{1}{k(k+1)} \quad \forall k \geqq 1 \quad \Rightarrow \quad \lambda_k = \frac{1}{k} \quad \forall k \geqq 1 \quad \Rightarrow \quad Q \notin \mathcal{B}_1[l_2],$$

$$\varepsilon_k = \frac{2k+1}{k^2(k+1)^2} \quad \forall k \geqq 1 \quad \Rightarrow \quad \lambda_k = \frac{1}{k^2} \quad \forall k \geqq 1 \quad \Rightarrow \quad Q \in \mathcal{B}_1[l_2].$$

This confirms the statement (b) in Remark 6. Now let $r_\sigma(A)^2 = \rho > 1$ and set $\varepsilon_k = \alpha^{k-1}$, $\forall k \geqq 1$, with $0 < \alpha < \rho^{-1} < 1$. Then $R = \operatorname{diag}(1, \alpha, \alpha^2, \cdots) > 0 \in \mathcal{B}_1[l_2]$, with $\operatorname{tr}(R) = (1-\alpha)^{-1}$, and

$$Q_i = \frac{1 - (\alpha\rho)^{i+1}}{1 - \alpha\rho} R > 0 \in \mathcal{B}_1[l_2] \quad \forall i \geqq 0,$$

with $\operatorname{tr}(Q_i) = [1 - (\alpha\rho)^{i+1}][(1-\alpha\rho)(1-\alpha)]^{-1}$. Thus

$$Q_i \xrightarrow{u} Q = (1 - \alpha\rho)^{-1} R > 0 \in \mathcal{B}_1[l_2],$$

with $\operatorname{tr}(Q) = [(1-\alpha\rho)(1-\alpha)]^{-1}$, since $\|Q - Q_i\| = (1-\alpha\rho)^{-1}(\alpha\rho)^{i+1} \to 0$ as $i \to \infty$. However $r_\sigma(A) > 1$, thus supporting the statement (d) in Remark 6.

Remark 7. By Theorem 1(a), Theorem 2, and Remark 1 one has

$$Q_i \xrightarrow{u} Q > 0 \in \mathcal{B}[H] \quad \Rightarrow \quad \dim(H) < \infty,$$

although (cf. Example 1)

$$Q_i \xrightarrow{s} Q > 0 \in \mathcal{B}[H] \quad \not\Rightarrow \quad \dim(H) < \infty.$$

If $\dim(H) < \infty$, then $\mathcal{B}_1[H] = \mathcal{B}[H]$, $P\sigma(A) = \sigma(A)$, strict positivity is equivalent to positivity, and uniform convergence is equivalent to strong convergence. Therefore, in such a case, it follows by Theorem 1 and Theorem 2 that

$$Q_i \to Q > 0 \in \mathcal{B}[H] \quad \Leftrightarrow \quad r_\sigma(A) < 1 \text{ and } Q > 0.$$

However the assumption $Q > 0$, which appears in both sides of the above statement, may not be dismissed. That is, even for finite-dimensional spaces,

$$Q_i \to Q \in \mathcal{B}[H] \quad \not\Rightarrow \quad r_\sigma(A) \leqq 1, \qquad r_\sigma(A) < 1 \quad \not\Rightarrow \quad Q > 0,$$

as it is readily verified. These finite-dimensional results can be extended to infinite-dimensional spaces, whenever A is compact, as follows.

COROLLARY 1. *If $A \in \mathcal{B}_\infty[H]$, then the following properties are equivalent*:

(a) $r_\sigma(A) < 1$ *and* $(A, BR^{1/2})$ *is* A-C.

(b) $Q_i \xrightarrow{u} Q > 0 \in \mathcal{B}_1[H]$.

(c) $Q_i \xrightarrow{s} Q > 0 \in \mathcal{B}[H]$.

Proof. (a)$\Rightarrow$(b) by Lemma 2(a) and Theorem 2, and (b)$\Rightarrow$(c) trivially, for any $A \in \mathcal{B}[H]$. Now assume that $A \in \mathcal{B}_\infty[H]$. Then (c)$\Rightarrow$(a), since $r_\sigma(A) = r_\sigma(A^*) = \max\{|\lambda|: \lambda \in P\sigma(A^*) \cup \{0\}\} < 1$, by Lemma 2(a). □

5. Concluding remarks. In this paper we have considered mean square stability for discrete bounded linear systems in Hilbert space driven by white noise. The evolution and convergence of the state correlation operators sequence were investigated in Proposition 1 and Lemma 2. It has been shown in Theorem 2 that uniform asymptotic stability is a sufficient condition for mean square stability, although the reverse is not necessarily true (cf. Remark 6), as it occurs in a finite-dimensional setting whenever $Q > 0$ (cf. Theorem 1 and Remark 7).

For compact operators the discrete-time stability problem is quite clear, being a straightforward generalization of the finite-dimensional case. Indeed, as recalled in Remark 4, for deterministic systems strong and uniform asymptotic stability are equivalent concepts whenever A is compact. Comparing Remark 7 with Corollary 1 it is readily verified that a similar situation actually happens for stochastic systems with a compact operator A.

Acknowledgments. The author gratefully thanks Dr. A. C. Gadelha Vieira (LCC/CNPq) for his helpful comments and stimulating discussions on the subject of this paper. Thanks are also due to Prof. Dr. Ruth F. Curtain (University of Groningen) for her suggestions, and to the anonymous referee for bringing the author's attention to part (b) of Lemma 2.

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, 2nd ed., Springer-Verlag, Berlin, 1980.

[2] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences, 8, Springer-Verlag, Berlin, 1978.

[3] P. L. FALB, *Infinite-dimensional filtering: the Kalman–Bucy filter in Hilbert space*, Inform. Control, 11 (1967), pp. 102–137.

[4] P. A. FUHRMANN, *On weak and strong reachability and controllability of infinite-dimensional linear systems*, J. Optim. Theory Appl., 9 (1972), pp. 77–89.

[5] ———, *On observability and stability in infinite-dimensional linear systems*, J. Optim. Theory Appl., 12 (1973), pp. 173–181.

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

[7] E. W. KAMEN AND W. L. GREEN, *Asymptotic stability of linear difference equations defined over a commutative Banach algebra*, J. Math. Anal. Appl., 75 (1980), pp. 584–601; *Addendum*, 84 (1981), pp. 295–298.

[8] C. S. KUBRUSLY, *Applied stochastic approximation algorithms in Hilbert space*, Int. J. Control, 28 (1978), pp. 23–31.

[9] H. J. KUSHNER, *Introduction to Stochastic Control Theory*, Holt, Rinehart & Winston, New York, 1971.

[10] J. P. LA SALLE, *The Stability of Dynamical Systems*, CBMS Regional Conference Series in Applied Mathematics, 25, Society for Industrial and Applied Mathematics, Philadelphia, 1976.

[11] A. W. NAYLOR AND G. R. SELL, *Linear Operator Theory in Engineering and Science*, 2nd ed., Springer-Verlag, Berlin, 1982.

[12] K. M. PRZYLUSKI, *The Lyapunov equation and the problem of stability for linear bounded discrete-time systems in Hilbert space*, Appl. Math. Optim., 6 (1980), pp. 97–112.

[13] J. H. VENTER, *On Dvoretzky stochastic approximation theorems*, Ann. Math. Statist. 37 (1966), pp. 1534–1544.

[14] J. WEIDMANN, *Linear Operators in Hilbert Spaces*, Springer-Verlag, Berlin, 1980.

[15] J. ZABCZYK, *Remarks on the control of discrete-time distributed parameter systems*, this Journal, 12 (1974), pp. 721–735.

[16] ———, *On optimal stochastic control systems in Hilbert space*, this Journal, 13 (1975), pp. 1217–1234.

[17] ———, *Stability properties of the discrete Riccati operator equation*, Kybernetica, 13 (1977), pp. 1–10.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

SUFFICIENCY OF EXACT PENALTY MINIMIZATION*

O. L. MANGASARIAN†

Abstract. By employing a recently obtained error bound for differentiable convex inequalities, it is shown that, under appropriate constraint qualifications, a minimum solution of an exact penalty function for a *single* value of the penalty parameter which exceeds a certain threshold, is also a solution of the convex program associated with the penalty function. No a priori assumption is made regarding the solvability of the convex program. If such a solvability assumption is made, then we show that a threshold value of the penalty parameter can be used which is smaller than both the above-mentioned value and that of Zangwill. These various threshold values of the penalty parameter also apply to the well-known big-M method of linear programming.

AMS (MOS) subject classifications. 90C30, 90C25

Key words. nonlinear programming, penalty functions, optimization, convex programming

1. Introduction.

Consider the convex program

$$\text{minimize } f(x) \quad \text{subject to } g(x) \leqq 0 \tag{1.1}$$

where $f: R^n \to R$, $g: R^n \to R^m$ are convex functions on the n-dimensional real Euclidean space R^n. It is well known [11], [6] that if (1.1) has a solution $\bar{x}$ and if the constraints of (1.1) satisfy a constraint qualification, then the exact penalty function

$$P(x, \alpha) := f(x) + \alpha e g(x)_+ = f(x) + \alpha \sum_{i=1}^{m} \max\{0, g_i(x)\} \tag{1.2}$$

where e is a vector of ones in R^m, has a global minimum at $\bar{x}$ for each value of $\alpha \geqq \bar{\alpha}$ for some threshold value $\bar{\alpha}$. In [11, p. 356] [2, Thm. 40] it was shown that

$$\bar{\alpha} = \bar{\alpha}_1 := \frac{f(x^1) - f(\bar{x}) + 1}{\min_{1 \leqq i \leqq m} -g_i(x^1)} \tag{1.3}$$

where x^1 is any point satisfying the Slater constraint qualification

$$g(x^1) < 0. \tag{1.4}$$

In [6, Thm. 4.9] it was shown that

$$\bar{\alpha} = \bar{\alpha}_2 := \|\bar{u}\|_\infty = \max_{1 \leqq i \leqq m} \bar{u}_i \tag{1.5}$$

where $\bar{u}$ is an optimal Lagrange multiplier for (1.1) provided that (1.4) holds. A minor modification of the proof of [6, Thm. 4.9] which invokes [10, Thm. 28.2] instead of [7, Thm. 5.4.8] extends (1.5) to the case where a relaxed Slater constraint qualification holds, that is

$$g_{I_1}(x^2) < 0, \qquad g_{I_2}(x^2) \leqq 0 \quad \text{for some } x^2 \tag{1.6}$$

where g_{I_1} is nonlinear and g_{I_2} is linear and $I_1 \cup I_2 = \{1, \cdots, m\}$. In contrast Zangwill's threshold (1.3) does not hold under the relaxed Slater constaint qualification (1.6) but must be replaced by a different value given by (2.2) below.

* Received by the editors December 6, 1983, and in revised form March 6, 1984. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041. This material is based on work sponsored by the National Science Foundation under grant MCS-8200632.

† Computer Sciences Department, University of Wisconsin, Madison, Wisconsin 53706.

What is not well known and constitutes a principal concern of this work are converses to the results stated above. In [11, p. 356], [2, Thm. 40] Zangwill shows that if we assume a priori that the minimization problem (1.1) has a solution, the Slater constraint qualification (1.4) is satisfied and $\bar{x}$ minimizes the exact penalty function (1.2) for some $\alpha \geqq \bar{\alpha}_1$, then $\bar{x}$ solves the minimization problem (1.1). Note the a priori assumptions that (1.1) is solvable and that it satisfies the Slater constraint qualification. By contrast in [6, Thm. 4.1] without any a priori assumptions regarding the solvability of the minimization problem (1.1) or the satisfaction of a constraint qualification it was shown that if (1.1) is feasible, that is $g(x) \leqq 0$ for some x, and if $\bar{x}$ minimizes $P(x, \alpha)$ for *all* values of $\alpha \geqq \bar{\alpha}$ for some $\bar{\alpha}$, then $\bar{x}$ solves the minimization problem (1.1). Note the distinction between these two sufficient conditions for $\bar{x}$ to solve the minimization problem (1.1). In Zangwill's result there are a priori assumptions that (1.1) is solvable and that its constraints satisfy the Slater constraint qualification, while the penalty function $P(x, \alpha)$ need be minimized for a single value of $\alpha \geqq \bar{\alpha}_1$. In [6, Thm. 4.1] no a priori assumption regarding the existence of a solution to (1.1) is made; however, feasibility of (1.1) is assumed and $\bar{x}$ must be a solution to $\min_{x \in R^n} P(x, \alpha)$ for *all* $\alpha \geqq \bar{\alpha}$ for some $\bar{\alpha}$, in order for $\bar{x}$ to be a solution to (1.1).

A primary purpose of this work is to combine the good features of these two results, namely the minimization of the penalty function for a *single* value of the penalty parameter and without an a priori assumption that the minimization problem has a solution. This is done in Theorem 3.1 where it is established that if for a single value of the penalty parameter $\alpha \geqq \bar{\alpha}_3$ for a well-defined $\bar{\alpha}_3$, $\bar{x}$ minimizes the exact penalty function $P(x, \alpha)$ over R^n, then $\bar{x}$ is also a global solution of the minimization problem (1.1). Although no a priori assumption regarding the solvability of (1.1) is made in Theorem 3.1, both the relaxed Slater constraint qualification (1.6) and a mild asymptotic constraint qualification (3.2) are needed in order to invoke the recent [8, Thm. 2.1] absolute error bound for convex differentiable inequalities which plays a key role in the derivation of Theorem 3.1. Another result of this work is a two-way improvement of Zangwill's sufficiency result in Theorem 2.1, where the threshold value of $\bar{\alpha}$ is decreased from $\bar{\alpha}_1$ of (1.3) to $\bar{\alpha}_2$ of (1.5) and the Slater constraint qualification (1.4) is replaced by the relaxed constraint qualification (1.6). We also give in Corollary 2.3 a finite counterpart of the threshold value $\bar{\alpha}_1$ of (1.3) when the Slater constraint qualification (1.4) is replaced by the relaxed qualification (1.6) which renders $\bar{\alpha}_1$ infinite. Table 1 below gives a general outline of the relations between the various sufficiency results derived here and elsewhere for exact penalty functions and indicates the key assumptions needed for the different results to hold.

TABLE 1

An outline of the key assumptions needed in the various sufficiency theorems establishing that each minimizer of an exact penalty function (1.2) *solves the minimization problem* (1.1).

Penalty function (1.2) minimized for:	A priori solvability of min. prob. (1.1): Assumed	A priori solvability of min. prob. (1.1): Not assumed	Constraint qualification:
All $\alpha \geqq \bar{\alpha}$		Han–Mangasarian [6, Thm. 4.1]	Not assumed
A single $\alpha \geqq \bar{\alpha}$	Zangwill [11, p. 356] Theorem 2.1	Theorem 3.1	Assumed

In § 3 of the paper we show that the big-M method of linear programming [1], [9] is in fact equivalent to an exact penalty problem and hence the threshold values of the penalty parameter developed in this work apply to it as well as to a big-M formulation for convex programs. Such threshold values do not seem to have been given for the big-M method for linear programs.

We briefly describe now our notation. For a vector x in the n-dimensional real Euclidean space R^n, x_+ will denote the vector in R^n with components $(x_+)_i = \max\{x_i, 0\}$, $i = 1, \cdots, n$. For a vector norm $\|x\|$ on R^n, $\|x\|'$ will denote the dual norm on R^n, that is $\|x\|' = \max_{\|y\|=1} xy$, where xy denotes the scalar product $\sum_{i=1}^n x_i y_i$. The Cauchy–Schwarz inequality $|xy| \leqq \|x\| \cdot \|y\|'$ for x and y in R^n follows immediately from this definition of the dual norm. For $1 \leqq p, q \leqq \infty$ and $(1/p)+(1/q) = 1$, the p-norm $\|x\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ and the q-norm are dual norms in R^n. For an $m \times n$ matrix A, A_i denotes the ith row, while $\|A\|_p$ denotes the matrix norm subordinate to the vector norm $\|\cdot\|_p$, that is $\|A\|_p = \max_{\|x\|_p=1} \|Ax\|_p$. The consistency condition $\|Ax\|_p \leqq \|A\|_p \|x\|_p$ follows immediately from this definition of a matrix norm. We shall also use $\|\cdot\|$ to denote an arbitrary vector norm and its subordinate matrix norm. A vector of ones in any real Euclidean space will be denoted by e. For a differentiable function $g: R^n \to R^m$, $\nabla g(x)$ will denote the $m \times n$ Jacobian matrix evaluated at the point x in R^n. For a subset $I \subset \{1, \cdots, m\}$, $g_I(x)$ or $g_{i \in I}(x)$ will denote those components $g_i(x)$ such that $i \in I$. Similarly $\nabla g_I(x)$ will denote the rows $(\nabla g(x))_i$ of $\nabla g(x)$ such that $i \in I$. The set of vectors in R^n with nonnegative components will be denoted by R^n_+.

2. Exact penalty characterization assuming solvability of the minimization problem. In this section we completely characterize solutions of the minimization problem (1.1) in terms of minimizers of the exact penalty function (1.2) for a single value of the penalty parameter exceeding the threshold $\bar{\alpha}_2$. This is done under the assumptions that the minimization problem is solvable and that it satisfies the relaxed Slater constraint qualification (1.6). The necessity part of the following result Theorem 2.1 is an improvement over both [6, Thm. 4.9] and Zangwill's theorem [11, p. 356] both of which require the Slater constraint qualification (1.4) instead of the relaxed qualification (1.6) needed here. This is a simple but important difference because it allows us to handle linearly constrained problems with no constraint qualification, and because Zangwill's threshold value $\bar{\alpha}_1$ becomes infinite under the relaxed constraint qualification (1.6). The new sufficiency part of Theorem 2.1 again improves over Zangwill's sufficiency result by using the relaxed Slater constraint qualification (1.6) instead of the Slater constraint qualification (1.4), and the smaller threshold value $\bar{\alpha}_2$ instead of $\bar{\alpha}_1$. It is interesting to note that the sufficiency part of Theorem 2.1 for the threshold value $\bar{\alpha}_2$ does not appear to have been given before even under the Slater constraint qualification. Now we state our result.

THEOREM 2.1 (Exact penalty characterization of solvable convex programs). *Let $f: R^n \to R$ and $g: R^n \to R^m$ be convex functions on R^n. Let either $(\bar{x}, \bar{u}) \in R^n \times R^m_+$ be a Karush–Kuhn–Tucker saddlepoint of the minimization problem* (1.1), *or let the relaxed Slater constraint qualification* (1.6) *hold and $\bar{x}$ be a solution of* (1.1). *A necessary (sufficient) condition for $\tilde{x} \in R^n$ to solve the minimization problem* (1.1) *is that $\tilde{x}$ minimizes $P(x, \alpha)$ over x in R^n for each (some) $\alpha \geqq \|\hat{u}\|_\infty (\alpha > \|\hat{u}\|_\infty)$ where $\hat{u} \in R^m_+$ is any (some) dual optimal multiplier for* (1.1).

Proof. Necessity. By assumption or by [10, Thm. 28.2] there exists a $\bar{u} \in R^m_+$ such that $(\bar{x}, \bar{u})$ is a Karush–Kuhn–Tucker saddlepoint of (1.1). For any other dual optimal multiplier $\hat{u}$, $(\bar{x}, \hat{u})$ is also a Karush–Kuhn–Tucker saddlepoint of (1.1) [4, p. 5]. Hence

for $x \in R^n$ and $\alpha \geqq \|\hat{u}\|_\infty$

$$P(\bar{x}, \alpha) = f(\bar{x}) = f(\bar{x}) + \hat{u}g(\bar{x}) \leqq f(x) + \hat{u}g(x)$$

$$\leqq f(x) + \hat{u}g(x)_+ \leqq f(x) + \|\hat{u}\|_\infty \|g(x)_+\|_1 \leqq P(x, \alpha).$$

Sufficiency. Let $\hat{u} \in R^m_+$ be some dual optimal multiplier for (1.1). Since $(\bar{x}, \bar{u})$ is a Karush–Kuhn–Tucker saddlepoint for (1.1) it follows by the necessity part of this theorem that for $\beta := \|\hat{u}\|_\infty$

$$P(\bar{x}, \beta) = \min_{x \in R^n} P(x, \beta).$$

Let $\tilde{x}$ be a solution of $\min_{x \in R^n} P(x, \alpha)$ for some $\alpha > \|\hat{u}\|_\infty = \beta$. Hence

$$f(\bar{x}) + \alpha e g(\bar{x})_+ \geqq f(\tilde{x}) + \alpha e g(\tilde{x})_+$$

and

$$f(\tilde{x}) + \beta e g(\tilde{x})_+ \geqq f(\bar{x}) + \beta e g(\bar{x})_+.$$

Addition of the last two inequalities gives upon noting that $g(\bar{x})_+ = 0$

$$(\alpha - \beta) e g(\tilde{x})_+ \leqq 0.$$

Since $\alpha > \beta$, this implies that $g(\tilde{x}) \leqq 0$ and hence $\tilde{x}$ is feasible for (1.1). For any other feasible point x

$$f(x) = P(x, \alpha) \geqq P(\tilde{x}, \alpha) = f(\tilde{x}). \qquad \square$$

The following corollary shows that under the Slater constraint qualification the threshold value $\bar{\alpha}_2 := \|\hat{u}\|_\infty$ of Theorem 2.1 is smaller than that of Zangwill's $\bar{\alpha}_1$ as defined in (1.3).

COROLLARY 2.2. *Let $f: R^n \to R$ and $g: R^n \to R^m$ be convex functions on R^n, let x^1 be any point in R^n satisfying the Slater constraint qualification $g(x^1) < 0$, and let $\bar{x}$ be a solution of the minimization problem* (1.1). *Then for any dual optimal multiplier $\hat{u} \in R^m_+$ for* (1.1)

$$\|\hat{u}\|_\infty \leqq \|\hat{u}\|_1 \leqq \frac{f(x^1) - f(\bar{x})}{\min_{1 \leqq i \leqq m} -g_i(x^1)} < \frac{f(x^1) - f(\bar{x}) + 1}{\min_{1 \leqq i \leqq m} -g_i(x^1)} =: \bar{\alpha}_1. \tag{2.1}$$

Proof. Since $\bar{x}$ is a solution of (1.1) and the Slater constraint qualification is satisfied, it follows that $\bar{x}$ and some $\bar{u} \in R^m_+$ constitute a Karush–Kuhn–Tucker saddlepoint for (1.1) and by [4, p. 5] so does $(\bar{x}, \hat{u})$. Consequently

$$f(\bar{x}) = f(\bar{x}) + \hat{u}g(\bar{x}) \leqq f(x^1) + \hat{u}g(x^1) \leqq f(x^1) - \|\hat{u}\|_1 \min_{1 \leqq i \leqq m} -g_i(x^1),$$

from which (2.1) follows. $\square$

We establish now another upper bound for the threshold value $\bar{\alpha}_2 := \|\hat{u}\|_\infty$ of Theorem 2.1 under the relaxed Slater constraint qualification (1.6).

COROLLARY 2.3. *Let $f: R^n \to R$ and $g: R^n \to R^m$ be differentiable convex functions on R^n, let x^2 be any point in R^n satisfying the relaxed Slater constraint qualification* (1.6), *and let $\bar{x}$ be a solution of the minimization problem* (1.1). *Then there exists a dual optimal*

multiplier $\hat{u} \in R^m_+$ *for* (1.1) *such that*

$$\|\hat{u}\|_\infty \leqq \|\hat{u}\|_1 \leqq \frac{f(x^2)-f(\bar{x})}{\min_{i\in I_1} -g_i(x^2)}$$

(2.2)
$$+\left(\|\nabla f(\bar{x})\|_1 + \frac{f(x^2)-f(\bar{x})}{\min_{i\in I_1} -g_i(x^2)}\|\nabla g_{I_1}(\bar{x})\|_1\right) \max_{\substack{g(x)\leqq 0 \\ I(x)\in J(x)}} \|A^T_{I(x)}(A_{I(x)}A^T_{I(x)})^{-1}\|_1,$$

where

$$J(x) = \{I \mid I \subset I_2,\ A_I x = b_I,\ A_{i\in I} \text{ lin. indep.}\} \tag{2.3}$$

and $g_{I_2}(x) = A_{I_2}x - b_{I_2}$.

Proof. Since $\bar{x}$ is a solution of (1.1) and the relaxed Slater constraint qualification is satisfied, it follows that $\bar{x}$ and some $\bar{u} \in R^m_+$ constitute a Karush–Kuhn–Tucker saddlepoint for (1.1). Since f and g are differentiable, it follows that

$$\nabla f(\bar{x}) + \bar{u}_{I_1}\nabla g_{I_1}(\bar{x}) + \bar{u}_{I_2}A_{I_2} = 0, \quad \bar{u}g(\bar{x}) = 0, \qquad g(\bar{x}) \leqq 0, \quad \bar{u} \geqq 0. \tag{2.4}$$

By the fundamental theorem on the existence of basic feasible solutions [3, Thm. 2.11] it follows that there exists $\hat{u} \in R^m_+$ such that $(\bar{x}, \hat{u})$ is a Karush–Kuhn–Tucker saddlepoint of (1.1) and

$$\nabla f(\bar{x}) + \hat{u}_{I_1}\nabla g_{I_1}(\bar{x}) + \hat{u}_{I(\bar{x})}A_{I(\bar{x})} = 0, \qquad \hat{u}_{i\notin I_1 \cup I(\bar{x})} = 0, \tag{2.5}$$

where $I(\bar{x})$ belongs to $J(\bar{x})$ as defined by (2.3). Hence

$$\hat{u}_{I(\bar{x})} = -(\nabla f(\bar{x}) + \hat{u}_{I_1}\nabla g_{I_1}(\bar{x}))A^T_{I(\bar{x})}(A_{I(\bar{x})}A^T_{I(\bar{x})})^{-1}. \tag{2.6}$$

Consequently

$$\|\hat{u}_{I(\bar{x})}\|_1 \leqq (\|\nabla f(\bar{x})\|_1 + \|\hat{u}_{I_1}\|_1\|\nabla g_{I_1}(\bar{x})\|_1)\|A^T_{I(\bar{x})}(A_{I(\bar{x})}A^T_{I(\bar{x})})^{-1}\|_1. \tag{2.7}$$

From the saddlepoint property we have that

$$f(\bar{x}) \leqq f(x^2) + \hat{u}_{I_1}g_{I_1}(x^2) + \hat{u}_{I(\bar{x})}g_{I(\bar{x})}(x^2) \leqq f(x^2) - \|\hat{u}_{I_1}\|_1 \min_{i\in I_1} -g_i(x^2).$$

Hence

$$\|\hat{u}_{I_1}\|_1 \leqq \frac{f(x^2)-f(\bar{x})}{\min_{i\in I_1} -g_i(x^2)}. \tag{2.8}$$

Combining (2.7) and (2.8) gives

$$\|\hat{u}\|_\infty \leqq \|\hat{u}\|_1 \leqq \frac{f(x^2)-f(\bar{x})}{\min_{i\in I_1} -g_i(x^2)} + \left(\|\nabla f(\bar{x})\|_1 + \frac{f(x^2)-f(\bar{x})}{\min_{i\in I_1} -g_i(x^2)}\|\nabla g_{I_1}(\bar{x})\|_1\right)\|A^T_{I(\bar{x})}(A_{I(\bar{x})}A^T_{I(\bar{x})})^{-1}\|_1.$$

Inequality (2.2) follows from the above upon replacing the last term by its maximum over all feasible x. □

It is evident that the last term in (2.2) may be difficult to compute because of its combinatorial aspect. However if there are only a few linear constraints, or if the point x^2 is interior to most of the linear constraints, in which case these constraints can be lumped with the nonlinear constraints, it may not be too difficult to compute the bound of (2.2). Obviously since $\bar{x}$ is unknown beforehand, $f(\bar{x})$ must be replaced by a lower

bound (as must be done for Zangwill's bound $\bar{\alpha}_1$) and $\|\nabla f(\bar{x})\|_1$ and $\|\nabla g_{I_1}(\bar{x})\|_1$, by upper bounds in (2.2).

3. Exact penalty characterization without assuming solvability of the minimization problem. In this section we characterize solutions of the minimization problem (1.1) in terms of minimizers of the exact penalty function (1.2) *without* any a priori assumption regarding the existence of solutions to (1.1) as was the case in the previous section. We do however need the relaxed Slater constraint qualification (1.6) and a mild asymptotic constraint qualification (3.2) below, which is automatically satisfied if all the constraints are linear. It is interesting to note that the threshold value $\bar{\alpha}_3$ of the penalty parameter in Theorem 3.1 below exceeds or equals the threshold value $\bar{\alpha}_2 := \|\bar{u}\|_\infty$ of Theorem 2.1.

THEOREM 3.1 (Exact penalty characterization of feasible convex programs). *Let $f: R^n \to R$ and $g: R^n \to R^m$ be differentiable convex functions on R^n. Let the relaxed Slater constraint qualification* (1.6) *hold, let*

$$0 \neq \beta := \sup_x \{\|\nabla f(x)\|_1 | g(x) \leqq 0\} < \infty \tag{3.1}$$

and let the following asymptotic constraint qualification [8] *hold*:

> *For each nonempty $I \subset \{1, \cdots, m\}$ and each sequence of points $\{x^i\}$ such that: $g(x^i) \leqq 0$, $g_I(x^i) = 0$ and $\nabla g_{j \in I}(x^i)$ are linearly independent, each accumulation point $(\overline{\nabla g}_{I_0}, \overline{\nabla g}_{I_1}, \overline{\nabla g}_{I_2})$ of the sequence $\{\nabla g_{j \in I_0}(x^i)/\|\nabla g_{j \in I_0}(x^i)\|, \nabla g_{I_1}(x^i), \nabla g_{I_2}(x^i)\}$ satisfies*
>
> $$\overline{\nabla g}_{I_0} z > 0, \overline{\nabla g}_{I_1} z > 0, \overline{\nabla g}_{I_2} z \geqq 0 \text{ for some } z \in R^n \tag{3.2}$$
>
> *where $I_0 \cup I_1 \cup I_2$ is a partition of I such that the sequence $\{\nabla g_j(x^i)\}$ is unbounded for $j \in I_0$ and bounded for $j \in I_1$, $g_{j \in I_0 \cup I_1}$ is nonlinear and $g_{j \in I_2}$ is linear.*

A necessary (sufficient) condition for $\bar{x} \in R^n$ to solve the minimization problem (1.1) *is that $\bar{x}$ minimizes $P(x, \alpha)$* over x in R^n *for all $\alpha \geqq \bar{\alpha}_3$ (some $\alpha > \bar{\alpha}_3$) where*

$$\bar{\alpha}_3 := \beta \sup_{w,p,I} \{\|w_I\|_\infty | g(p) \leqq 0,\ w_I > 0,\ g_I(p) = 0,\ \|w_I \nabla g_I(p)\|_1 = 1, \nabla g_{j \in I}(p) \text{ lin. indep.}, I \subset \{1, \cdots, m\}\}. \tag{3.3}$$

Proof. We first note that the finiteness of $\bar{\alpha}_3$ is ensured by the asymptotic constraint qualification [8, Thm. 2.1].

Necessity. Let $\bar{x}$ be a solution of (1.1) and let $\bar{u} \in R^m_+$ be an optimal dual multiplier for (1.1) chosen as indicated below. We will show that $\bar{\alpha}_3 \geqq \|\bar{u}\|_\infty$ and hence by the necessity part of Theorem 2.1, $\bar{x}$ minimizes $P(x, \alpha)$ for $\alpha \geqq \bar{\alpha}_3$. If $\nabla f(\bar{x}) = 0$, we take $\bar{u} = 0$ and evidently $\bar{\alpha}_3 \geqq \|\bar{u}\|_\infty = 0$. Suppose now $\nabla f(\bar{x}) \neq 0$. Take $\bar{u} = (\bar{u}_L, \bar{u}_K)$ where $\bar{u}_L > 0$ and corresponding to "basic" $g_{j \in L}(\bar{x}) = 0$ such that $\nabla g_{j \in L}(\bar{x})$ are linearly independent and $\bar{u}_K = 0$. Hence by the Karush–Kuhn–Tucker conditions [7]

$$\nabla f(\bar{x}) + \bar{u}_L \nabla g_L(\bar{x}) = 0$$

and consequently

$$\left\| \frac{\bar{u}_L}{\|\nabla f(\bar{x})\|_1} \nabla g_L(\bar{x}) \right\|_1 = 1.$$

Hence by the definition (3.3) of $\bar{\alpha}_3$ and the definition (3.1) of β

$$\bar{\alpha}_3 \geqq \beta \left\| \frac{\bar{u}_L}{\|\nabla f(\bar{x})\|_1} \right\|_\infty = \beta \frac{\|\bar{u}\|_\infty}{\|\nabla f(\bar{x})\|_1} \geqq \|\bar{u}\|_\infty =: \bar{\alpha}_2.$$

Sufficiency. Let $\bar{x}$ be a solution of $\min_{x \in R^n} P(x, \alpha)$ for some $\alpha > \bar{\alpha}_3$. We first show by contradiction, that $g(\bar{x}) \leqq 0$. For if $\bar{x}$ is infeasible, then by [8, Thm. 2.1] there exists a feasible $p(\bar{x})$ such that

$$\|\bar{x} - p(\bar{x})\|_\infty \leqq \frac{\bar{\alpha}_3}{\beta} \|g(\bar{x})_+\|_1 = \frac{\bar{\alpha}_3}{\beta} eg(\bar{x})_+. \tag{3.4}$$

Then for $\alpha > \bar{\alpha}_3$,

$$\begin{aligned}
f(p(\bar{x})) &= P(p(\bar{x}), \alpha) \\
&\geqq P(\bar{x}, \alpha) \qquad \text{(since } \bar{x} \text{ minimizes } P(x, \alpha) \text{ over } x \in R^n) \\
&= f(\bar{x}) + \alpha eg(\bar{x})_+ \\
&> f(\bar{x}) + \beta \|\bar{x} - p(\bar{x})\|_\infty \qquad \text{(by (3.4), } \alpha > \bar{\alpha}_3 \text{ and } g(\bar{x})_+ \neq 0) \\
&\geqq f(\bar{x}) + \|\nabla f(p(\bar{x}))\|_1 \|\bar{x} - p(\bar{x})\|_\infty \qquad \text{(by (3.1))} \\
&\geqq f(\bar{x}) - \nabla f(p(\bar{x}))(\bar{x} - p(\bar{x})) \qquad \text{(by the Cauchy–Schwarz inequality)} \\
&\geqq f(p(\bar{x})) \qquad \text{(by the convexity of } f)
\end{aligned}$$

which is a contradiction. Hence $g(\bar{x}) \leqq 0$ and $\bar{x}$ is feasible. For any other feasible x and $\alpha > \bar{\alpha}_3$

$$f(\bar{x}) = P(\bar{x}, \alpha) \leqq P(x, \alpha) = f(x)$$

and hence $\bar{x}$ solves (1.1). □

Obviously the threshold value $\bar{\alpha}_3$ given by (3.2) is difficult to compute in general. However besides providing an existence result for the minimization problem (1.1), it is useful to know that such a threshold value exists and to know how it depends on the problem parameters, especially when one is engaged in an unconstrained exact penalty function minimization either on R^n, as a substitute for the original constrained optimization problem, or on R^1 as part of an iterative method [5]. In both cases an α such that $\alpha > \bar{\alpha}_3$ would be a useful upper bound to the penalty parameters employed. This would avoid the use of arbitrarily large penalty parameters that may lead to numerical difficulties.

4. An application: The big-M method for convex programs. In linear programming, a well-known method [9], [1] for solving a linear program without an explicit phase I procedure is to add nonnegative artificial variables to the constraints and then add a penalty to the objective function involving the artificial variables. If the penalty parameter is "sufficiently large", then the artificial variables will be driven to zero and an optimal solution will be obtained, if one exists. In this section we will make the "sufficiently large" concept precise by using the results of the two previous sections and extend the idea of the big-M method to convex programs. We first state a simple lemma whose elementary proof we omit.

LEMMA 4.1. *Let $f: R^n \to R$, $g: R^n \to R^m$ and let $\alpha > 0$. Then the problems*

$$\min_{x \in R^n} f(x) + \alpha eg(x)_+ =: \min_{x \in R^n} P(x, \alpha), \tag{4.1}$$

$$\min_{(x,z) \in R^{n+m}} f(x) + \alpha ez \quad \text{s.t. } g(x) \leqq z,\ z \geqq 0 \tag{4.2}$$

are equivalent in the following sense: *For each solution* $\bar{x}$ *of* (4.1), $(\bar{x}, \bar{z} := g(\bar{x})_+)$ *solves* (4.2), *and for each solution* $(\tilde{x}, \tilde{z})$ *of* (4.2), $\tilde{x}$ *solves* (4.1).

The formulation of (4.2) is the big-M formulation and is used in linear programming because it is easy to obtain a feasible point for it by taking any x in R^n and $z := g(x)_+$. Formulation (4.2) can be used also for the very same reason in convex programming. Theorem 2.1 tells us that if we know a priori that problem (1.1) has a solution, f and g are convex and the relaxed Slater constraint qualification (1.6) is satisfied, then the penalty parameter α of the big-M formulation (4.2) must satisfy $\alpha > \bar{\alpha}_2 := \|\bar{u}\|_\infty$ where $\bar{u}$ is any optimal dual multiplier to (1.1). Note that if g is linear, then the relaxed Slater constraint qualification (1.6) is satisfied by any feasible point x. If we have no a priori knowledge that (1.1) is solvable, but that it is merely feasible, that f, g are differentiable and convex, and that (3.1) and the constraint qualifications (1.6) and (3.2) are satisfied, then the penalty parameter α of the big-M method (4.2) must satisfy $\alpha > \bar{\alpha}_3$ where $\bar{\alpha}_3$ is defined by (3.3). Note that if g is linear, then (1.6) and (3.2) are automatically satisfied, and if in addition f is nonconstant and linear, then (3.1) is also automatically satisfied.

REFERENCES

[1] M. S. BAZARAA AND J. J. JARVIS, *Linear Programming and Network Flows*, John Wiley, New York, 1977.

[2] A. V. FIACCO AND G. P. MCCORMICK, *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, John Wiley, New York, 1968.

[3] D. GALE, *The Theory of Linear Economic Models*, McGraw-Hill, New York, 1960.

[4] A. M. GEOFFRION, *Duality in nonlinear programming: A simplified applications-oriented development*, SIAM Rev., 13 (1971), pp. 1–37.

[5] S.-P. HAN, *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22 (1977), pp. 297–309.

[6] S.-P. HAN AND O. L. MANGASARIAN, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251–269.

[7] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[8] ———, *A condition number for differentiable convex inequalities*, Computer Sciences Technical Report 491, Univ. Wisconsin, Madison, revised July 1983; Math. Oper. Res., to appear.

[9] K. MURTY, *Linear and Combinatorial Programming*, John Wiley, New York, 1976.

[10] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[11] W. ZANGWILL, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

004

A HIGH ORDER TEST FOR OPTIMALITY OF BANG–BANG CONTROLS*

ALBERTO BRESSAN†

Abstract. For control systems of the form $\dot{x} = X(x) + \sum_{i=1}^{m} Y_i(x)u_i$, a strengthened version of the classical Pontryagin maximum principle is proved. The necessary condition for optimality given here is obtained using functional analytic techniques and quite general high order perturbations of the reference control. As shown by an example, this test is particularly effective when applied to bang–bang controls, a case where other high order tests do not provide additional information.

AMS subject classifications. 49B10, 49B27

Key words. admissible variational family, high-order tangent vector

1. Introduction. Let U be a closed convex subset of the Banach space $\mathscr{L}^1([0, T]; \mathbb{R}^m)$ and consider a continuously Fréchet differentiable mapping $\varphi : U \to \mathbb{R}^n$. Given $\bar{u} \in U$, in this paper we give a high-order sufficient condition for $\varphi(\bar{u})$ to belong to the interior of the image $\varphi(U)$. Problems of this kind arise frequently in control theory. Indeed, consider a control system of the form

$$\text{(S)} \qquad \begin{aligned} &\dot{x}(t) = f(x(t)) + G(x(t))u(t), \\ &x(0) = 0, \qquad u(t) \in \Omega \quad \text{for a.e. } t \in [0, T], \end{aligned}$$

where Ω is a compact convex subset of $\mathbb{R}^m$ and f, G are $\mathscr{C}^1$ mappings from $\mathbb{R}^n$ into $\mathbb{R}^n$ and $\mathbb{R}^n \times \mathbb{R}^m$ respectively. If T is small enough, then (S) yields a $\mathscr{C}^1$ map $\psi : u \to x(u, T)$ from the set U of admissible controls into $\mathbb{R}^n$. Here $x(u, T)$ is the point reached at time T by the trajectory of (S) corresponding to the control u. A classical problem is then the following: given an admissible control $\bar{u}$, decide whether $\bar{u}$ is time-optimal. This is often equivalent to showing that $x(\bar{u}, T)$ lies on the boundary of the reachable set $R(T)$.

A well-known necessary condition for optimality is given by the Pontryagin maximum principle (PMP) [2], [8]. Krener's high-order maximal principle (HMP) [6] provides further conditions, obtained from the study of more general one-parameter perturbations u_ξ of the control $\bar{u}$. If the first-order variation at the terminal point of the trajectory

$$\lim_{\xi \to 0} [x(u_\xi, T) - x(\bar{u}, T)]/\xi \tag{1.1}$$

vanishes, a high order tangent vector can be generated, and additional necessary conditions for extremality are found. This method yielded several new results [3], [4], [5], [6], especially concerning the problem of local stability. In this case, the reference control is $\bar{u}(t) \equiv 0$ and lies in the interior of $\Omega = [-1, 1]$. Hence there are several ways to locally perturb $\bar{u}$ and achieve a cancellation in the first order variation (1.1). The HMP can be here particularly effective. On the other hand, if $\bar{u}$ is bang–bang, $\bar{u}(t)$ already lies on the boundary of Ω, and only one-sided perturbations of $\bar{u}$ are admissible. As a result, in general there is no way of generating high order tangent vectors, as long as only the "instantaneous" control variations considered in [5], [6] are used. In order to develop a genuine high order test for optimality of bang–bang controls, it is

* Received by the editors April 21, 1983, and in revised form January 2, 1984.

† Istituto di Matematica Applicata, Università di Padova, 35100 Italy. This research was supported by the Consiglio Nazionale delle Ricerche, G.N.A.F.A. and by the U.S. Army under contract DAAG29-80-0041.

necessary to achieve the cancellation of the first-order variation (1.1) by perturbing $\bar{u}$ simultaneously in the neighborhoods of two or more distinct times. This leads us to consider more general control variations.

In the following, the variable t always denotes time, while ξ, c are used as variational parameters: $u(\xi, \cdot)$ or $u(c, \xi, \cdot)$ will denote controls in $\mathscr{L}^1([0, T]; \mathbb{R}^m)$ depending continuously on the parameters ξ, c. In the abstract setting considered in §§ 2, 3, the control u is regarded merely as a point in a Banach space E, and we use the shorter notation $u(\xi)$ or $u(c, \xi)$ to indicate its dependence on one or two parameters.

DEFINITION 1. A one-parameter admissible variational family of control functions (AVF) for a control $\bar{u}$ on $[0, T]$, generating a tangent vector $v \in \mathbb{R}^n$, is a continuous map ν: $\xi \to u(\xi, \cdot)$ from a nondegenerate interval $[0, \bar{\xi}]$ into $\mathscr{L}^1([0, T]; \mathbb{R}^m)$ such that

$$u(0, \cdot) = \bar{u}(\cdot), \qquad u(\xi, \cdot) \in U \quad \forall \xi \in [0, \bar{\xi}], \tag{1.2}$$

$$\lim_{\xi \to 0} [x(u(\xi, \cdot), T) - x(\bar{u}, T)]/\xi = v. \tag{1.3}$$

We say that the AVF ν has order h if there exist constants C_1, C_2 for which

$$0 < C_1 \leqq \xi^{-1} \| u(\xi, \cdot) - \bar{u}(\cdot) \|_{\mathscr{L}^1}^h \leqq C_2 \quad \forall \xi \in (0, \bar{\xi}]. \tag{1.4}$$

Notice that one can recover every high order tangent vector by means of the first-order derivative (1.3), via a suitable change of the parameter ξ. As shown in [5], this method differs from Krener's only in computational ease. The above class of AVF is at the same time simpler and more general than those studied in [5], [6], hence the corresponding family of tangent vectors can be much larger. One would like to use all of these vectors to derive a stronger HMP. Assume that, given suitable variational families ν_i for $\bar{u}$ $(i = 0, \cdots, k)$, the positive span of the corresponding tangent vectors v_i is all of $\mathbb{R}^n$. To conclude that $x(\bar{u}, T)$ lies in the interior of the reachable set $R(T)$, one has to construct approximate convex combinations of the v_i continuously depending on the parameters. More precisely, the ν_i should be *summable* in the sense of the next definition.

DEFINITION 2. Let $\mathscr{F} = \{\nu_0, \cdots, \nu_k\}$ be a finite collection of AVF for the control $\bar{u}$, generating the tangent vectors $v_0, \cdots, v_k$. Set

$$\Delta^k = \left\{ c = (c_0, \cdots, c_k);\ c_i \geqq 0,\ \sum_{i=0}^{k} c_i = 1 \right\}. \tag{1.5}$$

$\mathscr{F}$ is summable if there exist $\bar{\xi} > 0$ and a continuous map $(c, \xi) \to u(c, \xi, \cdot)$ from $\Delta^k \times [0, \bar{\xi}]$ into $\mathscr{L}^1([0, T]; \mathbb{R}^m)$ such that, for all $c \in \Delta^k$,

$$u(c, 0, \cdot) = \bar{u}(\cdot), \qquad u(c, \xi, \cdot) \in U \quad \forall \xi \in [0, \bar{\xi}], \tag{1.6}$$

$$\lim_{\xi \to 0} [x(u(c, \xi, \cdot), T) - x(\bar{u}(\cdot), T)]/\xi = \sum_{i=0}^{k} c_i v_i \tag{1.7}$$

uniformly on Δ^k.

This crucial property holds for variational families of the special kind considered in [5], [6], but is not satisfied by an arbitrary collection of AVF (see § 5 for a counterexample). Our key result is that if all but one of the $\nu_i \in \mathscr{F}$ have order 1, then $\mathscr{F}$ is summable. This is first proven in an abstract setting, then stated for the control system (S). We thus obtain a strengthened version of the PMP which is particularly effective when applied to bang–bang controls. Indeed, our single high order variational family is allowed to be quite arbitrary. An application of this technique is given in § 5.

2. Notation, statement of the main results. Consider a mapping φ from a neighborhood of a closed subset U of a Banach space E into $\mathbb{R}^n$ and denote by $D\varphi(u)$ its differential at u. We say that φ is $\mathscr{C}^1$ if the map $u \to D\varphi(u)$ from E into the space of continuous linear operators $L(E; \mathbb{R}^n)$ is continuous. For the definition of the operator norm on $L(E; \mathbb{R}^n)$ and for the basic properties of differentials our general reference is Diéudonne [1].

If $\bar{u} \in U$, by an admissible variational family (AVF) for $\bar{u}$, generating a tangent vector $v \in \mathbb{R}^n$, we mean a continuous map $\nu: \xi \to u(\xi)$ from $[0, 1]$ into U such that

$$u(0) = \bar{u}, \qquad u(\xi) \in U \quad \forall \xi \in [0, 1], \tag{2.1}$$

$$\lim_{\xi \to 0} [\varphi(u(\xi)) - \varphi(\bar{u})]/\xi = v. \tag{2.2}$$

If, for some $0 < C_1 \leqq C_2 < \infty$ and all $\xi \in (0, 1]$,

$$C_1 \leqq \xi^{-1} \|u(\xi) - \bar{u}\|^h \leqq C_2, \tag{2.3}$$

we say that ν has order h. We write $B(x, r)$ for the closed ball centered at x with radius r. The Euclidean norm on $\mathbb{R}^n$ and the operator norm on the set of $n \times m$ matrices are both written as $|\cdot|$, while double bars are used for the norms $\|\cdot\|$ in Banach spaces such as E or $L(E; \mathbb{R}^n)$. Int A, ∂A, $\overline{\text{co}}\, A$ denote the interior, the boundary and the convex closure of a set A. With these conventions we have the following theorem.

THEOREM 1. *Let U be a closed convex subset of a Banach space E, and let φ be a $\mathscr{C}^1$ mapping from a neighborhood of U into $\mathbb{R}^n$. Assume $\bar{u} \in U$ and let $\nu_i: \xi \to u_i(\xi)$ be* AVF *for $\bar{u}$ generating the tangent vectors v_i $(i = 0, \cdots, k)$. If $0 \in \text{int}\, \overline{\text{co}}\, \{v_0, \cdots, v_k\} \subseteq \mathbb{R}^n$ and $\nu_1, \cdots, \nu_k$ have order 1, then $\varphi(\bar{u}) \in \text{int}\, \varphi(U)$.*

From this result, a sharper form of Pontryagin's maximum principle for the system (S) can be derived. To fix the ideas, assume that f and G are $\mathscr{C}^1$ on $B(0, r) \subset \mathbb{R}^n$; f_x, G_x will denote the corresponding differentials. Let

$$\sup \{|f(x)| \vee |G(x)|;\, x \in B(0, r)\} < M_1,$$

$$\sup \{|w|;\, w \in \Omega\} < M_2, \qquad 0 < T < r(M_1 + M_1 M_2)^{-1}.$$

This guarantees that, for every control u in the admissible set

$$U = \{u \in \mathscr{L}^1([0, T]; R^m);\, u(t) \in \Omega \text{ a.e.}\},$$

$\|u\| < M_2 T$ and there exists a unique solution $t \to x(u, t)$ of (S) defined on $[0, T]$, taking values inside $B(0, r)$. Notice that the open ball $\mathscr{B} = \{u \in \mathscr{L}^1;\, \|u\| < M_2 T\}$ is a neighborhood of U. We assume that Ω is closed, bounded and convex, thus the same holds for U. The map $\psi: \mathscr{B} \to \mathscr{C}^0([0, T]; \mathbb{R}^n)$ that associates to each control u the corresponding solution $x(u, \cdot)$ of (S) is continuously Fréchet differentiable. Indeed, ψ is implicitly defined by the equation $\psi(u) = \Psi(u, \psi(u))$, with

$$\Psi(u, x)(t) = \int_0^t f(x(s))\, ds + \int_0^t G(x(s))u(s)\, ds. \tag{2.4}$$

The map Ψ can be thought of as the composition $\Psi_2 \cdot \Psi_1$, defined by

$$\Psi_1(u, x)(t) = (u(t), f(x(t)), G(x(t))),$$

$$\Psi_2(u, y_1, y_2)(t) = \int_0^t y_1(s)\, ds + \int_0^t y_2(s)u(s)\, ds.$$

Clearly Ψ_1 is $\mathscr{C}^1$ and Ψ_2 is bilinear. Hence Ψ is $\mathscr{C}^1$ and the same holds for ψ, because of the implicit function theorem ([1, p. 272]). An application of Theorem 1 yields Theorem 2.

THEOREM 2. *Let $\bar{u}$ be an admissible control for the system* (S) *and assume that $x(\bar{u}, T) \in \partial R(T)$. Then, for every tangent vector v_0 generated by a (possibly high order)* AVF ν_0 *for $\bar{u}$, there exists an absolutely continuous nontrivial n-vector valued function $t \to \lambda(t)$ on $[0, T]$ which satisfies*

$$\lambda(T) \cdot v_0 \leqq 0, \tag{2.5}$$

$$\dot{\lambda}(t) = -\lambda(t) \cdot [f_x(x(\bar{u}, t)) + G_x(x(\bar{u}, t))\bar{u}(t)], \tag{2.6}$$

$$\lambda(t) \cdot G(x(\bar{u}, t))\bar{u}(t) = \max \{\lambda(t) \cdot G(x(\bar{u}, t))w;\ w \in \Omega\}, \tag{2.7}$$

for almost every t in $[0, T]$.

3. Proof of Theorem 1. It is certainly not restrictive to assume that $k \geqq 1$ and that all vectors v_i are nontrivial. Relying on the fact that $\nu_1, \cdots, \nu_k$ have order 1 we first prove Lemma 1.

LEMMA 1. *The collection of admissible variational families $\mathscr{F} = \{\nu_0, \cdots, \nu_k\}$ is summable.*

Proof. Define the scalar function α by setting

$$\alpha(\xi) = \sup \{\|u_0(\zeta) - \bar{u}\|^{1/2};\ 0 \leqq \zeta \leqq \xi\}. \tag{3.1}$$

Clearly α is a continuous, nondecreasing function with $\alpha(0) = 0$. The existence of a first order tangent vector $v_1 \neq 0$ implies $D\varphi(\bar{u}) \neq 0$. By (2.2), for $\xi > 0$ small enough, we thus have

$$\|u_0(\xi) - \bar{u}\|/\xi \geqq |v_0|/2\|D\varphi(\bar{u})\|. \tag{3.2}$$

Therefore there exists a $\bar{\xi} > 0$ such that

$$\left(\frac{|v_0| \cdot \xi}{2\|D\varphi(\bar{u})\|}\right)^{1/2} \leqq \alpha(\xi) \leqq 1, \qquad \xi/\alpha(\xi) \leqq 1 \tag{3.3}$$

for all $\xi \in (0, \bar{\xi}]$. Define $u(c, \xi)$ on $\Delta^k \times [0, \bar{\xi}]$ by

$$u(c, \xi) = u_0(c_0\xi) + \sum_{i=1}^{k} c_i(\xi/\alpha(\xi))[u_i(\alpha(\xi)) - u_0(c_0\xi)]$$

if $0 < \xi \leqq \bar{\xi}$,

$$u(c, 0) = \bar{u}. \tag{3.4}$$

By (3.3), $u(c, \xi)$ is well defined and takes values inside U, being a convex combination of members of U. As $\xi \to 0$, $u_0(c, \xi)$ tends to $\bar{u}$ and each term inside the summation in (3.4) tends to zero uniformly w.r.t. c. Therefore u depends continuously on the parameters c, ξ. To show (1.7) we write

$$\frac{\varphi(u(c, \xi)) - \varphi(\bar{u})}{\xi} = \frac{\varphi(u_0(c_0\xi)) - \varphi(\bar{u})}{\xi} + \frac{\varphi(u(c, \xi)) - \varphi(u_0(c_0\xi))}{\xi}. \tag{3.5}$$

As $\xi \to 0$, the first term on the right-hand side of (3.5) converges to $c_0 v_0$. The second

term can be written as

$$\begin{aligned}(3.6)\qquad &\int_0^1 D\varphi(\theta u(c,\xi)+(1-\theta)u_0(c_0\xi))\circ[u(c,\xi)-u_0(c_0\xi)]\,d\theta\\ &=\int_0^1 [D\varphi(\bar u)+\chi(c,\xi,\theta)]\circ\left\{\sum_{i=1}^k c_i(1/\alpha(\xi))[u_i(\alpha(\xi))-u_0(c_0\xi)]\right\}d\theta.\end{aligned}$$

The continuous Fréchet differentiability of φ implies that $\chi(c,\xi,\theta)=D\varphi(\theta u(c,\xi)+(1-\theta)u_0(c_0\xi))-D\varphi(\bar u)$ is a continuous linear operator whose norm tends to zero uniformly in c, θ as $\xi\to 0$. Observe that

$$\begin{aligned}(3.7)\qquad (1/\alpha(\xi))\|u_i(\alpha(\xi))-u_0(c_0\xi)\|&\leqq(1/\alpha(\xi))\|u_i(\alpha(\xi))-\bar u\|+(1/\alpha(\xi))\|u_0(c_0\xi)-\bar u\|\\ &\leqq K_i+\|u_0(c_0\xi)-\bar u\|^{1/2}\end{aligned}$$

for some finite constants K_i $(i=1,\cdots,k)$, because the AVF ν_i have order 1 and by (3.1). The limit as $\xi\to 0$ of the last term in (3.5) is therefore given by

$$(3.8)\qquad \lim_{\xi\to 0} D\varphi(\bar u)\cdot\sum_{i=1}^k\left(\frac{c_i}{\alpha(\xi)}\right)(u_i(\alpha(\xi))-\bar u).$$

By the definition (2.2) of tangent vector, one has

$$\begin{aligned}(3.9)\qquad v_i&=\lim_{\xi\to 0}\frac{\varphi(u_i(\xi))-\varphi(\bar u)}{\xi}=\lim_{\xi\to 0}\frac{D\varphi(\bar u)\cdot(u_i(\xi)-\bar u)+o(\xi)}{\xi}\\ &=\lim_{\xi\to 0} D\varphi(\bar u)\cdot(u_i(\xi)-\bar u)/\xi.\end{aligned}$$

Indeed u_i is a first order AVF, hence the term $o(\xi)$, which is infinitesimal of higher order w.r.t. $\|u_i(\xi)-\bar u\|$ as $\xi\to 0$, is also of higher order w.r.t. ξ. Comparing (3.9) with (3.8) one concludes that

$$\begin{aligned}(3.10)\qquad &\lim_{\xi\to 0}[\varphi(u(c,\xi))-\varphi(\bar u)]/\xi\\ &\quad=c_0v_0+\sum_{i=1}^k c_i\cdot\lim_{\xi\to 0} D\varphi(\bar u)\circ[u_i(\alpha(\xi))-\bar u]/\alpha(\xi)=\sum_{i=0}^k c_iv_i,\end{aligned}$$

uniformly on Δ^k.

Using the above lemma, the proof of Theorem 1 can now be completed by an application of Brouwer's fixed point theorem.

LEMMA 2. *Assume that $\varpi\in\mathbb{R}^n$ and that $v_0,\cdots,v_k\in\mathbb{R}^n$ satisfy*

$$(3.11)\qquad 0\in\operatorname{int}\overline{\operatorname{co}}\{v_0,\cdots,v_k\}.$$

Let ω be a continuous map from $\Delta^k\times[0,\bar\xi]$ into $\mathbb{R}^n$ such that

$$\omega(c,0)=\varpi\quad\forall c\in\Delta^k,$$

$$\lim_{\xi\to 0}[\omega(c,\xi)-\varpi]/\xi=\sum_{j=0}^k c_jv_j$$

uniformly on Δ^k. Then the image $\omega(\Delta^k\times[0,\bar\xi])$ covers a whole neighborhood of ϖ in $\mathbb{R}^n$.

Proof. Clearly (3.11) implies $k\geqq n$. It is not restrictive to assume $k=n$. Indeed, if $k>n$, choose $n+1$ vectors $v_0',\cdots,v_n'\in\overline{\operatorname{co}}\{v_0,\cdots,v_k\}$ such that $0\in\operatorname{int}\overline{\operatorname{co}}\{v_0',\cdots,v_n'\}$. For $0\leqq i\leqq n$, let

$$v_i'=\sum_{j=0}^k a_{ij}v_j\quad\text{with } a_{ij}\geqq 0,\quad\sum_{j=0}^k a_{ij}=1.$$

The map $\omega' : \Delta^n \times [0, \bar{\xi}] \to \mathbb{R}^n$ defined by

$$\omega'(c', \xi) = \omega(c, \xi) \quad \text{with } c_j = \sum_{i=0}^{n} a_{ij} c'_i, \quad 0 \leqq j \leqq k,$$

is continuous and satisfies $\omega'(c', 0) = \varpi$ for all $c' \in \Delta^n$. Moreover

$$\lim_{\xi \to 0} [\omega'(c', \xi) - \varpi]/\xi = \sum_{j=0}^{k} \left(\sum_{i=0}^{n} a_{ij} c'_i \right) v_j = \sum_{i=0}^{n} c'_i v'_i$$

uniformly on Δ^n, and the image of ω' is clearly contained in the image of ω. By possibly replacing ω with ω', it therefore suffices to prove the lemma in the special case $k = n$.

Let $\delta = \operatorname{dist}(0, \partial \overline{\operatorname{co}}\{v_0, \cdots, v_n\})$ and choose ξ_0 so small that

$$\left| \omega(c, \xi_0) - \varpi - \xi_0 \sum_{i=0}^{n} c_i v_i \right| < \xi_0 \delta / 2 \tag{3.12}$$

for all $c \in \Delta^n$. Consider the injective map $\sigma : \Delta^n \to \mathbb{R}^n$ defined by

$$\sigma(c) = \varpi + \xi_0 \cdot \sum_{i=0}^{n} c_i v_i.$$

For $x \in B(\varpi, \xi_0 \delta)$ define $F(x) = \omega(\sigma^{-1}(x), \xi_0)$. By (3.12), $|F(x) - x| < \xi_0 \delta / 2$. For each $x_0 \in B(\varpi, \xi_0 \delta/2)$ an application of Brouwer's theorem ([8, p. 251]) now implies the existence of some $x \in B(\varpi, \xi_0 \delta)$ for which $F(x) = x_0$. We have thus shown that $B(\varpi, \xi_0 \delta/2) \subseteq \omega(\Delta^k \times [0, \bar{\xi}])$, proving the lemma.

By (3.10), Theorem 1 follows from Lemma 2 by setting $\varpi = \varphi(\bar{u})$, $\omega(c, \xi) = \varphi(u(c, \xi))$.

4. Proof of Theorem 2. Suppose that the conclusion is false. Then there exists an admissible variational family ν_0 for $\bar{u}$, possibly high order, that generates a tangent vector v_0 such that, for every absolutely continuous $\lambda(\cdot)$ satisfying (2.5) and (2.6), one has

$$\lambda(t) G(x(\bar{u}, t)) \bar{u}(t) < \max \{\lambda(t) G(x(\bar{u}, t)) w;\ w \in \Omega\} \tag{4.1}$$

for t in a subset $J \subseteq [0, T]$ having positive measure. For each vector $\eta \neq 0$ with $\eta \cdot v_0 \leqq 0$, let $\lambda_\eta(\cdot)$ be the unique solution of (2.6) for which $\lambda_\eta(T) = \eta$, and choose a control $u_\eta \in U$ such that

$$\lambda_\eta(t) G(x(\bar{u}, t)) u_\eta(t) = \max \{\lambda_\eta(t) G(x(\bar{u}, t)) w;\ w \in \Omega\} \tag{4.2}$$

for a.e. $t \in [0, T]$. The continuity of λ_η, G and $x(\bar{u}, \cdot)$ and a selection theorem [7] imply that such a measurable u_η exists. Define an AVF ν for $\bar{u}$ by setting

$$u(\xi, \cdot) = \xi u_\eta(\cdot) + (1 - \xi)\bar{u}(\cdot) \quad \forall \xi \in [0, 1]. \tag{4.3}$$

Then, for every ξ, $u(\xi) \in U$ because U is convex, and $\|u(\xi) - \bar{u}\|/\xi = \|u_\eta - \bar{u}\| \neq 0$, showing that ν has order one. Let $\Pi_T : \mathscr{C}^0[0, T] \to \mathbb{R}^n$ be the linear projection $x \to x(T)$. From the remarks made in § 2 it follows that the map $\xi \to x(u(\xi), T)$ is the composition of $\mathscr{C}^1$ mappings, hence the tangent vector generated by the AVF (4.3) exists and is given by

$$\begin{aligned} v &= \lim_{\xi \to 0} [x(u(\xi), T) - x(\bar{u}, T)]/\xi = \Pi_T \cdot D\psi(\bar{u}) \cdot (u_\eta - \bar{u}) \\ &= \int_0^T M(T, s) G(x(\bar{u}, s))(u_\eta(s) - \bar{u}(s))\, ds, \end{aligned} \tag{4.4}$$

where $s \to M(T, s)$ is the matrix fundamental solution of

$$\dot{z}(t) = -z(t) \cdot [f_x(x(\bar{u}, t)) + G_x(x(\bar{u}, t))\bar{u}(t)]$$

with $M(T, T) = I$, and ψ is the input–output map defined above (2.4). By (2.6) the inner product of η and v is

$$\begin{aligned} \eta \cdot v &= \int_0^T \lambda_\eta(T) M(T, s) G(x(\bar{u}, s))(u_\eta(s) - \bar{u}(s))\, ds \\ &= \int_0^T \lambda_\eta(s) G(x(\bar{u}, s))(u_\eta(s) - \bar{u}(s))\, ds > 0 \end{aligned}$$

because of (4.1). Hence, for every nontrivial vector η with $\eta \cdot v_0 \leqq 0$, there exists a first order tangent vector v for which $\eta \cdot v > 0$. A standard compactness argument yields the existence of finitely many first order tangent vectors $v_1, \cdots, v_k$ such that the positive span of $\{v_0, v_1, \cdots, v_k\}$ is the whole space $\mathbb{R}^n$. Theorem 1 applied to the $\mathscr{C}^1$ map $\varphi = \Pi_T \cdot \psi : u \to x(u, T)$ yields $x(\bar{u}, T) \in \operatorname{int} R(T)$, a contradiction.

5. Examples. The assumption on the order of the variational families in Theorem 1 is essential. Indeed, two arbitrary second order AVF need not be summable, as shown by the following example.

Example 1. Define a time-dependent system on $\mathbb{R}^3$ by

$$\begin{aligned} &(\dot{x}_1(t), \dot{x}_2(t), \dot{x}_3(t)) = (\varphi_2(t)x_3(t)u_1(t), \varphi_2(t)x_3(t)u_2(t), \varphi_1(t)u_3(t)), \\ &(x_1(0), x_2(0), x_3(0)) = (0, 0, 0), \end{aligned} \tag{5.1}$$

where $t \in [0, 3]$, the smooth function φ_1, φ_2 satisfy

$$\begin{aligned} &\varphi_1(t) = 0, \qquad \varphi_2(t) \geqq 0 \quad \text{for } t \in [1, 2], \\ &\varphi_2(t) = 0 \quad \text{for } t \in [0, 1] \cup [2, 3], \\ &\int_0^1 \varphi_1(t)\, dt = \int_1^2 \varphi_2(t)\, dt = -\int_2^3 \varphi_1(t)\, dt = 1 \end{aligned} \tag{5.2}$$

and the controls satisfy the constraints

$$0 \leqq u_i(t) \leqq 1 \quad (i = 1, 2), \qquad -\infty < u_3(t) < +\infty. \tag{5.3}$$

The reachable set at time $t = 3$ is then

$$R(3) = \{(x_1, x_2, x_3);\ x_1 x_2 \geqq 0\}. \tag{5.4}$$

Let $\bar{u}$ be the null control. Consider the two AVF for $\bar{u}$:

$$u^{(1)}(\xi, t) = (\xi^{1/2}, 0, \xi^{1/2}), \qquad u^{(2)}(\xi, t) = (0, \xi^{1/2}, -\xi^{1/2}),$$

constant on the time interval $[0, 3]$. Notice that for $i = 1, 2$

$$\|u^{(i)}(\xi) - \bar{u}\|^2 / \xi = \left(\int_0^3 u^{(i)}(\xi, t)\, dt \right)^2 \Big/ \xi = 18.$$

By setting $h = 2$, $C_1 = C_2 = 18$ in (1.4) one checks that $u^{(1)}$ and $u^{(2)}$ have order two. The endpoints of the corresponding trajectories are

$$x(u^{(1)}(\xi), 3) = (\xi, 0, 0), \qquad x(u^{(2)}(\xi), 3) = (0, -\xi, 0).$$

Hence $u^{(1)}$ and $u^{(2)}$ generate the tangent vectors

$$v_1 = (1, 0, 0), \qquad v_2 = (0, -1, 0). \tag{5.5}$$

Comparing (5.5) with (5.4), it is clear that these two AVFs cannot be summable. In this example, the set of high order tangent vectors of the special type considered in [6] is the cone $\Gamma = \{(0, 0, x_3);\ x_3 \in \mathbb{R}\}$. This is of course convex and coincides here with the first order tangent cone. Notice that the time dependency can be easily removed by adjoining a new variable $x_0 = t$.

We now illustrate a nontrivial application of Theorem 2 to the study of optimality of bang–bang controls.

Example 2. Consider the three-dimensional autonomous system with scalar control $u(t) \in [-1, 1]$:

$$(\dot{x}_1, \dot{x}_2, \dot{x}_3) = (u, x_1, x_2 + kx_1^2/2), \qquad (x_1(0), x_2(0), x_3(0)) = (0, 0, 0). \tag{5.6}$$

The adjoint equations for this system are

$$(\dot{\lambda}_1, \dot{\lambda}_2, \dot{\lambda}_3) = (-\lambda_2 - kx_1\lambda_3, -\lambda_3, 0). \tag{5.7}$$

If $|k| < 1$, then a theorem of H. Sussmann [9] yields the existence of a $T > 0$ such that every time optimal control $u(\cdot)$ on $[0, T]$ is bang–bang with at most two switchings. If $|k| > 1$, the above result does not apply. Indeed, for every $T > 0$, there exist bang–bang controls u that satisfy Pontryagin's necessary conditions for optimality and have an arbitrarily large number of switchings on $[0, T]$. In order to construct a regular feedback synthesis for (5.6) it is important to rule out the optimality of such controls. In this direction we prove the following proposition.

PROPOSITION 1. *Assume* $|k| > 1$. *Then every bang–bang control assuming the value* $+1$ *on a positive neighborhood of the origin is not optimal after its third switching time.*

Proof. Let $\bar{u}$ be a bang–bang control which is initially $+1$ and has at least 3 switchings, and let $0 < t_1 < t_2 < t_3$ be its first three switching times. Fix any $T > t_3$, smaller than the fourth switching time if there is any. We will prove that $x(\bar{u}, T) \in \operatorname{int} R(T)$. If the classical Pontryagin's necessary conditions do not hold for $\bar{u}$, we are done. Otherwise, let $\lambda(t) = (\lambda_1(t), \lambda_2(t), \lambda_3(t))$ be a nontrivial adjoint variable satisfying (2.6) and (2.7), given in this case by (5.7) and

$$\bar{u}(t) = \operatorname{sgn} \lambda_1(t) \quad \text{a.e. on } [0, T] \tag{5.8}$$

respectively. Our first task is to compute $\lambda(T)$. Set $t_0 = 0$, $t_4 = T$. From (5.7) it follows that the map $t \to \lambda(t)$ is $\mathscr{C}^1$ on $[0, T]$ and piecewise analytic on $[t_{i-1}, t_i]$ $(i = 1, \cdots, 4)$. In particular, we have

$$\lambda_3(t) = \lambda_3(0), \qquad \lambda_2(t) = \lambda_2(0) - t \cdot \lambda_3, \tag{5.9}$$

$$\ddot{\lambda}_1(t) = \lambda_3(1 - k \operatorname{sgn} \lambda_1(t)) \quad \text{a.e. on } [0, T], \tag{5.10}$$

$$\lambda_1(t_i) = 0 \qquad (i = 1, 2, 3). \tag{5.11}$$

Hence λ_1 is a polynomial of degree 2 in t on each subinterval $[t_{i-1}, t_i]$. If $\lambda_1(t) = 0$ for some $t \in (t_1, t_2)$, then we would have $\lambda(t) \equiv 0$, against the assumptions. Thus $\lambda_1(t) \neq 0$ for $t_1 < t < t_2$. By (5.11), $\ddot{\lambda}_1$ is not identically zero. Together with (5.10), this implies $\lambda_3 > 0$. Multiplying λ_3 by a positive scalar we can therefore assume $\lambda_3(t) \equiv 1$. This, together with (5.10) and (5.11), determines $\lambda_1(t)$ uniquely:

$$\lambda_1(t) = \frac{1+k}{2}(t - t_1)(t - t_2) \quad \text{for } t \in [t_1, t_2], \tag{5.12}$$

$$\lambda_1(t) = \frac{1-k}{2}(t - t_2)(t - t_3) \quad \text{for } t \in [t_2, t_3]. \tag{5.13}$$

The computation of $\dot\lambda_1(t_2)$ using alternatively (5.7), (5.12) and (5.13) yields

$$\begin{aligned}\dot\lambda_1(t_2) &= -\lambda_2(t_2) - kx_1(t_2) = -\lambda_2(t_2) - k(2t_1 - t_2)\\ &= \frac{k+1}{2}(t_2 - t_1) = \frac{k-1}{2}(t_3 - t_2).\end{aligned} \tag{5.14}$$

Notice that the above expressions coincide because λ is $\mathscr{C}^1$. From (5.7), (5.14) we deduce

$$\lambda_2(T) = t_1(1-3k)/2 + t_2(1+k)/2 - T. \tag{5.15}$$

For notational convenience, set $a = t_1$, $b = t_2 - t_1$, $c = t_3 - t_2$, $d = T - t_3$. So far, we have proven that, up to a positive scalar factor, there exists a unique adjoint variable $\lambda(t)$ that satisfies (2.6) and (2.7) on $[0, T]$. In particular, (5.15) and the last equality in (5.14) yield

$$\lambda(T) = \left(\lambda_1(T), -ka + \frac{k-1}{2}b - c - d, 1\right), \tag{5.16}$$

$$(k-1)b = (k-1)c. \tag{5.17}$$

The second part of the proof consists in the construction of a second order AVF for $\bar u$ generating at $t = T$ a tangent vector v having a positive inner product with $\lambda(T)$. A lengthy but elementary computation (see Appendix) shows that the control $\bar u$ steers the system from the origin to a point $x(\bar u, T)$ whose coordinates are

$$\begin{aligned} x_1(\bar u, T) &= a - b + c - d,\\ x_2(\bar u, T) &= T^2/2 - (b+c+d)^2 + (c+d)^2 - d^2,\\ x_3(\bar u, T) &= [T^3/2 - (b+c+d)^3 + (c+d)^3 - d^3]/3\\ &\quad + [a^3 + (b-a)^3 + (c-b+a)^3 + (d-c+b-a)^3/2]k/3. \end{aligned} \tag{5.18}$$

For $t \in [0, T]$ and $\xi > 0$ suitably small define

$$u(\xi, t) = 1 \quad \text{if } t \in [0, a + \xi^{1/2}c) \cup [a + b + \xi^{1/2}(b+c), T - d + \xi^{1/2}b),$$

$$u(\xi, t) = -1 \quad \text{if } t \in [a + \xi^{1/2}c, a + b + \xi^{1/2}(b+c)) \cup [T - d + \xi^{1/2}b, T].$$

The coordinates of $x(u(\xi, \cdot), T)$ are thus obtained from (5.18), replacing a, b, c, d by $a + \xi^{1/2}c$, $b + \xi^{1/2}b$, $c - \xi^{1/2}c$, $d - \xi^{1/2}b$ respectively.

Using (5.17), one checks that in the expression of $x(u(\xi, \cdot), T)$ all terms in $\xi^{1/2}$ cancel, hence the map $\xi \to u(\xi, \cdot)$ is an AVF for $\bar u$ of order 2. The computation of the corresponding tangent vector v defined by (1.3) yields (see Appendix)

$$v = (0, 2, b + c + 2d + k(2a - b + c))bc. \tag{5.19}$$

The inner product of (5.16) and (5.19) is

$$\lambda(T) \cdot v = (k-1)bc^2 > 0.$$

This shows that the necessary conditions for extremality stated in Theorem 2 do not hold for $\bar u$, hence $x(\bar u, T) \in \operatorname{int} R(T)$. For any $T > t_3$, $\bar u$ is not time optimal after T, therefore $\bar u$ is not optimal after its third switching time.

Appendix. For the control $\bar{u}$ considered in Example 2, the coordinates of the point $x(T, \bar{u})$ are:

$$x_1(T, \bar{u}) = \int_0^T \bar{u}(s)\, ds = a - b + c - d,$$

$$x_2(T, \bar{u}) = \int_0^T (T-s)\bar{u}(s)\, ds = T^2/2 - (b+c+d)^2 + (c+d)^2 - d^2,$$

$$\begin{aligned} x_3(T, \bar{u}) &= \int_0^T \frac{(T-s)^2}{2}\bar{u}(s)\, ds + \frac{k}{2}\int_0^T (x_1(s, \bar{u}))^2\, ds \\ &= \left[\int_0^a - \int_a^{a+b} + \int_{a+b}^{a+b+c} - \int_{a+b+c}^T\right]\frac{(T-s)^2}{2}\, ds \\ &\quad + \frac{k}{2}\left[\int_0^a s^2\, ds + \int_a^{a+b}(2a-s)^2\, ds + \int_{a+b}^{a+b+c}(s-2b+2a)^2\, ds \right. \\ &\qquad \left. + \int_{a+b+c}^T (2a-2b+2c-s)^2\, ds\right] \\ &= \left[\frac{T^3}{6} - \frac{(b+c+d)^3}{3} + \frac{(c+d)^3}{3} - \frac{d^3}{3}\right] \\ &\quad + \frac{k}{2}\left[\frac{2a^3}{3} + \frac{2(a-b)^3}{3} + \frac{2(c-b+a)^3}{3} + \frac{(d-c+b-a)^3}{3}\right]. \end{aligned}$$

The coordinates of $x(T, u(\xi))$ are:

$$\begin{aligned} x_1(T, u(\xi)) &= x_1(T, \bar{u}), \\ x_2(T, u(\xi)) &= T^2/2 - (b+c+d-c\xi^{1/2})^2 + (c+d-(b+c)\xi^{1/2})^2 - (d-b\xi^{1/2})^2 \\ &= x_2(T, \bar{u}) + 2bc\xi, \\ x_3(T, u(\xi)) &= \tfrac{1}{3}[T^3/2 - (b+c+d-c\xi^{1/2})^3 + (c+d-(b+c)\xi^{1/2})^3 - (d-b\xi^{1/2})^3] \\ &\quad + \frac{k}{3}[(a+c\xi^{1/2})^3 + (b-a+(b-c)\xi^{1/2})^3 + (c-b+a-b\xi^{1/2})^3 \\ &\qquad + (d-c+b-a)^3/2] \\ &= x_3(T, \bar{u}) + [b^2c + bc^2]\xi^{1/2} + k[b^2c - bc^2]\xi^{1/2} \\ &\quad + [bc^2 + b^2c + 2bcd]\xi + k[2abc - b^2c + bc^2] + O(\xi^{3/2}) \\ &= x_3(T, \bar{u}) + bc[b+c+2d+k(2a-b+c)]\xi + O(\xi^{3/2}), \end{aligned}$$

because, by (5.17), $b + c + k(b-c) = 0$. This yields (5.19).

REFERENCES

[1] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York, 1969.

[2] H. HERMES AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Academic Press, New York, 1969.

[3] H. HERMES, *Local controllability and sufficient conditions in singular problems*, J. Diff. Eq., 20 (1976), pp. 213–232.

[4] ———, *Controlled stability*, Ann. Mat. Pura ed Appl., CXIV (1977), pp. 103–119.

[5] H. HERMES, *Lie algebras of vector fields and local approximation of attainable sets*, this Journal, 16 (1978), pp. 715–727.

[6] A. J. KRENER, *The high order maximal principle and its applications to singular extremals*, this Journal, 15 (1977), pp. 256–293.

[7] K. KURATOWSKI AND C. RYLL-NARDZEWSKI, *A general theorem on selectors*, Bull. Acad. Pol. Sc. Math. Astr. Phys., 13, 6 (1965), pp. 397–403.

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

[9] H. SUSSMANN, *A bang–bang theorem with bounds on the number of switchings*, this Journal, 17 (1979), pp. 629–651.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

005

SECOND ORDER CONTROLLABILITY AND OPTIMIZATION WITH ORDINARY CONTROLS*

J. WARGA†

Abstract. Let Q be a convex subset of a real vector space and a topological space, $\mathcal{Z}$ a topological vector space, $\mathcal{U} \subset Q$, C a convex subset of $\mathcal{Z}$ with a nonempty interior, and $(\phi_0, \phi_1, \phi_2): Q \to \mathbb{R} \times \mathbb{R}^m \times \mathcal{Z}$ a continuous function with a second order "finite" Taylor approximation at a point $\bar{q}$. This is a typical framework for many smooth optimal control problems in which Q is the set of relaxed controls and $\mathcal{U}$ a sufficiently "large" set of ordinary controls. We study certain new second order conditions that are necessary for $\bar{q}$ to minimize ϕ_0 on the set $\mathcal{A} \cap \mathcal{U}$, where

$$\mathcal{A} \triangleq \{q \in Q | \phi_1(q) = 0, \phi_2(q) \in C\}.$$

We also prove that if similar conditions are not satisfied by the function (ϕ_1, ϕ_2) at $\bar{q}$ then there exist neighborhoods of the origins G_1 in $\mathbb{R}^m$ and G_2 in $\mathcal{Z}$ such that

$$\phi_1(\bar{q}) + G_1 \subset \{\phi_1(u) | u \in \mathcal{U}, \phi_2(u) + G_2 \subset C\}.$$

Finally, we show that if $\bar{q}$ minimizes ϕ_0 on $\mathcal{A} \cap \mathcal{U}$ but not on $\mathcal{A}$ then every $q_0 \in \mathcal{A}$ such that $\phi_0(q_0) < \phi_0(\bar{q})$ is an "abnormal second order extremal".

Key words. local controllability, inclusion restrictions, equality restrictions, optimization, optimal control, second order necessary conditions, relaxed controls, ordinary controls, abnormal extremal

1. Introduction. A large class of optimal control problems (defined by differential or functional-integral equations) can be described by the following optimization model: let Q be a convex subset of a real vector space, $\mathcal{U} \subset Q$, $\mathcal{Z}$ a topological vector space, $C \subset \mathcal{Z}$, and $(\phi_0, \phi_1, \phi_2): Q \to \mathbb{R} \times \mathbb{R}^m \times \mathcal{Z}$ a given function. A minimizing $\mathcal{U}$-solution is an element $\bar{u}$ that minimizes ϕ_0 on the set $\mathcal{A} \cap \mathcal{U}$, where

$$\mathcal{A} \triangleq \{q \in Q | \phi_1(q) = 0, \phi_2(q) \in C\}.$$

The function ϕ_1 is *strongly locally* $(\mathcal{U}, \phi_2, C)$*-controllable at* $\bar{q}$ if there exist neighborhoods of the origins G_1 in $\mathbb{R}^m$ and G_2 in $\mathcal{Z}$ such that

$$\phi_1(\bar{q}) + G_1 \subset \{\phi_1(u) | u \in \mathcal{U}, \phi_2(u) + G_2 \subset C\}.$$

In typical optimal control problems, Q is the set of relaxed controls, $\mathcal{U}$ is either the entire set Q or its subset consisting of ordinary (point-valued) controls, ϕ_0 is the objective function, $\phi_1(u) = 0$ describes constraints such as the endpoint restrictions, and $\phi_2(u) \in C$ describes the unilateral (state variable) restrictions.

In a previous paper [6] we have shown that, under rather general conditions that may be valid even without the customary differentiability assumptions, either (ϕ_0, ϕ_1) is strongly locally $(\mathcal{U}, \phi_2, C)$-controllable at $\bar{u}$ or $\bar{u}$ satisfies certain first order conditions that generalize Pontryagin's maximum principle. These conditions are thus necessary for $\bar{u}$ to be a minimizing $\mathcal{U}$-solution. We have also studied second order conditions for $\bar{u}$ to be a minimizing $\mathcal{U}$-solution in two special cases: when the restriction $\phi_2(u) \in C$ is absent and the problem is strongly normal [4]; and with the assumption that $\mathcal{U} = Q$ [7]. In the present paper we shall extend the results of [4] and [7], as well as certain results of Bernstein [1], by dropping the above mentioned special assumptions and deriving second order necessary conditions that are valid if ϕ_1 is not strongly locally

* Received by the editors September 6, 1983, and in revised form March 20, 1984. This research was supported in part by the National Science Foundation under grant MCS 8102079.
† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

$(\mathcal{U}, \phi_2, C)$-controllable at $\bar{u}$. We shall then illustrate by an example from optimal control that our results can prove strong local controllability at some (first order) extremal point in some cases when other methods apparently fail.

2. Assumptions and results. Let Q be a convex subset of a real vector space, $\bar{q} \in Q$, $\mathcal{Z}$ a topological vector space whose topological dual we denote by $\mathcal{Z}^*$, C a convex subset of $\mathcal{Z}$ with a nonempty interior, and $\Phi \triangleq (\phi_0, \phi_1, \phi_2): Q \to \mathbb{R} \times \mathbb{R}^m \times \mathcal{Z}$. Let

$$\mathcal{T}_k \triangleq \left\{ (\theta_1, \cdots, \theta_k) \in \mathbb{R}^k \mid \theta_j \geqq 0, \sum_{j=1}^{k} \theta_j \leqq 1 \right\},$$

and assume that, for every choice of a positive integer k and of $q_1, \cdots, q_k \in Q$, the function

$$\theta \to \psi(\theta) \triangleq \Phi\left(\bar{q} + \sum_{j=1}^{k} \theta_j (q_j - \bar{q}) \right) : \mathcal{T}_k \to \mathbb{R} \times \mathbb{R}^m \times \mathcal{Z}$$

admits a *second order Taylor approximation* at 0, i.e., there exist (a restriction of) a linear operator $\psi'(0): \mathcal{T}_k \to \mathbb{R} \times \mathbb{R}^m \times \mathcal{Z}$ and (a restriction of) a symmetric bilinear operator $\psi''(0): \mathcal{T}_k^2 \to \mathbb{R} \times \mathbb{R}^m \times \mathcal{Z}$ such that

$$\lim |\theta|^{-2} (\psi(\theta) - [\psi(0) + \psi'(0)\theta + \tfrac{1}{2}\psi''(0)\theta\theta]) = 0 \quad \text{as } \theta \to 0, \ \theta \in \mathcal{T}_k.$$

This is the case, in particular, if $\mathcal{Z}$ is normed and ψ admits a second order derivative $\psi''(0)$ (relative to $\mathcal{T}_k$) at 0 (see [1, Thm. A.1]).

We shall describe these properties of Φ by borrowing a related expression used by Neustadt [2, I.7.4, p. 45] and saying that Φ has a *second order finite Taylor approximation at* $\bar{q}$. These properties imply, in particular, that there exist (a restriction of) a linear operator $\Phi'(\bar{q}): Q - \bar{q} \to \mathbb{R} \times \mathbb{R}^m \times \mathcal{Z}$ and (a restriction of) a symmetric bilinear operator $\Phi''(\bar{q}): Q - \bar{q} \to \mathbb{R} \times \mathbb{R}^m \times \mathcal{Z}$ such that

$$\Phi'(\bar{q})(q_1 - \bar{q}) = \psi'(0)(1, 0, \cdots, 0),$$

$$\Phi''(\bar{q})(q_1 - \bar{q})(q_2 - \bar{q}) = \psi''(0)(1, 0, \cdots, 0)(0, 1, 0, \cdots, 0),$$

$$\Phi''(\bar{q})(q_1 - \bar{q})^2 = \psi''(0)(1, 0, \cdots, 0)(1, 0, \cdots, 0).$$

We refer to $\Phi'(\bar{q})$ and $\Phi''(\bar{q})$ as *finite derivatives.*

In some of our theorems we shall require certain additional assumptions. We shall say that $(f, \mathcal{U}, \bar{q})$ satisfies Condition 2.1 if the following assumptions are satisfied.

Condition 2.1. Q has a uniform structure defined on it, f is a continuous function from Q to some topological space, $\mathcal{U} \subset Q$ and, for every choice of a positive integer k, of $q_1, \cdots, q_k \in Q$ and of $\theta \in \mathcal{T}_k$, there exists a sequence $(u_n(\theta))$ in $\mathcal{U}$ such that

$$\lim_n u_n(\theta) = \bar{q} + \sum_{j=1}^{k} \theta_j (q_j - \bar{q}) \text{ uniformly for } \theta \in \mathcal{T}_k,$$

$$\theta \to u_n(\theta): \mathcal{T}_k \to \mathcal{U} \text{ is continuous for each } n = 1, 2, \cdots.$$

We recall that, in particular, $\mathcal{U}$ has the properties described in Condition 2.1 if Q is a set of relaxed controls and $\mathcal{U}$ an "abundant" subset such as the set of all ordinary controls [3, IV.3, pp. 279 ff].

We write

$$\phi \triangleq (\phi_1, \phi_2), \qquad \Phi \triangleq (\phi_0, \phi_1, \phi_2),$$

$$\mathcal{A} \triangleq \{ q \in Q \mid \phi_1(q) = 0, \ \phi_2(q) \in C \},$$

and denote the interior, the closure and the boundary of a set A by A^0, $\bar{A}$ and ∂A. We identify $\mathbb{R}^k$ with its dual and thus write $l_1 x$ for $l_1 \cdot x$ if $l_1, x \in \mathbb{R}^k$. We write $|x|$ for $\sum_{j=1}^k |x_j|$ if $x = (x_1, \cdots, x_k) \in \mathbb{R}^k$ and define the norm $|B|$ of a $k \times k$ matrix B with columns $b^1, \cdots, b^k$ accordingly, that is,

$$|B| = \max_{i \leqq j \leqq k} |b^j|.$$

THEOREM 2.2. *Let $\bar{q} \in Q$, $\phi_2(\bar{q}) \in C$, and let $Y \subset Q - \bar{q}$ be such that*

(a1) *$y \in Y$ implies*

$$\phi_1'(\bar{q})y = 0, \qquad \phi_2'(\bar{q})y \in C - \phi_2(\bar{q})$$

and

(a2) *$y_1, y_2 \in Y$, $y_1 \neq y_2$ implies*

$$\phi_1''(\bar{q})y_1y_2 = 0, \qquad \phi_2''(\bar{q})y_1y_2 \in C - \phi_2(\bar{q}).$$

Then either

(b1) *there exists $l = (l_1, l_2) \in \mathbb{R}^m \times \mathscr{Z}^*$ such that $l \neq 0$,*

(b1, 1) $l\phi'(\bar{q})h \geqq 0 \quad \forall h \in Q - \bar{q}, \qquad l_2 c \leqq 0 \quad \forall c \in C - \phi_2(\bar{q})$

and

(b1, 2) $l\phi''(\bar{q})y^2 \geqq 0 \quad \forall y \in Y,$

or

(b2) *there exist neighborhoods G_1 respectively G_2 of 0 in $\mathbb{R}^m$ respectively $\mathscr{Z}$ such that*

$$\phi_1(\bar{q}) + G_1 \subset \{\phi_1(q) | q \in Q, \phi_2(q) + G_2 \subset C\}.$$

Next assume that $(\phi, \mathscr{U}, \bar{q})$ satisfies Condition 2.1. *Then either statement* (b1) *is valid or*

(b3) *there exist neighborhoods G_1 respectively G_2 of 0 in $\mathbb{R}^m$ respectively $\mathscr{Z}$ such that*

$$\phi_1(\bar{q}) + G_1 \subset \{\phi_1(u) | u \in \mathscr{U}, \phi_2(u) + G_2 \subset C\}.$$

Remark. Conditions (b1) and (b2) (respectively (b1) and (b3)) are not exclusive.

As a corollary of Theorem 2.2, we shall derive Theorems 2.3–2.5 below. Theorem 2.3 had been derived before but Theorems 2.4 and 2.5 represent new results.

THEOREM 2.3 [7, Thm. A]. *Assume that $\bar{q}$ minimizes ϕ_0 on $\mathscr{A}$. Let $Y \subset Q - \bar{q}$ be such that*

(a1) *$y \in Y$ implies*

$$\phi_0'(\bar{q})y \leqq 0, \quad \phi_1'(\bar{q})y = 0, \quad \phi_2'(\bar{q})y \in C - \phi_2(\bar{q})$$

and

(a2) *$y_1, y_2 \in Y$, $y_1 \neq y_2$ implies*

$$\phi_0''(\bar{q})y_1y_2 \leqq 0, \quad \phi_1''(\bar{q})y_1y_2 = 0, \quad \phi_2''(\bar{q})y_1y_2 \in C - \phi_2(\bar{q}).$$

Then there exists $l = (l_0, l_1, l_2) \in [0, \infty) \times \mathbb{R}^m \times \mathscr{Z}^*$ such that $l \neq 0$,

(b1, 1) $l\Phi'(\bar{q})h \geqq 0 \quad \forall h \in Q - \bar{q}, \qquad l_2 c \leqq 0 \quad \forall c \in C - \phi_2(\bar{q})$

and

(b1, 2) $l\Phi''(\bar{q})y^2 \geqq 0 \quad \forall y \in Y.$

THEOREM 2.4. *Let* $(\Phi, \mathcal{U}, \bar{q})$ *satisfy Condition* 2.1 *and* $\bar{q}$ *minimize* ϕ_0 *on* $\mathcal{A} \cap \mathcal{U}$. *Let* Y *be as in Theorem* 2.3. *Then the conclusions of Theorem* 2.3 *remain valid.*

THEOREM 2.5. *Assume that* $\bar{q}$ *minimizes* ϕ_0 *on* $\mathcal{A} \cap \mathcal{U}$ *but not on* $\mathcal{A}$. *Let* $q_0 \in \mathcal{A}$ *be such that* $\phi_0(q_0) < \phi_0(\bar{q})$, Φ *has a second order finite derivative at* q_0, *and* $(\Phi, \mathcal{U}, q_0)$ *satisfies Condition* 2.1, *and let* $Y_0 \subset Q - q_0$ *be such that*

(a1) $y \in Y_0$ *implies*

$$\phi_0'(q_0)y \leqq 0, \quad \phi_1'(q_0)y = 0, \quad \phi_2'(q_0)y \in C - \phi_2(q_0)$$

and

(a2) $y_1, y_2 \in Y_0$, $y_1 \neq y_2$ *implies*

$$\phi_0''(q_0)y_1y_2 \leqq 0, \quad \phi_1''(q_0)y_1y_2 = 0, \quad \phi_2''(q_0)y_1y_2 \in C - \phi_2(q_0).$$

Then the conclusions of Theorem 2.3 *remain valid with* $\bar{q}$ *replaced by* q_0 *and with* $l_0 = 0$.

Remark 2.6.

(a) We observe that Theorems 2.2–2.5 remain valid if all references to $\mathcal{Z}$, C and ϕ_2 are deleted. This can be seen by setting

$$\mathcal{Z} = C = \mathbb{R}, \qquad \phi_2(q) = 0 \quad \text{for } q \in Q.$$

Then the second relation in statement (b1, 1) of Theorem 2.2 implies that $l_2 = 0$ and the inclusions in statement (b2) (respectively (b3)) read

$$\phi_1(\bar{q}) + G_1 \subset \phi_1(Q) \quad (\text{respectively } \phi_1(\bar{q}) + G_1 \subset \phi_1(\mathcal{U})).$$

Theorems 2.3 and 2.4 are similarly modified. Similarly, Theorems 2.3–2.5 remain valid if we delete references to ϕ_1. This can be seen by setting

$$\hat{Q} = Q \times \mathbb{R}, \quad \hat{\phi}_0(q, a) = \phi_0(q), \quad \phi_2(q, a) = \phi_2(q),$$
$$\hat{\phi}_1(q, a) = a \quad \text{for } \hat{q} = (q, a) \in \hat{Q},$$

and applying these theorems with Q, ϕ_i replaced by $\hat{Q}$, $\hat{\phi}_i$. Then statement (b1, 1) of Theorem 2.3 yields $l_1 a \geqq 0$, $\forall a \in \mathbb{R}$; hence $l_1 = 0$.

(b) If Q is the set of relaxed controls then statement (b1, 1) of Theorems 2.3 and 2.4 is a first order necessary condition for minimum that generalizes Pontryagin's maximum principle to problems with restricted state variables (see e.g. [3, Thms. V.2.3 and V.3.2, pp. 303ff and 310ff]). If $\bar{q}$ satisfies statement (b1, 1) then it is customary to refer to it as an extremal. In view of Theorems 2.3–2.5, we might say that $\bar{q}$ is a *first order extremal* if, for some $l \neq 0$, it satisfies statement (b1, 1) of Theorem 2.3; and say that $\bar{q}$ is a *second order extremal* if, for every choice of a set $Y \subset Q - \bar{q}$ satisfying conditions (a1) and (a2), $\bar{q}$ satisfies statements (b1, 1) and (b1, 2) for an appropriate l. If, for every such Y, there exists an associated $l = (l_0, l_1, l_2)$ with $l_0 = 0$ then we shall say that $\bar{q}$ is an *abnormal second order extremal.* Thus, Theorem 2.5 can be summarized by saying that every $q_0 \in \mathcal{A}$ with $\phi_0(q_0) < \phi_0(\bar{u})$ is an abnormal second order extremal.

3. An example. Let $\mathcal{U}$ be the set of all (Lebesgue) measurable functions $u : [0, 1] \to [-1, 1]$. For each $u \in \mathcal{U}$, let

$$\phi_1(u) = (x_1(1), x_2(1)), \qquad \phi_2(u) = x_3(1),$$

where $t \to x(t) = (x_1, x_2, x_3)(t)$ is the absolutely continuous solution of the differential equations

$$\dot{x}_1 = u(t), \quad \dot{x}_2 = x_3 u(t), \quad \dot{x}_3 = x_1^2 - u(t)^2 \quad \text{a.e. in } [0, 1],$$
$$(x_1, x_2, x_3)(0) = (0, 0, 0).$$

We shall apply Theorem 2.2 to show that ϕ_1 is strongly locally $(\mathcal{U}, \phi_2, (-\infty, 0])$-controllable at $\bar{u} = 0$ (the null function) despite the fact that $\bar{u}$ is a (first order) extremal satisfying Pontryagin's maximum principle.

In order to apply Theorem 2.2, we must embed $\mathcal{U}$ in an appropriate set Q and extend the definition of $\phi = (\phi_1, \phi_2)$ to all of Q. The most obvious way of doing this is to define Q as $\mathcal{U}$ which is a convex subset of the vector space of all measurable functions from $[0, 1]$ to $[-1, 1]$. While this approach will not yield the desired result, we shall follow it briefly to illustrate the importance of choosing an appropriate linear structure on $\mathcal{U}$.

It is easily seen that ϕ has second order finite derivatives, with

$$\phi'(0)u_1 = x_\alpha(1; 0, 0), \quad \phi''(0)u_1u_2 = x_{\alpha\beta}(1; 0, 0), \quad \phi''(0)u_1^2 = x_{\alpha\alpha}(1; 0, 0),$$

where the subscripts α and β denote partial differentiation and $x(t) = x(t; \alpha, \beta)$ is the solution of

$$\begin{aligned}
&\dot{x}_1 = \alpha u_1(t) + \beta u_2(t), \qquad \dot{x}_2 = x_3 \cdot (\alpha u_1(t) + \beta u_2(t)), \\
&\dot{x}_3 = x_1^2 - [\alpha u_1(t) + \beta u_2(t)]^2 \quad \text{a.e. in } (0, 1], \\
&x(0) = (0, 0, 0).
\end{aligned}$$

An easy calculation shows that, for all $u_1 \in \mathcal{U}$,

$$\phi'(0)u_1 = \left(\int_0^1 u_1(t)\, dt, 0, 0 \right),$$

$$\phi''(0)u_1^2 = \left(0, 0, 2 \int_0^1 \left[\left(\int_0^t u_1(t')\, dt' \right)^2 - u_1(t)^2 \right] dt \right).$$

If we choose $l_1 = (0, 1)$, $l_2 = 0$, $l = (l_1, l_2)$ then statements (b1, 1) and (b1, 2) of Theorem 2.2 are trivially satisfied for any choice of the set Y and therefore the validity of statement (b3) remains open to question.

Since the simplest approach (of setting $Q = \mathcal{U}$ i.e. using Lagrangian variations) fails, we turn to alternatives. A very general set of variations (that includes Lagrangian variations) is the set of hybrid relaxed-Lagrangian variations introduced in [5]. However, for our present example, it will suffice to consider the smaller set Q_{relaxed} of relaxed controls [3, Ch. IV] which are functions σ from $[0, 1]$ to the class of Radon probability measures on $[-1, 1]$ that are measurable in an appropriate sense and whose set is endowed with an appropriate topology. It follows from [3, IV.3.9 and VI.2.1, pp. 285 and 353] that $(\phi, \mathcal{U}, \bar{u})$ satisfies Condition 2.1. As usual, any $u \in \mathcal{U}$ is identified with the relaxed control $t \to \delta_{u(t)}$, where δ_r is the Dirac measure concentrated at r (and then the null function 0 is identified with the constant function δ_0). The corresponding definition of $\phi(\sigma) = (\phi_1, \phi_2)(\sigma)$ for $\sigma \in Q_{\text{relaxed}}$ is

$$\phi_1(\sigma) = (x_1(1), x_2(1)), \qquad \phi_2(\sigma) = x_3(1)$$

where

$$\dot{x}_1 = \int r\sigma(t)(dr), \quad \dot{x}_2 = x_3 \int r\sigma(t)(dr), \quad \dot{x}_3 = x_1^2 - \int r^2\sigma(t)(dr) \quad \text{a.e. in } [0, 1],$$

$$(x_1, x_2, x_3)(0) = (0, 0, 0).$$

Since the set $\mathcal{U}$ is now embedded in Q_{relaxed}, its induced linear structure is that of relaxed controls, that is, of measure-valued functions. Therefore, for all $y_1, y_2 \in$

$Q_{\text{relaxed}} - \delta_0$ with $y_i = \sigma_i - \delta_0$,

$$\phi'(\delta_0)y_1 = x_\alpha(1;0,0), \quad \phi''(\delta_0)y_1y_2 = x_{\alpha\beta}(1;0,0), \quad \phi''(\delta_0)y_1^2 = x_{\alpha\alpha}(1;0,0),$$

where $x(t;\alpha,\beta)$ is the solution of

$$\dot{x}_1 = \alpha\eta_1(t) + \beta\eta_2(t), \quad \dot{x}_2 = x_3[\alpha\eta_1(t) + \beta\eta_2(t)], \quad \dot{x}_3 = x_1^2 - \alpha\omega_1(t) - \beta\omega_2(t) \quad \text{a.e. in } [0,1],$$

$$x(0) = (0,0,0),$$

and where η_i, ω_i are defined by

$$\eta_i(t) \triangleq \int r\sigma_i(t)(dr), \qquad \omega_i(t) \triangleq \int r^2\sigma_i(t)(dr).$$

If we set

$$a_i(t) \triangleq \int_0^t \eta_i(\tau)\, d\tau, \qquad b_i(t) \triangleq \int_0^t \omega_i(\tau)\, d\tau,$$

then an easy computation yields

$$\phi_1'(\delta_0)y_1 = (a_1(1), 0), \qquad \phi_2'(\delta_0)y_1 = -b_1(1),$$

$$\phi_1''(\delta_0)y_1y_2 = \left(0, -\int_0^1 [b_1(t)\eta_2(t) + b_2(t)\eta_1(t)]\, dt\right),$$

$$\phi_2''(\delta_0)y_1y_2 = 2\int_0^1 a_1(t)a_2(t)\, dt,$$

$$\phi_1''(\delta_0)y_1^2 = \left(0, -2\int_0^1 b_1(t)\eta_1(t)\, dt\right), \qquad \phi_2''(\delta_0)y_1^2 = 2\int_0^1 a_1(t)^2\, dt.$$

Now let

$$Y = \{\sigma_1 - \delta_0, \sigma_2 - \delta_0\} = \{y_1, y_2\},$$

where

$$\sigma_1(t) = \begin{cases} \delta_1 & \text{for } 0 \leqq t \leqq \frac{1}{2}, \\ \delta_{-1} & \text{for } \frac{1}{2} < t \leqq 1, \end{cases} \qquad \sigma_2(t) = \begin{cases} \delta_{-1} & \text{for } 0 \leqq t \leqq \frac{1}{2}, \\ \delta_1 & \text{for } \frac{1}{2} < t \leqq 1. \end{cases}$$

We verify that Y satisfies conditions (a1) and (a2) of Theorem 2.2; indeed, we have

$$\eta_1(t) = 1 \quad \text{for } t \leqq \tfrac{1}{2}, \quad \eta_1(t) = -1 \quad \text{for } t > \tfrac{1}{2}, \quad \eta_2(t) = -\eta_1(t) \quad \forall t,$$

$$\omega_1(t) = \omega_2(t) = 1 \quad \forall t,$$

$$a_1(t) = t \quad \text{for } t \leqq \tfrac{1}{2}, \quad a_1(t) = 1 - t \quad \text{for } t > \tfrac{1}{2}, \quad a_2 = -a_1, \quad b_i(t) = t \quad \forall t, i.$$

Thus

$$\phi_1'(\delta_0)y_i = (0,0), \qquad \phi_2'(\delta_0)y_i = -1 \leqq 0 \quad \text{for } i = 1, 2,$$

$$\phi_1''(\delta_0)y_1y_2 = (0,0), \qquad \phi_2''(\delta_0)y_1y_2 = \int_0^1 a_1(t)a_2(t)\, dt = -\int_0^1 a_1(t)^2\, dt \leqq 0.$$

If statement (b1, 1) of Theorem 2.2 is satisfied and if we define η, ω, a, b in terms of σ as we defined η_i, ω_i, a_i, b_i in terms of σ_i and set $l_1 = (\lambda_1, \lambda_2)$ then

$$\lambda_1 a(1) - l_2 b(1) \geqq 0 \quad \forall \sigma \in Q_{\text{relaxed}}, \qquad l_2 c \leqq 0 \quad \forall c \leqq 0;$$

hence $l_2 \geqq 0$ and

$$-l_2 \int_0^1 \omega(t)\, dt + \lambda_1 \int_0^1 \eta(t)\, dt = \int_0^1 dt \int (-l_2 r^2 + \lambda_1 r)\sigma(t)(dr) \geqq 0.$$

If we do not have $\lambda_1 = l_2 = 0$ then $-l_2 r^2 + \lambda_1 r$ achieves a negative value at some point $\rho \in [-1, 1]$ and we may then set $\sigma(t) = \delta_\rho$ which yields a result contradicting the above relation. This shows that $\lambda_1 = l_2 = 0$; hence $\lambda_2 \neq 0$. Thus statement (b1, 2) yields

$$l\phi''(\delta_0)y_1^2 = -2\lambda_2 \int_0^1 b_1(t)\eta_1(t)\, dt = -2\lambda_2 \left[\int_0^{1/2} t\, dt - \int_{1/2}^1 t\, dt \right] = \tfrac{1}{2}\lambda_2 \geqq 0,$$

$$l\phi''(\delta_0)y_2^2 = -\tfrac{1}{2}\lambda_2 \geqq 0,$$

two inequalities that contradict $\lambda_2 \neq 0$.

We conclude that statements (b1, 1) and (b1, 2) of Theorem 2.2 cannot be simultaneously satisfied, which implies that statement (b3) is valid.

4. Proofs. We shall require a lemma which follows simply from the classical theorem about the separation of convex sets and which is similar to [3, Lemma V.2.1, p. 299].

LEMMA 4.1. *Let W be a convex subset of $\mathbb{R}^m \times \mathscr{Z}$ containing 0 and C' an open convex subset of $\mathscr{Z}$, with $0 \in \bar{C}'$. Then either there exists $l = (l_1, l_2) \in \mathbb{R}^m \times \mathscr{Z}$ such that $l \neq 0$,*

$$lw \geqq 0 \quad \forall w \in W, \qquad l_2 c \leqq 0 \quad \forall c \in C'$$

or there exist points $\xi^i = (\xi_1^i, \xi_2^i) \in W$ and numbers β_i for $i = 0, \cdots, m$ such that the set $\{(1, \xi_1^i) | i = 0, \cdots, m\}$ is linearly independent in $\mathbb{R}^{m+1}$ and

$$\beta_i > 0, \quad \sum_{i=0}^m \beta_i = 1, \quad \sum_{i=0}^m \beta_i \xi_1^i = 0, \quad \xi_2^i \in C'.$$

Proof. Let $W_1 \triangleq \{\xi_1 | (\xi_1, \xi_2) \in W)$. Then either

(1) $0 \in \partial W_1$; or

(2) $0 \in W_1^0$ and there exists some $\bar{\xi} \in W$ with $\bar{\xi}_1 = 0$, $\bar{\xi}_2 \in C'$; or

(3) $0 \in W_1^0$ and every ξ with $\xi_1 = 0$, $\xi_2 \in C'$ is outside W.

If (1) holds then the first alternative is satisfied with $l_2 = 0$. If (2) holds then there exist $\eta^i \in W$ and $\beta_i > 0$ for $i = 0, \cdots, m$ such that the η_1^i are vertices of a simplex in $\mathbb{R}^m$ containing 0 in its interior—hence the set $\{(1, \eta_1^i) | i = 0, \cdots, m\}$ is linearly independent—and

$$\sum_{i=0}^m \beta_i = 1, \qquad \sum_{i=0}^m \beta_i \eta_1^i = 0.$$

If we choose $\alpha \in (0, 1]$ small enough and set

$$\xi^i = \alpha\eta^i + (1-\alpha)\bar{\xi} \quad \text{for } i = 0, \cdots, m$$

then ξ^i will have the required properties.

Finally, we consider the case (3), and set

$$W_2 \triangleq \{w_2 | (0, w_2) \in W\}.$$

Then W_2 is a nonempty convex subset of $\mathscr{Z}$ and $W_2 \cap C' = \varnothing$. Since $0 \in W_2 \cap \overline{C'}$, it follows that there exists $\lambda_2 \in \mathscr{Z}^*$ such that $\lambda_2 \neq 0$ and

$$\lambda_2 w_2 \leqq 0 \leqq \lambda_2 c \quad \forall w_2 \in W_2, c \in C'. \tag{4}$$

We denote by 0_m the origin of $\mathbb{R}^m$, and set

$$W^{\#} \triangleq \{(w_1, \lambda_2 w_2) | (w_1, w_2) \in W\},$$
$$H \triangleq \{0_m, \alpha) | \alpha > 0\}.$$

Then, by (4), $W^{\#}$ and H are disjoint nonempty convex subsets of $\mathbb{R}^m \times \mathbb{R}$ and there exists $\mu = (\mu_1, \mu_0) \in \mathbb{R}^m \times \mathbb{R}$ such that

$$\mu \neq 0, \qquad \mu_1 w_1 + \mu_0 \lambda_2 w_2 \geqq \mu_0 \alpha \quad \forall (w_1, w_2) \in W \text{ and } \alpha > 0. \tag{5}$$

We set $l_1 = \mu_1$, $l_2 = \mu_0 \lambda_2$ and conclude that $\mu_0 \leqq 0$, $l = (l_1, l_2) \neq 0$ and, by (4) and (5),

$$lw \geqq 0, \qquad l_2 c \leqq 0 \quad \forall w \in W, c \in C'. \qquad \text{Q.E.D.}$$

Proof of Theorem 2.2. We first assume that $(\phi, \mathcal{U}, \bar{q})$ satisfies Condition 2.1, and set

$$C_1 \triangleq C - \phi_2(\bar{q}), \qquad C' \triangleq C_1^0,$$
$$W \triangleq \{\phi''(\bar{q}) \sum{}' \tau_y^2 y^2 + 2\phi'(\bar{q})h | y \in Y, \tau_y \in [0, 1], h \in Q - \bar{q}\},$$

where $\sum'$ denotes finite sums in which different terms contain distinct elements y. Clearly, W is convex, $0 \in W$ and $0 \in \bar{C}'$. Thus, by Lemma 4.1, either there exists $l = (l_1, l_2) \in \mathbb{R}^m \times \mathscr{Z}^*$ such that

$$l \neq 0, \quad lw \geqq 0 \quad \forall w \in W, \quad l_2 c \leqq 0 \quad \forall c \in \bar{C}_1,$$

or there exist $\xi^i \triangleq (\xi_1^i, \xi_2^i) \in W$ and β_i for $i = 0, \cdots, m$ such that the set $\{(1, \xi_1^i) | i = 0, \cdots, m\}$ is linearly independent in $\mathbb{R}^{m+1}$ and

$$\beta_i > 0, \quad \sum_{i=0}^{m} \beta_i = 1, \quad \sum_{i=0}^{m} \beta_i \xi_1^i = 0, \quad \xi_2^i \in C'.$$

If the first alternative holds then

$$lw = l(\sum{}' \tau_y^2 \phi''(\bar{q}) y^2 + 2\phi'(\bar{q})h) \geqq 0 \quad \forall w \in W \text{ and } h \in Q - \bar{q}.$$

Then relation (b1, 1) is obtained by setting $w = 2\phi'(\bar{q})h$ and relation (b1, 2) by setting $w = \phi''(\bar{q})y^2$.

Now assume that the second alternative holds, and let

$$\xi^i = \phi''(\bar{q}) \sum_{j=1}^{k_i} \tau_{i,j}^2 y_{i,j}^2 + 2\phi'(\bar{q})h_i \quad \text{for } i = 0, \cdots, m.$$

If we represent all the distinct $y_{i,j}$ as $y_1, \cdots, y_k$ and redefine $\tau_{i,j}$ appropriately, we can write that

$$\xi^i = \phi''(\bar{q}) \sum_{j=1}^{k} \tau_{i,j}^2 y_j^2 + 2\phi'(\bar{q})h_i,$$

where k may be assumed $\geqq 1$ (because $\tau_{i,j}$ may have value 0). We observe that for

$$\alpha \in [0, (2k)^{-1}], \quad \theta = (\theta_0, \cdots, \theta_m), \quad \theta_j \geqq 0, \quad |\theta| = 1,$$

we have

$$\sum_{i=0}^{m} \theta_i \tau_{i,j}^2 \leqq 1,$$

and therefore

$$\tilde{q} \triangleq \bar{q} + \alpha \sum_{j=1}^{k} \left(\sum_{i=0}^{m} \theta_i \tau_{i,j}^2 \right)^{1/2} y_j + \alpha^2 \sum_{i=0}^{m} \theta_i h_i \in Q$$

and

$$\phi(\tilde{q}) = \phi(\bar{q}) + \alpha\phi'(\bar{q}) \sum_{j=1}^{k} \left(\sum_{i=0}^{m} \theta_i \tau_{i,j}^2 \right)^{1/2} y_j$$
$$+ \tfrac{1}{2}\alpha^2 \left[\phi''(\bar{q}) \left(\sum_{j=1}^{k} \left(\sum_{i=0}^{m} \theta_i \tau_{i,j}^2 \right)^{1/2} y_j \right)^2 + 2\phi'(\bar{q}) \sum_{i=0}^{m} \theta_i h_i \right] + \psi(\alpha, \theta),$$

where $\psi(\alpha, \theta) = o(\alpha^2)$ uniformly for all θ with $|\theta| = 1$. (This last assertion can be verified by applying the Taylor approximation to $\phi(\tilde{q})$ as a function of (α, u, v) about $(0, 0, 0)$, where $u = (u_1, \cdots, u_k)$, $v = (v_1, \cdots, v_m)$, u_j is the coefficient of y_j in $\tilde{q}$ and $v_i = \alpha\theta_i$). Since

$$\phi_1'(\bar{q})y_j = 0, \quad \phi_2'(\bar{q})y_j \in C_1, \quad \phi_1''(\bar{q})y_p y_r = 0, \quad \phi_2''(\bar{q})y_p y_r \in C_1 \quad \text{if } p \neq r,$$

we have

$$\tilde{a}_2 \triangleq \alpha\phi_2'(\bar{q}) \sum_{j=1}^{k} \left(\sum_{i=0}^{m} \theta_i \tau_{i,j}^2 \right)^{1/2} y_j$$
$$+ \tfrac{1}{2}\alpha^2 \phi_2''(\bar{q}) \sum_{\substack{p,r=1 \\ p \neq r}}^{k} \left(\sum_{i=0}^{m} \theta_i \tau_{i,p}^2 \right)^{1/2} \left(\sum_{i=0}^{m} \theta_i \tau_{i,r}^2 \right)^{1/2} y_p y_r \in C_1 \tag{1}$$

and

$$\phi(\tilde{q}) = \phi(\bar{q}) + \tfrac{1}{2}\alpha^2 \left[\phi''(\bar{q}) \sum_{j=1}^{k} \sum_{i=0}^{m} \theta_i \tau_{i,j}^2 y_j^2 + 2\phi'(\bar{q}) \sum_{i=0}^{m} \theta_i h_i \right] + \psi(\alpha, \theta) + (0, \tilde{a}_2)$$
$$= \phi(\bar{q}) + \tfrac{1}{2}\alpha^2 \sum_{i=0}^{m} \theta_i \xi^i + \psi(\alpha, \theta) + (0, \tilde{a}_2). \tag{2}$$

We recall that $\xi_2^i \in C_1^0$ and there exists therefore a neighborhood G of the origin in $\mathscr{Z}$ such that

$$\xi_2^i + G + G + G \subset C_1 \quad \text{for } i = 0, \cdots, m.$$

We may find $\alpha_0 \in (0, (2k)^{-1}]$ such that, for $\psi = (\psi_1, \psi_2)$,

$$\psi_2(\alpha, \theta) \in \tfrac{1}{2}\alpha^2 G, \qquad \tilde{a}_2 \in \tfrac{1}{2}C_1 \quad \text{if } 0 \leqq \alpha \leqq \alpha_0, |\theta| = 1;$$

hence, by (2),

$$\phi_2(\tilde{q}) + \tfrac{1}{2}\alpha^2(G + G) \subset C \quad \text{if } 0 \leqq \alpha \leqq \alpha_0 \text{ and } |\theta| = 1. \tag{3}$$

Let $\beta_{\min}$ ($\beta_{\max}$) denote the minimum (maximum) of $\{\beta_0, \cdots, \beta_m\}$, and let H denote the nonsingular $(m+1) \times (m+1)$ matrix with columns $(1, \xi_1^i)$ for $i = 0, \cdots, m$. We observe that there exists $\alpha_1 \in (0, \alpha_0]$ such that

$$|\psi_1(\alpha, \theta)| \leqq s_1 \triangleq \tfrac{1}{24}|H^{-1}|^{-1}\alpha_1^2\beta_{\min} \quad \text{if } 0 \leqq \alpha \leqq \alpha_1, |\theta| = 1. \tag{4}$$

We have $\tilde{q} \in Q$ if $\alpha \in (0, \alpha_1]$ and $|\theta| = 1$; and $\omega \in \mathscr{T}_{k+m+1}$ if $\omega = (\omega_1, \cdots, \omega_{k+m+1})$ and

$$\omega_j = \alpha \sum_{i=0}^{m} \theta_i \tau_{i,j}^2 \quad \text{for } j = 1, \cdots, k, \qquad \omega_{k+i+1} = \alpha^2\theta_i \quad \text{for } i = 0, \cdots, m.$$

Because Condition 2.1 is assumed satisfied, there exists a sequence $(u_n(\omega))$ in $\mathscr{U}$ such that: $\lim_n u_n(\omega) = \tilde{q}$ uniformly for all ω corresponding to $\alpha \in [0, \alpha_1]$ and $|\theta| = 1$; the functions

$$(\alpha, \theta) \to \omega \to u_n(\omega)$$

are continuous for each n; and there exists N such that

$$|e_1(\alpha, \theta)| \leqq s_1, \qquad e_2(\alpha, \theta) \in s_2 G \quad \text{if } 0 \leqq \alpha \leqq \alpha_1, |\theta| = 1, \tag{5}$$

where s_1 was defined in (4) and

$$e(\alpha, \theta) \triangleq (e_1, e_2)(\alpha, \theta) \triangleq \phi(u_N(\omega)) - \phi(\tilde{q}), \qquad s_2 \triangleq s_1 + \tfrac{1}{3}\alpha_1^2. \tag{6}$$

Let $\gamma \triangleq \frac{2}{3}\alpha_1^2$ and

$$X \triangleq \{x = (x_0, \cdots, x_m) \in \mathbb{R}^{m+1} \,|\, |x_i - \gamma\beta_i| \leqq \tfrac{1}{2}\gamma\beta_{\min} \text{ for } i = 0, \cdots, m\}.$$

The set X is compact and convex and

$$x \in X \text{ implies } 0 < \tfrac{1}{2}\gamma\beta_i \leqq x_i \leqq \tfrac{3}{2}\gamma\beta_i, \quad \tfrac{1}{3}\alpha_1^2 \leqq |x| \leqq \alpha_1^2. \tag{7}$$

For each $z \in \mathbb{R}^m$ with $|z| \leqq s_1$, the function

$$x \to \gamma\beta - 2H^{-1}(-s_1, \psi_1(|x|^{1/2}, |x|^{-1}x) + e_1(|x|^{1/2}, |x|^{-1}x) - z)$$

of X into $\mathbb{R}^{m+1}$ is continuous and, by (4)–(7), maps X into itself. Therefore this function has a fixed point $\hat{x} = \hat{\alpha}^2\hat{\theta}$, where $\hat{\alpha} = |\hat{x}|^{1/2}$, $\hat{\theta} = |\hat{x}|^{-1}\hat{x}$, $|\hat{\theta}| = 1$.

Let $\hat{q}$, $\hat{\omega}$ be defined the same way as $\tilde{q}$, ω but with $\hat{\alpha}$, $\hat{\theta}$ replacing α, θ, and let $\hat{u} \triangleq u_N(\hat{\omega})$. We have

$$\hat{\alpha}^2\hat{\theta} = \gamma\beta - 2H^{-1}(-s_1, \psi_1(\hat{\alpha}, \hat{\theta}) + e_1(\hat{\alpha}, \hat{\theta}) - z);$$

hence

$$\tfrac{1}{2}\hat{\alpha}^2 H\hat{\theta} = \tfrac{1}{2}\gamma H\beta - (-s_1, \psi_1(\hat{\alpha}, \hat{\theta}) + e_1(\hat{\alpha}, \hat{\theta}) - z)$$

which yields

$$\tfrac{1}{2}\hat{\alpha}^2 = \tfrac{1}{2}\gamma + s_1 = s_2 \tag{8}$$

and

$$\tfrac{1}{2}\hat{\alpha}^2 \sum_{i=0}^{m} \hat{\theta}_i\xi_1^i + \psi_1(\hat{\alpha}, \hat{\theta}) + e_1(\hat{\alpha}, \hat{\theta}) = \tfrac{1}{2}\gamma \sum_{i=0}^{m} \beta_i\xi_1^i + z = z.$$

Therefore, by (2),

$$\phi_1(\hat{u}) = \phi_1(\hat{q}) + e_1(\hat{\alpha}, \hat{\theta}) = \phi_1(\bar{q}) + \tfrac{1}{2}\hat{\alpha}^2 \sum_{i=0}^{m} \hat{\theta}_i\xi_1^i + \psi_1(\hat{\alpha}, \hat{\theta}) + e_1(\hat{\alpha}, \hat{\theta}) = \phi_1(\bar{q}) + z. \tag{9}$$

Furthermore, by (5) and (6),

$$\phi_2(\hat{u}) = \phi_2(\hat{q}) + e_2(\hat{\alpha}, \hat{\theta}) \in \phi_2(\hat{q}) + s_2 G;$$

hence, by (3) and (8),

$$\phi_2(\hat{u}) + s_2 G \subset \phi_2(\hat{q}) + s_2(G + G) \subset C. \tag{10}$$

If we denote by G_1 the ball in $\mathbb{R}^m$ about 0 with radius s_1 and set $G_2 \triangleq s_2 G$ then statement (b3) follows from relations (9) and (10).

The proof of the first part of the theorem is a simplified version of the preceding, with no reference to $u_n(\omega)$ and with $e(\alpha, \theta)$, $\hat{u}$ replaced by 0, $\hat{q}$. Thus the only elements of Q involved in discussing the second alternative are of the form $\tilde{q}$, and there is no need to use Condition 2.1. Q.E.D.

Proof of Theorem 2.3. Let

$$\hat{\phi}_1(q) = \phi_1(q), \qquad \hat{\phi}_2(q) = (\phi_2(q), \phi_0(q) - \phi_0(\bar{q})) \quad \text{for } q \in Q,$$

$$\hat{\mathscr{Z}} = \mathscr{Z} \times \mathbb{R}, \quad \hat{C} = C \times (-\infty, 0], \quad \hat{\phi} = (\hat{\phi}_1, \hat{\phi}_2).$$

We observe that if we replace ϕ, $\mathscr{Z}$, C by $\hat{\phi}$, $\hat{\mathscr{Z}}$, $\hat{C}$ then the assumptions of the first part of Theorem 2.2 remain satisfied and therefore one of the alternatives (b1) or (b2) of that theorem is valid. The second of these alternatives implies that there exist neighborhoods G_1 and $G_2 \times (-s, s)$ of the origins in $\mathbb{R}^m$ and $\hat{\mathscr{Z}}$ such that

$$0 \in G_1 \subset \{\phi_1(q) | q \in Q, \phi_2(q) + G_2 \subset C, \phi_0(q) - \phi_0(\bar{q}) + s < 0\}.$$

Thus there exists $q_0 \in Q$ such that

$$\phi_0(q_0) < \phi_0(\bar{q}) - s < \phi_0(\bar{q}), \quad \phi_1(q_0) = 0, \quad \phi_2(q_0) \in C,$$

contrary to assumption. Therefore alternative (b1) applies to $\hat{\phi}$, $\hat{\mathscr{Z}}$, $\hat{C}$ and implies the existence of $\hat{l} = (\hat{l}_1, (\hat{l}_2, \hat{l}_0)) \in \mathbb{R}^m \times \mathscr{Z} \times \mathbb{R}$ such that $\hat{l} \neq 0$,

(1) $\quad \hat{l}\hat{\phi}'(\bar{q})h = [\hat{l}_1\phi_1'(\bar{q}) + \hat{l}_2\phi_2'(\bar{q}) + \hat{l}_0\phi_0'(\bar{q})]h \geqq 0 \quad \forall h \in Q - \bar{q},$

(2) $\quad (\hat{l}_2, \hat{l}_0)(c, \alpha) = \hat{l}_2 c + \hat{l}_0 \alpha \leqq 0 \quad \forall c \in C - \phi_2(\bar{q}), \alpha \leqq 0,$

(3) $\quad \hat{l}\hat{\phi}''(\bar{q})y^2 = [\hat{l}_1\phi_1''(\bar{q}) + \hat{l}_2\phi_2''(\bar{q}) + \hat{l}_0\phi_0''(\bar{q})]y^2 \geqq 0 \quad \forall y \in Y.$

Relation (2) implies

$$\hat{l}_0 \geqq 0, \quad \hat{l}_2 c \leqq 0 \quad \forall c \in C - \phi_2(\bar{q}),$$

and the remaining assertions of Theorem 2.3 follow from (1) and (3). Q.E.D.

Proof of Theorem 2.4. The proof is the same as for Theorem 2.3 except that the reference to (b2) is replaced by reference to (b3) which implies

$$0 \in G_1 \subset \{\phi_1(u) | u \in \mathscr{U}, \phi_2(u) + G_2 \subset C, \phi_0(u) - \phi_0(\bar{q}) + s < 0\}. \qquad \text{Q.E.D.}$$

Proof of Theorem 2.5. We apply the second part of Theorem 2.2, with $\bar{q}$, $\mathscr{Z}$, C, ϕ replaced by q_0, $\hat{\mathscr{Z}}$, $\hat{C}$, $\hat{\phi}$, where

$$\hat{\mathscr{Z}} = C \times \mathbb{R}, \qquad \hat{C} = C \times (-\infty, \phi_0(\bar{q})),$$

$$\hat{\phi}_1(q) = \phi_1(q), \qquad \hat{\phi}_2(q) = (\phi_2(q), \phi_0(q)).$$

Statement (b3) implies then the existence of neighborhoods G_1 and $G_2 \times (-s, s)$ of the origins in $\mathbb{R}^m$ and $\mathscr{Z} \times \mathbb{R}$ such that

$$0 \in G_1 \subset \{\phi_1(u) | u \in \mathscr{U}, \phi_2(u) + G_2 \subset C, \phi_0(u) + s < \phi_0(\bar{q})\}.$$

Thus there exists $u_1 \in \mathscr{U}$ such that

$$\phi_0(u_1) < \phi_0(\bar{q}) - s, \quad \phi_1(u_1) = 0, \quad \phi_2(u_1) \in C,$$

contrary to assumption. Therefore alternative (b1) of Theorem 2.2 is valid and implies the existence of $\hat{l} = (\hat{l}_1, (\hat{l}_2, \hat{l}_0)) \in \mathbb{R}^m \times \mathscr{Z} \times \mathbb{R}$, $\hat{l} \neq 0$ that satisfies the relations

(1) $\quad \hat{l}\hat{\phi}'(q_0)h \geqq 0 \quad \forall h \in Q - q_0, \qquad \hat{l}\hat{\phi}''(q_0)y^2 \geqq 0 \quad \forall y \in Y_0,$

(2) $\quad (\hat{l}_2, \hat{l}_0)(c, \alpha) = \hat{l}_2 c + \hat{l}_0 \alpha \leqq 0 \quad \forall c \in C - \phi_0(q_0), \alpha \in (-\infty, \phi_0(\bar{q}) - \phi_0(q_0)).$

Relation (2) implies that

$$\hat{l}_0 = 0, \qquad \hat{l}_2 c \leqq 0 \quad \forall c \in C - \phi_2(q_0),$$

and the remaining assertions of Theorem 2.5 follow from (1). Q.E.D.

REFERENCES

[1] DENNIS S. BERNSTEIN, *A systematic approach to higher order necessary conditions in optimization theory*, this Journal, 22 (1984), pp. 211–238.

[2] L. W. NEUSTADT, *Optimization—A Theory of Necessary Conditions*, Princeton Univ. Press, Princeton, NJ, 1976.
[3] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.
[4] ———, *A second order condition that strengthens Pontryagin's maximum principle*, J. Diff. Eqs., 28 (1978), pp. 284–307.
[5] ———, *A hybrid relaxed-Lagrangian second order condition for minimum*, Differential Games and Control Theory, Vol. 3, P. T. Liu and E. Roxin, eds., Marcel Dekker, New York, 1979.
[6] ———, *Optimization and controllability without differentiability assumptions*, this Journal, 21 (1983), pp. 837–855.
[7] ———, *Second order necessary conditions in optimization*, this Journal, 22 (1984), pp. 524–528.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

006

SOME RESULTS ON BELLMAN EQUATION IN HILBERT SPACES*

G. DA PRATO†

Abstract. We give an existence result on the Bellman equation related to an infinite dimensional control problem.

Key words. Bellman equation, dynamic programming, nonlinear semigroup

1. Introduction.

This paper deals with the evolution equation

$$\begin{aligned} &\phi_t = \tfrac{1}{2}\operatorname{Tr}(S\phi_{xx}) + \langle Ax, \phi_x\rangle - F(x, \phi_x), \\ &\phi(0, x) = \phi_0(x), \end{aligned} \tag{1.1}$$

as well as with the stationary equation

$$\lambda\phi - \tfrac{1}{2}\operatorname{Tr}(S\phi_{xx}) - \langle Ax, \phi_x\rangle + F(x, \phi_x) = 0, \qquad \lambda > 0. \tag{1.2}$$

Here A is the infinitesimal generator of a strongly continuous semi-group in H, F a mapping from $H \times H$ into $\mathbb{R}$, ϕ a mapping from $[0, T]\times H$ into $\mathbb{R}$ (ϕ_t and ϕ_x denote derivatives with respect to t and x).

Equations (1.1) and (1.2) are relevant in the study of dynamic programming in the control of stochastic differential equations (see for instance [3], [7]). In [1] (1.1) is studied in the particular case

$$F(x, \phi_x) = \tfrac{1}{2}|\phi_x|^2 - g(x). \tag{1.3}$$

In this case it is possible to prove the existence and uniqueness of ϕ if ϕ_0 and g are convex (with polynomial growth to infinity). In applications to control theory, the hypothesis of convexity is fulfilled if the state equation is linear and the cost functional is convex. In this paper we give an approach to (1.1) and (1.2) without convexity hypotheses.

We remark that, using abstract Gauss measure, some results have been proved in [9] in the particular case when $A = 0$.

Our method consists first in solving the linear problem

$$\begin{aligned} &\phi_t = \tfrac{1}{2}\operatorname{Tr}(S\phi_{xx}) + \langle Ax, \phi_x\rangle, \\ &\phi(0, x) = \phi_0(x), \end{aligned} \tag{1.4}$$

then in considering the nonlinear term as a perturbation of the linear one. Section 2 is devoted to problem (2.4) and § 3 to (1.1), (1.2) (using the theory of nonlinear semigroups). Finally in § 4 we present an application of our results to a problem of stochastic control.

2. The linear problem.

We are here concerned with the problem

$$\begin{aligned} &\phi_t = \tfrac{1}{2}\operatorname{Tr}(S\phi_{xx}) + \langle Ax, \phi_x\rangle, \\ &\phi(0, x) = \phi_0(x). \end{aligned} \tag{2.1}$$

Let us list the following hypotheses:

H1) S is a self-adjoint, positive nuclear operator in a separable Hilbert space H.

* Received by the editors October 13, 1983.

† Scuola Normale Superiore, 56100, Pisa, Italy.

S is given by

$$Sx = \sum_{i=1}^{\infty} \lambda_i \langle x, e_i \rangle e_i \tag{2.2}$$

where $\{e_i\}$ is a complete orthonormal system in H and $\lambda_i > 0$, $i = 1, 2, \cdots$ ($\langle \cdot \rangle$ denotes the inner product and $|\cdot|$ the norm in H).

H2) $A: D_A \subset H \to H$ is the infinitesimal generator of a strongly continuous, linear semi-group e^{tA} in H. Moreover $|e^{tA}| \leqq 1$ and $\{e_i\} \subset D_A$.

We shall denote by $C_b(H)$ the set of all mappings $\psi: H \to \mathbb{R}$ uniformly continuous and bounded. $C_b(H)$, endowed with the norm

$$\|\psi\|_\infty = \sup_{x \in H} |\psi(x)|, \tag{2.3}$$

is a Banach space. By $C_b^h(H)$, $h = 1, 2, \cdots$, we mean the set of all mappings $\psi: H \to \mathbb{R}$ uniformly continuous and bounded, with all derivatives of order less than or equal to h.

Let $\{\beta_i\}$ be a sequence of mutually independent real Brownian motions in a probability space (Ω, ε, P). Set

$$W_t = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\, \beta_i(t) e_i; \tag{2.4}$$

then it is well known (see for instance [5]) that W_t is a H-valued Brownian motion with covariance operator S.

To solve (2.1) we consider the following approximating problem:

$$\begin{aligned} &\phi_t^n = \tfrac{1}{2} \operatorname{Tr}(S_n \phi_{xx}^n) + \langle A_n x, \phi_x^n \rangle, \\ &\phi^n(0, x) = \phi_0(x), \qquad x \in H_n \end{aligned} \tag{2.5}$$

where $H_n = P_n(H)$, $P_n x = \sum_{i=1}^{\infty} \langle x, e_i \rangle e_i$, $S_n = SP_n$, $A_n = P_n A P_n$. Note that A_n is bounded by virtue of hypothesis H2b.

The following lemma is standard (since problem (2.5) is finite dimensional).

LEMMA 2.1. *Assume that $\phi_0 \in C_b^2(H)$. Then problem* (2.5) *has a unique solution ϕ^n given by*

$$\phi^n(t, x) = E\phi_0(e^{tA_n} x + X_t^n) \quad \forall x \in H_n, \tag{2.6}$$

where

$$X_t^n = \int_0^t e^{(t-s)A_n}\, dW_s^n, \qquad W_s^n = P_n W_s \tag{2.7}$$

(*E means expectation*).

In the sequel we set

$$(T_t^n \psi)(x) = E\psi(e^{tA_n} x + X_t^n) \quad \forall x \in H_n \tag{2.8}$$

for any $\psi \in C_b(H)$. It is easy to check that T_t^n is a strongly continuous semi-group of contractions in $C_b(H_n)$ whose infinitesimal generator $\mathscr{A}_n$ is given by

$$\mathscr{A}^n \psi = \tfrac{1}{2} \operatorname{Tr}(S_n \psi_{xx}) + \langle A_n x, \psi_x \rangle \quad \forall \psi \in C_b^2(H_n). \tag{2.9}$$

Note now that X_t^n is a Gaussian random variable in H_n whose covariance Σ_t^n is given

by

$$\Sigma_t^n x = \int_0^t e^{sA_n^*} S_n e^{sA_n} x \, ds \quad \forall x \in H_n. \tag{2.10}$$

It follows:

$$(T_t^n \psi)(x) = (2\pi)^{-n/2} \det (\Sigma_t^n)^{-1/2} \int_{H_n} \exp\left(-\tfrac{1}{2}\langle (\Sigma_t^n)^{-1} y_n, y_n\rangle\right) \psi(e^{tA_n} x + y)\, dy \quad \forall \psi \in C_b(H_n). \tag{2.11}$$

Observe that, due to the hypothesis that $\lambda_i > 0$, we have $\det(\Sigma_t^n) \neq 0$.

We will compute now the derivative of $T_t^n \psi$.

LEMMA 2.2. *For any $\psi \in C_b(H_n)$, $t > 0$ and $x \in H_n$ the derivative of T_t^n with respect to x exists and is given by*

$$\frac{d}{dx}(T_t^n \psi)(x) = E(e^{tA_n^*} (\Sigma_t^n)^{-1} X_t^n \psi(e^{tA_n} x + X_t^n)). \tag{2.12}$$

Proof. Setting in (2.11) $z = e^{tA_n} x + y$, we get

$$(T_t^n \psi)(x) = (2\pi)^{-n/2} \det (\Sigma_t^n)^{-1/2} \int_{H_n} \exp\left(-\tfrac{1}{2}\langle (\Sigma_t^n)^{-1} (z - e^{tA_n} x), z - e^{tA_n} x\rangle\right) \psi(z)\, dz \tag{2.13}$$

from which

$$\begin{aligned} \left(\frac{dT_t^n \psi}{dx}\right)(x) &= (2\pi)^{-n/2} \det (\Sigma_t^n)^{-1/2} \int_{H_n} \exp\left(-\tfrac{1}{2}\langle (\Sigma_t^n)^{-1} (z - e^{tA_n} x), z - e^{tA_n} x\rangle\right) \\ &\qquad \cdot e^{tA_n^*} (\Sigma_t^n)^{-1} (z - e^{tA_n} x) \psi(z)\, dz \\ &= \int_{H_n} e^{tA_n^*} (\Sigma_t^n)^{-1} y \psi(e^{tA_n} x + y) f_n(y)\, dy \end{aligned} \tag{2.14}$$

where f_n is the n-dimensional density of X_t^n. Thus (2.12) follows. □

For any $\psi \in C_b(H)$ we now set

$$(T_t \psi)(x) = E\psi(e^{tA} x + X_t), \qquad t > 0, \quad x \in H, \tag{2.15}$$

where

$$X_t = \int_0^t e^{(t-s)A} \, dW_s. \tag{2.16}$$

LEMMA 2.3. *Let $\psi \in C_b(H)$, $\psi^n(x) = \psi(P_n x)$; then the following statements hold*:

a) $(T_t^n \psi^n)(x) \to T_t \psi(x)$ $\forall x \in H$;

b) $T_t \psi \in C_b(H)$;

c) *T_t is a semi-group of contractions in $C_b(H)$.*

Proof. We have

$$|T_t \psi(x) - T_t^n \psi^n(x)| \leqq E|\psi(e^{tA} x + X_t) - \psi(e^{tA_n} P_n x + X_t^n)|.$$

Now $e^{tA_n} P_n x \to e^{tA} x$ by the Trotter–Kato theorem; moreover $X_t^n \to X_t$ in probability

since

$$E|X_t - X_t^n|^2 = \sum_{i=n+1}^{\infty} \lambda_i \int_0^t |e^{(t-s)A} e_i|^2 \, ds$$
$$+ \sum_{i=1}^{n} \lambda_i \int_0^t |e^{(t-s)A} e_i - e^{(t-s)A_n} e_i|^2 \, ds$$
$$\leqq t \sum_{i=n+1}^{\infty} \lambda_i + \sum_{i=1}^{\infty} \lambda_i \int_0^t |e^{(t-s)A_n} e_i - e^{(t-s)A_n} e_i|^2 \, ds \to 0$$

as $n \to \infty$. (Recall that $\sum_{i=1}^{\infty} \lambda_i = \mathrm{Tr}\,(S) < +\infty$.) Conclusion a) follows from the Lebesgue theorem. The statements b) and c) are straightforward. □

We will study now the differentiability of T_t. From (2.12) it appears (for $n \to \infty$) that we have no chance to define $(d/dx)(T_t\psi)$ for every $\psi \in C_b(H)$. To this end we need some additional hypotheses and a new definition of differentiality. The situation is similar to the Gross theory for the heat equation in Hilbert spaces (when $A = 0$, see [8]).

We set

$$\Lambda_t^n = S_n \, e^{tA_n^*} (\Sigma_t^n)^{-1} \tag{2.17}$$

and assume:

H3) a) There exists the limit

$$\lim_{n \to \infty} \Lambda_t^n P_n x = \Lambda_t x \quad \forall x \in H.$$

b) There exists a constant $\gamma > 0$ such that

$$|\Lambda_t^n| \leqq \frac{\gamma}{t} \quad \forall t > 0.$$

Let us give an example in which H3 is fulfilled.

Example 2.4. Assume that

$$Ae_i = -\mu_i e_i, \qquad \mu_i \geqq 0, \quad i = 1, 2, \cdots. \tag{2.18}$$

Then

$$\Sigma_t^n e_i = \int_0^t e^{-2\mu_i t} \lambda_i \, dt \, e_i, \qquad i = 1, 2, \cdots, \tag{2.19}$$

so that

$$\Lambda_t^n e_i = \frac{2 \, e^{-t\mu_i} \mu_i}{1 - e^{-2t\mu_i}} e_i, \qquad i = 1, 2, \cdots. \tag{2.20}$$

Now the limit in H3a exists; in fact

$$|\Lambda_t^{n+p} P_{n+p} x - \Lambda_t^n P_n x|^2 = \sum_{i=n+1}^{n+p} \left| \frac{2\mu_i \, e^{-t\mu_i}}{1 - e^{-2\mu_i t}} \right|^2 |\langle x, e_i \rangle|^2$$
$$\leqq \frac{\gamma}{t} \sum_{i=n+1}^{n+p} |\langle x, e_i \rangle|^2 \tag{2.21}$$

where

$$\gamma = \sup_{\alpha > 0} \frac{\alpha \, e^{-\alpha/2}}{1 - e^{-\alpha}}. \tag{2.22}$$

H3a, b follow easily from (2.21). □

Let us now define differentiability.

DEFINITION 2.5. We assume that $\psi \in C_b(H)$ is S-differentiable if:

a) For any $x, y \in H$ there exists the limit

$$\lim_{h\to 0} \frac{1}{h}(\psi(x+hSy)-\psi(x)) = L_x(y); \tag{2.23}$$

b) $L_x(y)$ is linear, continuous in y.

If ψ is S-differentiable we denote by $S\psi_x$ the element of H defined by

$$L_x(y) = \langle S\psi_x(x), y\rangle. \tag{2.24}$$

We shall denote by $C_S^1(H)$ the set of all mappings ψ in $C_b(H)$ such that

i) ψ is S-differentiable,

ii) $S\psi_x \in C_b(H)$,

and $C_S^1(H)$, endowed with the norm

$$\|\psi\|_{C_S^1(H)} = \|\psi\|_\infty + \|S\psi_x\|_\infty, \tag{2.25}$$

is a Banach space.

We are ready now to prove the main result of this section.

PROPOSITION 2.6. *Assume that* H1, H2 *and* H3 *are fulfilled. Let* $\psi \in C_b(H)$ *and* $t>0$; *then* $T_t\psi \in C_S^1(H)$ *and*

$$S(T_t\psi)_x(x) = E(\Lambda_t X_t \psi(e^{tA}x + X_t)) = \lim_{n\to\infty} S_n(T_t^n\psi)_x(P_n x). \tag{2.26}$$

Moreover

$$\begin{aligned} &\|S_n(T_t^n\psi)_x\|_\infty \leqq \frac{\gamma}{\sqrt{t}}\sqrt{\mathrm{Tr}\,(S)}, \\ &\|S(T_t\psi)_x\|_\infty \leqq \frac{\gamma}{\sqrt{t}}\sqrt{\mathrm{Tr}\,(S)}, \end{aligned} \tag{2.27}$$

where γ *is the constant in* H3b.

Proof. For any $x, y \in H$ we set

$$F(h) = (T_t\psi)(x+hSy), \tag{2.28}$$

$$F_n(h) = (T_t^n\psi)(P_n x + hS_n y). \tag{2.29}$$

Clearly $F_n(h) \to F(h)$ uniformly in $[0, 1]$. Moreover from Lemma 2.2 we have

$$F_n'(h) = \langle E(\Lambda_t^n X_t^n \psi(e^{tA_n}(P_n x + hS_n y) + X_t^n)), y\rangle \tag{2.30}$$

so that, as $h \to 0$,

$$F_n'(h) \to \langle E(\Lambda_t X_t \psi(e^{tA}(x+hSy)+X_t)), y\rangle \quad \text{uniformly in } [0, 1]. \tag{2.31}$$

Thus $F(h)$ is differentiable in h and equality (2.26) follows. Concerning (2.27) we have

$$\begin{aligned} \|S(T_t\psi)_x\|_\infty &\leqq \frac{\gamma}{t}\|\psi\|_\infty (E(|X_t|^2)^{1/2} \\ &= \left(\sum_{i=1}^\infty \lambda_i \int_0^t |e^{(t-s)A}e_i|^2\, ds\right)^{1/2} \leqq \frac{\gamma}{\sqrt{t}}\sqrt{\mathrm{Tr}\, S}. \end{aligned} \tag{2.32}$$

□

We remark now that the semi-group T_t on $C_b(H)$ is not strongly continuous (when H is infinite-dimensional and A is unbounded). Since we cannot use the Hille–Yosida theorem, we use the following procedure to define the "infinitesimal generator" of T_t.

We set

$$(F_\lambda \psi)(x) = \int_0^\infty e^{-\lambda t}(T_t\psi)(x)\,dt = \int_0^\infty e^{-\lambda t} E\psi(e^{tA}x + X_t)\,dt \quad \forall \psi \in C_b(H), x \in H. \tag{2.33}$$

Clearly there exists a linear operator $\mathscr{A}$ in $C_b(H)$ such that

$$R(\lambda, \mathscr{A})\psi = F_\lambda \psi \quad \forall \lambda > 0; \tag{2.34}$$

moreover

$$\|R(\lambda, \mathscr{A})\|_\infty \leqq \frac{1}{\lambda} \quad \forall \lambda > 0 \tag{2.35}$$

so that $\mathscr{A}$ is m-dissipative in $C_b(H)$. $\mathscr{A}$ can be viewed as the abstract realization of the linear operator

$$\tfrac{1}{2}\operatorname{Tr}(S\psi_{xx}) + \langle Ax, \psi_x\rangle.$$

The following corollary is straightforward:

COROLLARY 2.7. *Assume that* H1, H2 *and* H3 *are fulfilled. Let* $\psi \in C_b(H)$ *and* $\lambda > 0$. *Then* $R(\lambda, \mathscr{A})\psi \in C_S^1(H)$ *and*

$$S(R(\lambda, \mathscr{A})\psi)_x(x) = \lim_{n\to\infty} S_n(R(\lambda, \mathscr{A}_n)\psi_n)_x(P_n x), \tag{2.36}$$

where the operators $\mathscr{A}_n$ *and* $\mathscr{A}$ *are defined by* (2.9) *and* (2.34) *respectively. Moreover,*

$$\|S(R(\lambda, \mathscr{A})\psi)_x\|_\infty \leqq \frac{\gamma\Gamma(1/2)\sqrt{\operatorname{Tr}(S)}}{\sqrt{\lambda}} = \frac{\gamma'}{\sqrt{\lambda}}. \tag{2.37}$$

3. The nonlinear problem. We consider here the problem.

$$\begin{aligned} &\phi_t = \tfrac{1}{2}\operatorname{Tr}(S\phi_{xx}) + \langle Ax, \phi_x\rangle - F(S\phi_x), \\ &\phi(0, x) = \phi_0(x). \end{aligned} \tag{3.1}$$

Denote by $\operatorname{Lip}(H)$ the set of all mappings $\psi: H \to \mathbb{R}$ Lipschitz continuous and set

$$\|F\|_L = \sup\left\{\frac{|f(x)-f(y)|}{|x-y|}, x, y \in H, x \neq y\right\}. \tag{3.2}$$

Let $F \in \operatorname{Lip}(H, H)$ and $\mathscr{B}$ be the mapping in $C_b(H)$ defined by

$$\mathscr{B}\phi = -F(S\phi_x) \quad \forall \phi \in C^1_{S(H)}. \tag{3.3}$$

We are going to prove that $\mathscr{A} + \mathscr{B}$ is m-dissipative, and then we shall invoke the Crandall–Ligget theorem [4] to solve (3.1).

Let us also introduce the approximating operator

$$\mathscr{B}_n\phi = -F(S_n\phi_x) \quad \forall \phi \in C_b^1(H_n). \tag{3.4}$$

LEMMA 3.1. *Assume that the hypotheses* H1, H2 *and* H3 *hold. Let* $F \in \mathrm{Lip}(H)$; *then* $\mathscr{A}_n + \mathscr{B}_n$ *is m-dissipative. Moreover, if*

$$\lambda > 4(\gamma' \|F\|_L)^2 \tag{3.5}$$

we have

$$\|\mathscr{B}_n(R(\lambda, \mathscr{A}_n))\|_L \leqq \tfrac{1}{2} \tag{3.6}$$

and

$$(\lambda - \mathscr{A}_n - \mathscr{B}_n)^{-1} g = R(\lambda, \mathscr{A}_n)(1 - \mathscr{B}_n(R(\lambda, \mathscr{A}_n)))^{-1} g \quad \forall g \in C_b(H_n). \tag{3.7}$$

Proof. The dissipativity of $\mathscr{A}_n + \mathscr{B}_n$ can be easily checked (it is a finite-dimensional operator). For m-dissipativity it suffices to show (see for instance [6]) that $\lambda - \mathscr{A}_n - \mathscr{B}_n$ is surjective for some $\lambda > 0$. To this purpose choose $g \in C_b(H_n)$ and consider the equation

$$\lambda\phi - \mathscr{A}_n\phi - \mathscr{B}_n\phi = g, \qquad \lambda > 0. \tag{3.8}$$

If we set $\psi = \lambda\phi - \mathscr{A}_n\phi$, (3.8) is equivalent to

$$\psi - \Sigma_n(\psi) = g \tag{3.9}$$

where

$$\Sigma_n\psi = -F(S_n(R(\lambda, \mathscr{A}_n)\psi_x)). \tag{3.10}$$

Recalling (2.27) we have

$$\|\Sigma_n\|_L \leqq \|F\|_L \frac{\gamma'}{\sqrt{\lambda}} \tag{3.11}$$

and the conclusion follows from the contraction principle. □

The proof of the following lemma is quite similar so it will be omitted.

LEMMA 3.2. *Under the same hypotheses of Lemma* 3.1, *if* (3.5) *holds then* $(\lambda - \mathscr{A} - \mathscr{B})^{-1}$ *exists and is given by*

$$(\lambda - \mathscr{A} - \mathscr{B})^{-1} g = R(\lambda, \mathscr{A})(1 - \mathscr{B}R(\lambda, \mathscr{A}))^{-1} g \quad \forall g \in C_b(H). \tag{3.12}$$

Note that at this stage we cannot assert that $\mathscr{A} + \mathscr{B}$ is m-dissipative (we did not prove that $\mathscr{A} + \mathscr{B}$ is dissipative). This will be proved by the following proposition.

PROPOSITION 3.3. *Assume that hypotheses* H1, H2, H3 *hold. Let* $F \in \mathrm{Lip}(H)$; *then* $\mathscr{A} + \mathscr{B}$ *is m-dissipative. Moreover, for any* $g \in C_b(H)$ *we have*

$$((\lambda - \mathscr{A} - \mathscr{B})^{-1} g)(x) = \lim_{n \to \infty} ((\lambda - \mathscr{A}_n - \mathscr{B}_n)^{-1} g_n)(x) \quad \forall x \in H, \tag{3.13}$$

$$S((\lambda - \mathscr{A} - \mathscr{B})^{-1} g)_x(x) = \lim_{n \to \infty} S_n((\lambda - \mathscr{A}_n - \mathscr{B}_n)^{-1} g_n)_x(x) \quad \forall x \in H, \tag{3.14}$$

where

$$g_n(x) = g(P_n x). \tag{3.15}$$

Proof. Set

$$\begin{aligned} \psi_n &= (1 - \mathscr{B}_n R(\lambda, \mathscr{A}_n))^{-1} g_n, \\ \psi &= (1 - \mathscr{B}R(\lambda, \mathscr{A}))^{-1} g. \end{aligned} \tag{3.16}$$

By virtue of Corollary 2.7, in order to prove (3.13) and (3.14) it suffices to prove that

$$\psi(x) = \lim_{n\to\infty} \psi_x(x) \quad \forall x \in H. \tag{3.17}$$

By the contraction principle we have

$$\begin{aligned} &\psi_n = \lim_{m\to\infty} \psi_n^m \quad \text{in } C_b(H_n), \\ &\psi = \lim_{m\to\infty} \psi^m \quad \text{in } C_b(H) \end{aligned} \tag{3.18}$$

where

$$\begin{aligned} &\psi_n^0 = g_n, \qquad \psi^0 = g, \\ &\psi_n^{m+1} = g_n + \Sigma_n(\psi_n^m), \qquad \psi^{m+1} = g + \Sigma(\psi^m). \end{aligned} \tag{3.19}$$

However, since g_n does not go to g in $C_b(H)$ (as $n \to 0$), the conclusion (3.17) does not follow immediately.

Fix now $x \in H$; then we have

$$\begin{aligned} |\psi(x) - \psi_n(P_n x)| \leqq &|\psi(x) - \psi^m(x)| + |\psi^m(x) - \psi_n^m(P_n x)| \\ &+ |\psi_n(P_n x) - \psi_n^m(P_n x)|. \end{aligned} \tag{3.20}$$

The first and the third term of the right-hand side of (3.20) go to zero (as $m \to \infty$) uniformly in n; moreover, for any fixed m we have $|\psi^m(x) - \psi_n^m(P_n x)| \to 0$ as $n \to \infty$; thus (3.17) is proved. Now dissipativity of $\mathscr{A} + \mathscr{B}$ follows from (3.13), and m-dissipativity from Lemma 3.2.

Let now $\rho \in \mathrm{Lip}\,(H, H)$ and set

$$\begin{aligned} &\mathscr{C}\phi = \langle \rho(x), S\phi_x \rangle \quad \forall \phi \in C_S^1(H), \\ &\mathscr{C}_n\phi = \langle S_n\rho(x), \phi_x \rangle \quad \forall \phi \in C_S^1(H_n). \end{aligned} \tag{3.21}$$

Then by similar arguments we can prove the following.

PROPOSITION 3.4. *Assume that hypotheses* H1, H2, H3 *hold. Let* $F \in \mathrm{Lip}\,(H)$, $\rho \in \mathrm{Lip}\,(H, H)$; *then* $\mathscr{A} + \mathscr{B} + \mathscr{C}$ *is m-dissipative. Moreover, for any* $g \in C_b(H)$ *we have*

$$((\lambda - \mathscr{A} - \mathscr{B} - \mathscr{C})^{-1} g)(x) = \lim_{n\to\infty} ((\lambda - \mathscr{A}_n - \mathscr{B}_n - \mathscr{C}_n)^{-1} g_n)(x) \quad \forall x \in H, \tag{3.22}$$

$$S((\lambda - \mathscr{A} - \mathscr{B} - \mathscr{C})^{-1} g)_x(x) = \lim_{n\to\infty} S((\lambda - \mathscr{A}_n - \mathscr{B}_n - \mathscr{C}_n)^{-1} g_n)_x(x) \quad \forall x \in H, \tag{3.23}$$

where g_n *is given by* (3.15).

Remark 3.5. Under the hypotheses of Proposition 3.4 we draw the following conclusions.

a) For any $\lambda > 0$, $g \in C_b(H)$ the equations

$$\lambda\phi - \tfrac{1}{2}\,\mathrm{Tr}\,(S\phi_{xx}) - \langle Ax, \phi_x \rangle + F(S\phi_x) - \langle \rho(x), S\phi_x \rangle = g, \tag{3.24}$$

$$\lambda\phi^n - \tfrac{1}{2}\,\mathrm{Tr}\,(S_n\phi_{xx}^n) - \langle A_n x, \phi_x^n \rangle + F(S_n\phi_x^n) - \langle S_n\rho(x), \phi_x^n \rangle = g \tag{3.25}$$

have unique solutions ϕ and ϕ^n; moreover

$$\phi^n(P_n x) \to \phi(x), \quad (S_n\phi_x^n)(P_n x) \to (S\phi_x)(x) \quad \forall x \in H. \tag{3.26}$$

b) $\mathscr{A} + \mathscr{B} + \mathscr{C}$ verifies the hypotheses of the Crandall–Liggett theorem; thus we

can conclude that the problem

$$\begin{aligned} &\phi_t = \tfrac{1}{2}\operatorname{Tr}(S\phi_{xx}) + \langle Ax, \phi_x\rangle - F(S\phi_x) + \langle \rho(x), S\phi_x\rangle = g, \\ &\phi(0, x) = \phi_0 \in C_b(H) \end{aligned} \tag{3.27}$$

has a unique weak solution. □

4. An application to control theory. We shall study the following control problem. Minimize

$$J(x, u) = E\int_0^\infty e^{-\lambda t}(g(y(s) + \tfrac{1}{2}|u(s)|^2))\, ds, \qquad \lambda > 0 \text{ fixed}, \tag{4.1}$$

over all $u \in U$ subject to the state equation

$$\begin{aligned} &dy = (Ay + S\rho(y) + Su)\, dt + dW_t, \\ &y(0) = x. \end{aligned} \tag{4.2}$$

U (the control space) is the set of all stochastic processes u adapted to W_t and such that $|u(t)| \leqq R$ where $R > 0$ is fixed. We shall assume in the whole of this section that hypotheses H1, H2, H3 hold and moreover that $\rho \in \operatorname{Lip}(H)$.

Let $J(x) = \inf_{u \in U} J(x, u)$ be the value function of problem (4.1). The corresponding Bellman equation is see for instance [3]:

$$\lambda\phi - \tfrac{1}{2}\operatorname{Tr}(S\phi_{xx}) - \langle Ax, \phi_x\rangle - \langle \rho(x), S\phi_x\rangle + F(S\phi_x) = g(x), \tag{4.3}$$

where

$$F(x) = \begin{cases} \tfrac{1}{2}|x|^2 & \text{if } |x| \leqq R, \\ R|x| - \dfrac{R^2}{2} & \text{if } |x| \geqq R. \end{cases} \tag{4.4}$$

Clearly $F \in \operatorname{Lip}(H)$, so that by Proposition (3.4) (see also Remark 3.5a), (4.3) has a unique solution $\phi \in C_S^1(H)$. Moreover, by (3.26) ϕ can be approximated by the solution ϕ^n to the equation

$$\lambda\phi^n - \tfrac{1}{2}\operatorname{Tr}(S_n\phi^n_{xx}) - \langle A_n x, \phi^n_x\rangle - \langle S_n\rho(x), \phi^n_x\rangle + F(S_n\phi^n_x) = g(x), \qquad x \in H_n. \tag{4.5}$$

Let us also consider the approximating state equations

$$\begin{aligned} &dy_n = (A_n y_n + S_n\rho(y_n) + S_n u)\, dt + dW^n_t, \\ &y_n(0) = x \in H_n. \end{aligned} \tag{4.6}$$

LEMMA 4.1. *Let $x \in H$, $u \in U$, y be the corresponding solution of* (4.2) *and ϕ the solution of* (4.3). *Then the following identity holds,*

$$\begin{aligned} &\phi(x) + \tfrac{1}{2}E\int_0^t [|u + S\phi_x|^2 - \chi(|S\phi_x| - R)]\, ds \\ &\qquad = E\int_0^t (g(y(s) + \tfrac{1}{2}|u(s)|^2)\, ds + e^{-\lambda t}(y(t)), \end{aligned} \tag{4.7}$$

where

$$\chi(\alpha) = \begin{cases} 0 & \text{if } \alpha \leqq 0, \\ \alpha^2 & \text{if } \alpha \geqq 0. \end{cases} \tag{4.8}$$

Proof. Let y_n be the solution of (4.6) and ϕ^n the solution of (4.5). By the Itô formula we have

$$d e^{-\lambda t}\phi^n(y_n) = \{F(S_n\phi_x^n) + \langle S_n u, \phi_x^n\rangle - g(y_n)\}\, dt + \langle \phi_x^n, dW_t\rangle. \tag{4.9}$$

By integrating and taking expectations, we get

$$\begin{aligned} &\phi^n(x_n) + \tfrac{1}{2}E\int_0^t [|u_n + S_n\phi_x^n|^2 - \chi(|S_n\phi_x^n| - R)]\, ds \\ &\quad = E\int_0^t (g(y_n) + \tfrac{1}{2}|u_n(s)|^2)\, ds + e^{-\lambda t}\phi^n(y_n(t)), \end{aligned} \tag{4.10}$$

where $u_n = P_n u$, and (4.9) follows by letting n go to infinity.

PROPOSITION 4.2. *The solution ϕ to* (4.3) *coincides with the value function J of problem* (4.1). *Moreover, there exists a unique optimal control u^* for problem* (4.1) *which is related to the optimal state by the synthesis formula*:

$$u^*(t) = -h(S\phi_x(y^*(t))), \qquad t \geqq 0, \tag{4.11}$$

where

$$h(z) = \begin{cases} |z| & \text{if } |z| \leqq R, \\ \dfrac{z}{|z|} R & \text{if } |z| \geqq R. \end{cases} \tag{4.12}$$

Proof. First of all we remark that the following inequality holds

$$|u + S\phi_x|^2 - \chi(|S\phi_x| - R) \geqq 0, \tag{4.13}$$

the equality being fulfilled if

$$u = -h(S\phi_x). \tag{4.14}$$

Thus, from (4.7) it follows that $\phi(x) \leqq J(x)$. To prove the converse let $\bar{y}$ be the solution of the closed loop equation

$$\begin{aligned} &d\bar{y} = (A\bar{y} + Sp(\bar{y}) - h(S\phi_x(\bar{y})))\, dt + dW_t, \\ &\bar{y}(0) = x. \end{aligned} \tag{4.15}$$

The existence and uniqueness of Eq. (4.15) are standard because $h \in \mathrm{Lip}\,(H)$ and $S\phi_x \in C_b(H)$. By setting $u = \bar{u}$, $y = \bar{y}$ in (4.7), and letting λ go to infinity, we obtain

$$\phi(x) = E\int_0^\infty (g(\bar{y}(s)) + \tfrac{1}{2}|\bar{u}(s)|^2)\, ds \tag{4.16}$$

so that $(\bar{u}, \bar{y})$ is an optimal couple for problem (4.1). Finally let $(\tilde{u}, \tilde{y})$ be another optimal couple; again by (4.7) we get

$$E\int_0^\infty [|\tilde{u} + S\phi_x(\tilde{y})|^2 - \chi(|S\phi_x(\tilde{y})| - R)]\, ds = 0 \tag{4.17}$$

which implies $\tilde{u} = -h(S\phi_x(\tilde{y}))$; due to the uniqueness of (4.15) we have $\tilde{u} = \bar{u}$. □

REFERENCES

[1] V. BARBU AND G. DA PRATO, *Solution of the Bellman equation associated with an infinite dimensional control problem and synthesis of optimal control*, this Journal, 21 (1983), pp. 531–550.

[2] ———, *Hamilton–Jacobi Equations in Hilbert Spaces*, Pitman, London, 1983.

[3] A. BENSOUSSAN, *Stochastic Control by Functional Analysis Methods*, North-Holland, Amsterdam, 1982.
[4] M. CRANDALL AND T. M. LIGGETT, *Generation of semi-groups of non-linear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.
[5] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1978.
[6] G. DA PRATO, *Applications croissants et équations d'évolution dans les espaces de Banach*, Academic Press, New York, 1976.
[7] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.
[8] L. GROSS, *Potential theory on Hilbert space*, J. Func. Anal., 1 (1967), pp. 123–181.
[9] T. HAVÂRNEAU, *Existence for the dynamic programming equations of control diffusion processes in Hilbert spaces*, Nonlinear Anal. Theory, Methods and Applications, to appear.
[10] E. PARDOUX, *Intégrales stochastiques Hilbertiennes*, Cahiers de Mathématique de la Décision, Université Paris Dauphine, 1971.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

A CONTINUOUSLY DIFFERENTIABLE EXACT PENALTY FUNCTION FOR NONLINEAR PROGRAMMING PROBLEMS WITH INEQUALITY CONSTRAINTS*

G. DI PILLO† AND L. GRIPPO‡

Abstract. In this paper it is shown that, given a nonlinear programming problem with inequality constraints, it is possible to construct a continuously differentiable exact penalty function whose global or local unconstrained minimizers correspond to global or local solutions of the constrained problem.

Key words. constrained optimization, nonlinear programming, exact penalty function methods

1. Introduction. A potential attractive idea in the field of nonlinear programming is that of constructing an exact penalty function; that is a function whose minimizing points are also solutions of the constrained problem.

Nondifferentiable exact penalty functions which possess such a property have been widely investigated in recent years (see, e.g. [1]–[5]). However, in order to enable conventional unconstrained techniques to be employed, it is necessary that the exact penalty function is sufficiently smooth, particularly in that the gradient exists and is continuous everywhere.

For problems involving only equality constraints it has been shown [6]–[8] that a continuously differentiable exact penalty function can actually be constructed, by replacing the Lagrange multiplier in the augmented Lagrangian of Hestenes and Powell with a multiplier which is a continuously differentiable function of the problem variables.

An extension of this approach to inequality constraints has been proposed in [9] by defining the multiplier function as the vector of Kuhn–Tucker multipliers associated with a quadratic programming subproblem. In this way, however, the resulting exact penalty function has still first derivative discontinuities that might affect the efficiency of the unconstrained minimization method employed.

A different approach has been considered in [10]–[11], by replacing the search for a saddle point of the ordinary Lagrangian with the search for an unconstrained minimum of a continuously differentiable exact augmented Lagrangian in the extended space of the problem variables and the associated multipliers. Extensions and computational applications of this technique have been investigated in [12]–[16].

In this paper we show that a continuously differentiable exact penalty function which depends only on the problem variables can be constructed also for the inequality constrained case. More specifically, we prove that, under mild regularity assumptions the augmented Lagrangian obtained by employing the multiplier function introduced by Glad and Polak in [17], turns out to be an exact penalty function.

The paper is organized as follows. Section 2 contains the problem formulation and the definition of the exact penalty function. In § 3 we investigate the relationships between stationary points of this function and Kuhn–Tucker pairs of the original problem. The main results are given in §§ 4 and 5, where the equivalence between the unconstrained minimization of the exact penalty function and the solution of the constrained problem is established, respectively, for global and local solutions.

* Received by the editors January 27, 1983, and in revised form September 20, 1983.

† Dipartimento di Informatica e Sistemistica, Università di Roma La Sapienza, Via Eudossiana 18, 00184 Roma, Italy.

‡ Istituto di Analisi dei Sistemi ed Informatica del C.N.R., Viale Manzoni 30, 00185 Roma, Italy.

2. Problem formulation. We consider the general nonlinear programming problem:

Problem P.

$$\text{minimize } f(x),\ x \in R^n$$
$$\text{subject to } g(x) \leqq 0,$$

where $f: R^n \to R^1$ and $g: R^n \to R^m$. We assume, unless otherwise stated, that f and g are two times continuously differentiable on R^n.

We denote by $L(x, \lambda)$ the Lagrangian function for problem P:

$$L(x, \lambda) = f(x) + \lambda' g(x)$$

and by $\nabla_x L(x, \lambda)$ the gradient of $L(x, \lambda)$ with respect to x.

A Kuhn–Tucker (K–T) pair for problem P is a pair $(\bar{x}, \bar{\lambda}) \in R^n \times R^m$ which satisfies the conditions

$$\nabla_x L(\bar{x}, \bar{\lambda}) = 0,$$
$$g(\bar{x}) \leqq 0,$$
$$\bar{\lambda} \geqq 0,$$
$$G(\bar{x})\bar{\lambda} = 0,$$

where

$$G(x) \triangleq \operatorname{diag}(g_i(x)).$$

Given any $x \in R^n$, let $I_0(x)$, $I_\pi(x)$ and $I_\nu(x)$ be the index sets defined by

$$I_0(x) \triangleq \{i: g_i(x) = 0\},$$
$$I_\pi(x) \triangleq \{i: g_i(x) \geqq 0\},$$
$$I_\nu(x) \triangleq \{i: g_i(x) < 0\}.$$

We assume that the following hypothesis is satisfied:

Assumption A. For any $x \in R^n$, the gradients $\nabla g_i(x)$, $i \in I_0(x)$ are linearly independent.

Moreover, we shall make use, when needed, of the following assumption, where X is a given subset of R^n.

Assumption B. At any point $x \in X$ where

$$\sum_{i \in I_\pi(x)} \nabla g_i(x) g_i(x) = 0,$$

we have $g_i(x) = 0$ for all $i \in I_\pi(x)$.

Note that Assumption A reduces, over the feasible set, to a well-known regularity condition. Assumption B is much weaker than the assumption considered in [17], where it is assumed that

$$\sum_{i \in I_\pi(x)} \nabla g_i(x) \lambda_i = 0$$

with $\lambda_i \geqq 0$ for all $i \in I_\pi(x)$ implies that $\lambda_i = 0$ for all $i \in I_\pi(x)$.

As regards Assumption B, we can state the following proposition:

PROPOSITION 1. *Let $\tilde{x}$ be a feasible point for problem* P. *Then there exists a neighbourhood X of $\tilde{x}$, where Assumption* B *is satisfied.*

Proof. Recalling Assumption A, we have that there must exist a neighbourhood X_1 of $\tilde{x}$ such that the gradients $\nabla g_i(x)$, $i \in I_0(\tilde{x})$ are linearly independent for all $x \in X_1$.

On the other hand, by continuity we can find a neighbourhood X of $\tilde{x}$, $X \subseteq X_1$ such that $I_\pi(x) \subseteq I_0(\tilde{x})$ for all $x \in X$. Therefore, for all $x \in X$,

$$\sum_{i \in I_\pi(x)} \nabla g_i(x) g_i(x) = 0$$

implies $g_i(x) = 0$ for all $i \in I_\pi(x)$. □

In order to define an exact penalty function for problem P, we first introduce, by means of the following proposition, the multiplier function $\lambda(x)$ proposed in [17].

PROPOSITION 2. *For any given $x \in R^n$ and $\gamma > 0$ there exists a unique minimizer $\lambda(x) \in R^m$ of the quadratic function in λ, $\varphi(\lambda; x) \triangleq \|\nabla_x L(x, \lambda)\|^2 + \gamma^2 \|G(x)\lambda\|^2$ over R^m, given by*

$$\lambda(x) = -M^{-1}(x) \frac{\partial g(x)}{\partial x} \nabla f(x) \tag{1}$$

where

$$M(x) \triangleq \frac{\partial g(x)}{\partial x} \frac{\partial g(x)'}{\partial x} + \gamma^2 G^2(x).$$

Moreover, if $(\bar{x}, \bar{\lambda})$ is a K–T *pair for problem* P, *then $\lambda(\bar{x}) = \bar{\lambda}$.*

Proof. Let

$$I_\sigma(x) \triangleq \{i, 1 \leqq i \leqq m, i \notin I_0(x)\},$$

$$A_0(x) \triangleq (\partial g_i(x)/\partial x), \qquad i \in I_0(x),$$

$$A_\sigma(x) \triangleq (\partial g_i(x)/\partial x), \qquad i \in I_\sigma(x),$$

$$G_\sigma(x) \triangleq \operatorname{diag}(g_i(x)), \qquad i \in I_\sigma(x).$$

Then, by Assumption A, we have

$$\operatorname{rank} \begin{bmatrix} \dfrac{\partial g(x)}{\partial x} & \gamma G(x) \end{bmatrix} = \operatorname{rank} \begin{bmatrix} A_0(x) & 0 & 0 \\ A_\sigma(x) & 0 & \gamma G_\sigma(x) \end{bmatrix} = m$$

so that the matrix

$$M(x) = \begin{bmatrix} \dfrac{\partial g(x)}{\partial x} & \gamma G(x) \end{bmatrix} \begin{bmatrix} \dfrac{\partial g(x)'}{\partial x} \\ \gamma G(x) \end{bmatrix}$$

is nonsingular and positive definite. This implies that the vector $\lambda(x)$ given by (1) is the unique minimizer of the quadratic function $\varphi(\lambda; x)$.

Finally, if $(\bar{x}, \bar{\lambda})$ is a K–T pair for problem P, then $\varphi(\bar{\lambda}; \bar{x}) = 0$ and hence $\lambda(\bar{x}) = \bar{\lambda}$. □

We consider now an equality constrained problem, equivalent to Problem P, which is obtained by employing squared slack variables.

Problem Q.

$$\underset{z}{\text{minimize}} f(x), \; z = (x', y')',$$

$$x \in R^n, \; y = (y_1, \cdots, y_m)' \in R^m$$

$$\text{subject to } g_i(x) + y_i^2 = 0, \; i = 1, \cdots, m.$$

The augmented Lagrangian function for Problem Q is given by:

$$L_a(z, \lambda) = f(x) + \lambda'(g(x) + Yy) + \frac{1}{\varepsilon} \|g(x) + Yy\|^2$$

where $\varepsilon > 0$ and $Y = \operatorname{diag}(y_i)$.

By substituting for λ the multiplier function $\lambda(x)$ defined by (1) and minimizing $L_a(z, \lambda(x))$ with respect to y, we obtain the function

$$U(x; \varepsilon) \triangleq \min_y L_a(z, \lambda(x)) = f(x) + \lambda(x)'(g(x) + Y(x; \varepsilon)y(x; \varepsilon)) + \frac{1}{\varepsilon} \|g(x) + Y(x; \varepsilon)y(x; \varepsilon)\|^2 \tag{2}$$

where

$$y_i^2(x; \varepsilon) \triangleq -\min\left[0, \frac{2g_i(x) + \varepsilon\lambda_i(x)}{2}\right], \qquad i = 1, \cdots, m$$

and

$$Y(x; \varepsilon) \triangleq \operatorname{diag}(y_i(x; \varepsilon)).$$

We note that function U can be rewritten, after simple calculations, in the form of the augmented Lagrangian function considered in [17].

As a consequence of (2), we have

PROPOSITION 3. *Let x be a feasible point for problem* P; *then for any $\varepsilon > 0$,*

$$U(x; \varepsilon) \leqq f(x).$$

Proof. Let x be such that $g(x) \leqq 0$. Suppose first that $y_i^2(x; \varepsilon) = 0$; this implies, by definition of y_i^2, that

$$2g_i(x) + \varepsilon\lambda_i(x) \geqq 0$$

so that, since $g_i(x) \leqq 0$, we have:

$$\frac{1}{\varepsilon} g_i^2(x) + \lambda_i(x)g_i(x) \leqq 0.$$

Now assume that $y_i^2(x; \varepsilon) > 0$; in this case

$$g_i(x) + y_i^2(x; \varepsilon) = -\frac{\varepsilon}{2}\lambda_i(x)$$

whence

$$\frac{1}{\varepsilon}(g_i(x) + y_i^2(x; \varepsilon))^2 + \lambda_i(x)(g_i(x) + y_i^2(x; \varepsilon)) = -\frac{\varepsilon}{4}\lambda_i^2(x) \leqq 0.$$

Therefore, recalling (2), we have $U(x; \varepsilon) \leqq f(x)$. □

PROPOSITION 4. *The function $U(x; \varepsilon)$ defined by* (2) *is continuously differentiable at any $x \in R^n$ and its gradient $\nabla U(x; \varepsilon)$ is given by*

$$\nabla U(x; \varepsilon) = \nabla f(x) + \frac{\partial g(x)'}{\partial x}\lambda(x) + \frac{\partial \lambda(x)'}{\partial x}(g(x) + Y(x; \varepsilon)y(x; \varepsilon)) + \frac{2}{\varepsilon}\frac{\partial g(x)'}{\partial x}(g(x) + Y(x; \varepsilon)y(x; \varepsilon)) \tag{3}$$

where

$$\frac{\partial \lambda(x)}{\partial x} = -M^{-1}(x)\left[\frac{\partial g(x)}{\partial x}\nabla_x^2 L(x, \lambda(x)) + \sum_{j=1}^{m} e_j \nabla_x L(x, \lambda(x))' \frac{\partial^2 g_j(x)}{\partial x^2} + 2\gamma^2 \Lambda(x) G(x) \frac{\partial g(x)}{\partial x}\right], \tag{4}$$

with

$$\nabla_x L(x, \lambda(x)) \triangleq [\nabla_x L(x, \lambda)]_{\lambda=\lambda(x)},$$

$$\nabla_x^2 L(x, \lambda(x)) \triangleq [\nabla_x^2 L(x, \lambda)]_{\lambda=\lambda(x)},$$

$$\Lambda(x) = \operatorname{diag}(\lambda_i(x)),$$

e_j *the jth column of the* $m \times m$ *identity matrix.*

Proof. It can be easily verified that

$$U(x; \varepsilon) = f(x) + \lambda(x)'g(x) + \frac{1}{\varepsilon}\|g(x)\|^2 - \frac{1}{\varepsilon}\sum_{i=1}^{m}\left[\min\left[0, \frac{2g_i(x) + \varepsilon\lambda_i(x)}{2}\right]\right]^2$$

which, together with the assumptions made on the problem functions, implies that U is continuously differentiable.

Noting that, by the definition of U, we have:

$$\nabla U(x; \varepsilon) = [\nabla_x L_a(z, \lambda)]_{\substack{\lambda=\lambda(x) \\ y=y(x;\varepsilon)}} + \frac{\partial y(x; \varepsilon)'}{\partial x}[\nabla_y L_a(z, \lambda)]_{\substack{\lambda=\lambda(x) \\ y=y(x;\varepsilon)}}$$

$$+ \frac{\partial \lambda(x)'}{\partial x}[\nabla_\lambda L_a(z, \lambda)]_{\substack{\lambda=\lambda(x) \\ y=y(x;\varepsilon)}}$$

and that, by definition of $y(x; \varepsilon)$,

$$[\nabla_y L_a(z, \lambda)]_{\substack{\lambda=\lambda(x) \\ y=y(x;\varepsilon)}} = 0,$$

we obtain (3).

As regards (4), we observe first that, by (1), we have

$$\frac{\partial g(x)}{\partial x}\nabla_x L(x, \lambda(x)) + \gamma^2 G^2(x)\lambda(x) = 0.$$

Making use of a dyadic expansion, we can write

$$\sum_{j=1}^{m} e_j \frac{\partial g_j(x)}{\partial x}\nabla_x L(x, \lambda(x)) + \gamma^2 G^2(x)\lambda(x) = 0,$$

whence, by differentiation, we obtain

$$\sum_{j=1}^{m} e_j\left(\nabla_x L(x, \lambda(x))' \frac{\partial^2 g_j(x)}{\partial x^2} + \frac{\partial g_j(x)}{\partial x}\left[\nabla^2 L(x, \lambda(x)) + \left[\frac{\partial^2 L(x, \lambda)}{\partial \lambda\, \partial x}\right]_{\lambda=\lambda(x)} \frac{\partial \lambda(x)}{\partial x}\right]\right)$$
$$+ \gamma^2 G^2(x)\frac{\partial \lambda(x)}{\partial x} + 2\gamma^2 \Lambda(x) G(x)\frac{\partial g(x)}{\partial x} = 0. \tag{5}$$

Since

$$\left[\frac{\partial^2 L(x,\lambda)}{\partial\lambda\,\partial x}\right]_{\lambda=\lambda(x)}=\frac{\partial g(x)'}{\partial x}$$

(4) follows directly from (5). □

In the sequel we shall investigate to which extent the function U can be considered as a (continuously differentiable) exact penalty function for the inequality constrained problem *P*.

3. Preliminary results. We establish here the relationships between stationary points of U and K–T pairs of problem P.

THEOREM 1. *Let* $(\bar{x},\bar{\lambda})$ *be a* K–T *pair of Problem* P. *Then, for any* $\varepsilon>0$:

(a) $\bar{x}$ *is a stationary point of* U,

(b) $\bar{\lambda}=\lambda(\bar{x})$,

(c) $U(\bar{x};\varepsilon)=f(\bar{x})$.

Proof. (b) has already been proved in Proposition 2. Moreover (b), the K–T conditions and the definition of $y(x;\varepsilon)$ imply

$$\nabla_x L(\bar{x},\lambda(\bar{x}))=0$$

and

$$g(\bar{x})+Y(\bar{x};\varepsilon)y(\bar{x};\varepsilon)=0.$$

Then, (a) follows from (3) and (c) follows from (2). □

The following lemmas are needed in the proof of a converse result.

LEMMA 1. *Let* $\tilde{x}$ *be a feasible point for problem* P. *Then, there exist values* $\tilde{\varepsilon}>0$ *and* $\rho>0$ *such that the matrix*

$$K(x;\varepsilon)\triangleq\frac{\partial g(x)}{\partial x}\frac{\partial g(x)'}{\partial x}-\gamma^2 G(x)Y^2(x;\varepsilon)$$

is nonsingular for all $\varepsilon\in[0,\tilde{\varepsilon}]$ *and all* x: $\|x-\tilde{x}\|\leqq\rho$.

Proof. By definition of $y_i^2(x;\varepsilon)$ we have that, if $\tilde{x}$ is feasible and $\varepsilon=0$, it results:

$$Y^2(\tilde{x};0)=-G(\tilde{x}).$$

Therefore $K(\tilde{x};0)=M(\tilde{x})$, which is nonsingular by Proposition 2. The existence of $\tilde{\varepsilon}$ and ρ follows from the continuity assumptions. □

LEMMA 2. *Let* $\bar{x}$ *be a stationary point of* U. *Then*

$$[K(\bar{x};\varepsilon)+\varepsilon B(\bar{x})](g(\bar{x})+Y(\bar{x};\varepsilon)y(\bar{x};\varepsilon))=0, \tag{6}$$

where $K(x;\varepsilon)$ *is the matrix defined in Lemma* 1 *and*

$$B(x)\triangleq\frac{1}{2}\left[\frac{\partial g(x)}{\partial x}\frac{\partial\lambda(x)'}{\partial x}-\gamma^2 G(x)\Lambda(x)\right].$$

Proof. It can be easily verified that, by the definition of $y_i^2(x;\varepsilon)$,

$$Y^2(x;\varepsilon)\lambda(x)=-\frac{2}{\varepsilon}Y^2(x;\varepsilon)(g(x)+Y(x;\varepsilon)y(x;\varepsilon)). \tag{7}$$

Moreover, by the definition of $\lambda(x)$, we have

$$\begin{aligned}\frac{\partial g(x)}{\partial x}\nabla_x L(x,\lambda(x)) &= -\gamma^2 G^2(x)\lambda(x)\\ &= -\gamma^2 G(x)(G(x)+Y^2(x;\varepsilon))\lambda(x)+\gamma^2 G(x)Y^2(x;\varepsilon)\lambda(x)\\ &= -\gamma^2 G(x)\Lambda(x)(g(x)+Y(x;\varepsilon)y(x;\varepsilon))\\ &\quad +\gamma^2 G(x)Y^2(x;\varepsilon)\lambda(x).\end{aligned} \tag{8}$$

Therefore, by (7) and (8), we get

$$\begin{aligned}\frac{\partial g(x)}{\partial x}\nabla_x L(x,\lambda(x)) = -\gamma^2\Big[G(x)\Lambda(x)+\frac{2}{\varepsilon}G(x)Y^2(x;\varepsilon)\Big]\\ \cdot(g(x)+Y(x;\varepsilon)y(x;\varepsilon)).\end{aligned} \tag{9}$$

Finally, since $\partial g(\bar{x})/\partial x\,\nabla U(\bar{x};\varepsilon)=0$, from (3) and (9) we obtain (6). □

LEMMA 3. *Let $\bar{x}$ be a stationary point of U and assume that*

$$g(\bar{x})+Y(\bar{x};\varepsilon)y(\bar{x};\varepsilon)=0.$$

Then $(\bar{x},\lambda(\bar{x}))$ is a K–T *pair of Problem* P.

Proof. By (3) and the assumptions made, we have

$$\nabla_x L(\bar{x},\lambda(\bar{x}))=0.$$

On the other hand, by the definition of $\lambda(x)$, we have

$$\frac{\partial g(x)}{\partial x}\nabla_x L(\bar{x},\lambda(\bar{x}))+\gamma G^2(\bar{x})\lambda(\bar{x})=0.$$

Hence, we have

$$G(\bar{x})\lambda(\bar{x})=0.$$

Finally, if $g_i(\bar{x})=0$ for some i, we have, by assumption, $y_i^2(\bar{x};\varepsilon)=0$ and, by the definition of $y_i(x;\varepsilon)$, $\lambda_i(\bar{x})\geqq 0$. □

At this point we can state the following result.

THEOREM 2. *Let X be a compact subset of R^n and suppose that assumption* B *holds on X. Then, there exists an $\bar{\varepsilon}>0$ such that for all $\varepsilon\in(0,\bar{\varepsilon}]$, if $\bar{x}\in X$ is a stationary point of $U(x;\varepsilon)$, the pair $(\bar{x},\lambda(\bar{x}))$ satisfies also the* K–T *conditions for problem* P.

Proof. We proceed by contradiction. Assume that the theorem is false. Then, for any integer k, there exists an $\varepsilon_k\leqq 1/k$ and a point $\bar{x}_k\in X$ such that $\bar{x}_k$ is a stationary point of $U(x;\varepsilon_k)$ but the pair $(\bar{x}_k,\lambda(\bar{x}_k))$ does not satisfy the K–T conditions for problem P.

By the compactness of X, there exists a subsequence, that we relabel $\{\bar{x}_k\}$, such that $\lim_{k\to\infty}\bar{x}_k=\tilde{x}\in X$.

Moreover, since $\nabla U(\bar{x}_k;\varepsilon_k)=0$, recalling (3) and taking into account the continuity assumptions, we have

$$\frac{\partial g(\tilde{x})'}{\partial x}(g(\tilde{x})+Y(\tilde{x};0)y(\tilde{x};0))=0,$$

that is, by definition of $y(x;\varepsilon)$;

$$\sum_{i\in I_\pi(\tilde{x})}\nabla g_i(\tilde{x})g_i(\tilde{x})=0.$$

Therefore, Assumption B implies $g_i(\tilde{x}) = 0$ for all $i \in I_\pi(\tilde{x})$, so that $\tilde{x}$ is a feasible point for problem P.

It follows from Lemma 1 that the matrix $K(\tilde{x}; 0)$ is nonsingular. Then, by the continuity assumptions, we have that there must exist an $\bar{\varepsilon} > 0$ and a value $\rho > 0$ such that the matrix $K(x; \varepsilon) + \varepsilon B(x)$ is nonsingular for all $\varepsilon \in [0, \bar{\varepsilon}]$ and x: $\|x - \tilde{x}\| \leqq \rho$.

This implies, by Lemma 2 that, for sufficiently large values of k, say $k \geqq \hat{k}$, we must have,

$$g(\bar{x}_k) + Y(\bar{x}_k; \varepsilon_k) y(\bar{x}_k; \varepsilon_k) = 0, \qquad k \geqq \hat{k}.$$

Using Lemma 3 it follows that for $k \geqq \hat{k}$, $(\bar{x}_k, \lambda(\bar{x}_k))$ satisfies the K–T conditions for problem P, and this establishes the contradiction. □

4. Global optimality results. Given a set $X \subseteq R^n$, let us denote by $\mathcal{P}(X)$ the set of global minimum points of problem P on X, that is the set of points $\bar{x}$ such that:

$$f(\bar{x}) \leqq f(x) \quad \text{for all } x \in X,\ g(x) \leqq 0,$$

and by $\bar{f}(X)$ the value of f on $\mathcal{P}(X)$.

Similarly, we denote by $\mathcal{U}(X; \varepsilon)$ the set of global minimum points of $U(x; \varepsilon)$ on X, that is the set of points $\bar{x}$ such that:

$$U(\bar{x}; \varepsilon) \leqq U(x, \varepsilon) \quad \text{for all } x \in X,$$

and by $\bar{U}(X; \varepsilon)$ the value of $U(x; \varepsilon)$ on $\mathcal{U}(X, \varepsilon)$.

In the sequel we shall investigate the relationships between $\mathcal{P}(X)$ and $\mathcal{U}(X; \varepsilon)$.

We state first the following lemma.

LEMMA 4. *Assume that* $\mathcal{P}(X) \subseteq \text{int}(X)$; *then, for any given* $\varepsilon > 0$, $\mathcal{U}(X; \varepsilon) \subseteq \mathcal{P}(X)$ *implies* $\mathcal{U}(X; \varepsilon) = \mathcal{P}(X)$.

Proof. If $\bar{x} \in \mathcal{P}(X) \subseteq \text{int}(X)$, Assumption A implies that there exists a multiplier $\bar{\lambda} \in R^m$ such that $(\bar{x}, \bar{\lambda})$ is a K–T pair of problem P. Then, by Theorem 1 we have

$$U(\bar{x}; \varepsilon) = \bar{f}(X) \quad \text{for all } \bar{x} \in \mathcal{P}(X).$$

Since $\mathcal{U}(X; \varepsilon) \subseteq \mathcal{P}(X)$ we have $\bar{U}(X; \varepsilon) = \bar{f}(X)$ so that $U(\bar{x}; \varepsilon) = \bar{U}(X; \varepsilon)$ for all $\bar{x} \in \mathcal{P}(X)$, which proves that $\mathcal{P}(X) \subseteq \mathcal{U}(X; \varepsilon)$. □

Then we can state:

THEOREM 3. *Let X be a compact subset of R^n and assume that $\mathcal{P}(X) \subseteq \text{int}(X)$. Then, there exists an $\bar{\varepsilon} > 0$ such that, for all $\varepsilon \in (0, \bar{\varepsilon}]$ $\mathcal{U}(X; \varepsilon) = \mathcal{P}(X)$.*

Proof. Recalling Lemma 4, it is sufficient to prove that there exists an $\bar{\varepsilon} > 0$ such that, for all $\varepsilon \in (0, \bar{\varepsilon}]$, $\mathcal{U}(X; \varepsilon) \subseteq \mathcal{P}(X)$. This will be shown by contradiction. Assume that the assertion is false; then for any integer k there must exist an $\varepsilon_k \leqq 1/k$ and a point $x_k \in \mathcal{U}(X; \varepsilon_k)$, $x_k \notin \mathcal{P}(X)$ such that, for all $\bar{x} \in \mathcal{P}(X)$

$$U(x_k; \varepsilon_k) \leqq U(\bar{x}; \varepsilon_k) = f(\bar{x}) = \bar{f}(X), \tag{10}$$

where the first equality follows from Theorem 1.

Since X is compact there exists a convergent subsequence, which we relabel $\{x_k\}$, such that $\lim_{k \to \infty} x_k = \hat{x} \in X$.

By (10), we have:

$$\limsup_{k \to \infty} U(x_k; \varepsilon_k) \leqq \bar{f}(X),$$

which implies, by (2):

$$g(\hat{x}) + Y(\hat{x}, 0)y(\hat{x}, 0) = 0, \quad \text{and} \quad f(\hat{x}) \leqq \bar{f}(X),$$

so that we must have $\hat{x} \in \mathcal{P}(X)$.

On the other hand, recalling Proposition 1, there exists a closed sphere $S(\hat{x}) \subseteq \text{int}(X)$, where Assumption B is satisfied and then, by Theorem 2, there exists a value $\hat{\varepsilon} > 0$ such that for all $\varepsilon_k \in (0, \hat{\varepsilon}]$, the pair $(x_k, \lambda(x_k))$, with $x_k \in S(\hat{x})$, satisfies the K–T conditions for problem P. Thus, for $\varepsilon_k \in (0, \hat{\varepsilon}]$, we have by Theorem 1:

$$U(x_k; \varepsilon_k) = f(x_k) \quad \text{and} \quad g(x_k) \leqq 0$$

and, by (10): $f(x_k) \leqq \bar{f}(X)$. This establishes the contradiction with the assumption that $x_k \notin \mathcal{P}(X)$. □

COROLLARY 1. *Let X be a compact subset of R^n and assume that the feasible set of problem* P *is contained in the interior of X, that is*:

$$\{x: g(x) \leqq 0\} \subseteq \text{int}\,(X).$$

Then there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon}]$ $\mathcal{U}(X; \varepsilon) = \mathcal{P}(X)$.

COROLLARY 2. *Let X be a compact subset of R^n and assume that $\bar{x} \in \text{int}\,(X)$ is the unique global minimum point of problem* P *on X. Then there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon}]$, $\bar{x}$ is the unique global minimum point of $U(x; \varepsilon)$ on X.*

5. Local optimality results. In this section we consider the relationships between local solutions of problem P and local minimum points of U.

A first result is a direct consequence of Theorem 3.

THEOREM 4. *Let $\bar{x}$ be a local minimum point of problem* P *and suppose that there exists a closed sphere $S(\bar{x})$ such that $\bar{x} \in \mathcal{P}(S(\bar{x})) \subseteq \text{int}\,(S(\bar{x}))$. Then there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon}]$, $\bar{x}$ is a local minimum point of $U(x; \varepsilon)$ and $\mathcal{U}(S(\bar{x}); \varepsilon) = \mathcal{P}(S(\bar{x}))$.*

In particular, we have

COROLLARY 3. *Let $\bar{x}$ be an isolated local minimum point of problem* P. *Then there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon}]$, the point $\bar{x}$ is an isolated local minimum point of $U(x; \varepsilon)$.*

A converse result is stated in the following theorem.

THEOREM 5. *Let X be a compact subset of R^n and suppose that Assumption* B *holds on X.*

Then there exists an $\bar{\varepsilon} > 0$ such that, for all $\varepsilon \in (0, \bar{\varepsilon}]$, if $\bar{x} \in X$ is a local unconstrained minimum point of $U(x; \varepsilon)$, $\bar{x}$ is a local minimum point of problem P *and $\lambda(\bar{x})$ is the associate* K–T *multiplier.*

Proof. By Theorem 2, there exists an $\bar{\varepsilon} > 0$ such that for all $\varepsilon \in (0, \bar{\varepsilon}]$, if $\bar{x}$ is a local unconstrained minimum point of $U(x; \varepsilon)$, the pair $(\bar{x}, \lambda(\bar{x}))$ satisfies the K–T conditions for problem P. Therefore, assuming $\varepsilon \in (0, \bar{\varepsilon}]$ and recalling Theorem 1, we have:

$$U(\bar{x}, \varepsilon) = f(\bar{x})$$

so that, since $\bar{x}$ is a local minimum point of $U(x; \varepsilon)$, there exists a neighbourhood Ω of $\bar{x}$ such that

$$f(\bar{x}) \leqq U(x; \varepsilon) \quad \text{for all } x \in \Omega.$$

Thus, recalling Proposition 3, we have:

$$f(\bar{x}) \leqq U(x; \varepsilon) \leqq f(x) \quad \text{for all } x \in \Omega,\ g(x) \leqq 0. \qquad \square$$

Further local optimality results, based on the consideration of second order derivatives, can be derived under the assumption that the functions f and g are three times continuously differentiable and that strict complementarity holds at K–T pairs $(\bar{x}, \bar{\lambda})$; that is: $\bar{\lambda}_i > 0$ if $g_i(\bar{x}) = 0$.

We introduce first the following notations. Let $\bar{x}$ be a feasible point for Problem P; we denote:

$$
\begin{aligned}
&g_0(x) \triangleq (g_i(x)), && i \in I_0(\bar{x}),\\
&g_\nu(x) \triangleq (g_i(x)), && i \in I_\nu(\bar{x}),\\
&\lambda_0(x) \triangleq (\lambda_i(x)), && i \in I_0(\bar{x}),\\
&\lambda_\nu(x) \triangleq (\lambda_i(x)), && i \in I_\nu(\bar{x}),\\
&y_0(x;\varepsilon) \triangleq (y_i(x;\varepsilon)), && i \in I_0(\bar{x}),\\
&y_\nu(x;\varepsilon) \triangleq (y_i(x;\varepsilon)), && i \in I_\nu(\bar{x}),\\
&Y_0(x;\varepsilon) \triangleq \operatorname{diag}(y_i(x;\varepsilon)), && i \in I_0(\bar{x}),\\
&Y_\nu(x;\varepsilon) \triangleq \operatorname{diag}(y_i(x;\varepsilon)), && i \in I_\nu(\bar{x}).
\end{aligned}
$$

We can state:

PROPOSITION 5. *Let $(\bar{x}, \bar{\lambda})$ be a* K–T *pair for problem* P *and assume that strict complementarity holds at $(\bar{x}, \bar{\lambda})$.*

Then, for any $\varepsilon > 0$, the function $U(x;\varepsilon)$ is twice continuously differentiable in a neighbourhood of $\bar{x}$, and the Hessian matrix of $U(x;\varepsilon)$ evaluated at $\bar{x}$ is given by

$$
\begin{aligned}
\nabla^2 U(\bar{x};\varepsilon) = \nabla_x^2 L(\bar{x}, \lambda(\bar{x})) &+ \frac{\partial \lambda_0(\bar{x})'}{\partial x}\frac{\partial g_0(\bar{x})}{\partial x} + \frac{\partial g_0(\bar{x})'}{\partial x}\frac{\partial \lambda_0(\bar{x})}{\partial x}\\
&+ \frac{2}{\varepsilon}\frac{\partial g_0(\bar{x})'}{\partial x}\frac{\partial g_0(\bar{x})}{\partial x} - \frac{\varepsilon}{2}\frac{\partial \lambda_\nu(\bar{x})'}{\partial x}\frac{\partial \lambda_\nu(\bar{x})}{\partial x},
\end{aligned}
\tag{11}
$$

where

$$
\nabla_x^2 L(x, \lambda(x)) \triangleq (\nabla_x^2 L(x, \lambda))_{\lambda = \lambda(x)}.
$$

Proof. Let $(\bar{x}, \bar{\lambda})$ be a K–T pair for problem P and let $I_0(\bar{x})$, $I_\nu(\bar{x})$ be the corresponding index sets. Then, under the strict complementarity assumption, it follows that there exists a neighbourhood Ω of $\bar{x}$ such that:

$$
\begin{aligned}
&y_0(x;\varepsilon) = 0,\\
&g_\nu(x) + Y_\nu(x;\varepsilon) y_\nu(x,\varepsilon) = -\frac{\varepsilon}{2}\lambda_\nu(x),
\end{aligned}
\tag{12}
$$

for all $x \in \Omega$.

As a consequence, recalling Proposition 4, we can write, for $x \in \Omega$:

$$
\begin{aligned}
\nabla U(x;\varepsilon) = \nabla f(x) &+ \frac{\partial g_0(x)'}{\partial x}\lambda_0(x) + \frac{\partial g_\nu(x)'}{\partial x}\lambda_\nu(x)\\
&+ \frac{\partial \lambda_0(x)'}{\partial x} g_0(x) + \frac{\partial \lambda_\nu(x)'}{\partial x}(g_\nu(x) + Y_\nu(x;\varepsilon) y_\nu(x;\varepsilon)) + \frac{2}{\varepsilon}\frac{\partial g_0(x)'}{\partial x} g_0(x)\\
&+ \frac{2}{\varepsilon}\frac{\partial g_\nu(x)'}{\partial x}(g_\nu(x) + Y_\nu(x;\varepsilon) y_\nu(x;\varepsilon)),
\end{aligned}
$$

from which, by (12), we get:

$$\nabla U(x;\varepsilon)=\nabla f(x)+\frac{\partial g_0(x)'}{\partial x}\lambda_0(x)+\frac{\partial \lambda_0(x)'}{\partial x}g_0(x)-\frac{\varepsilon}{2}\frac{\partial \lambda_\nu(x)'}{\partial x}\lambda_\nu(x)$$
$$+\frac{2}{\varepsilon}\frac{\partial g_0(x)'}{\partial x}g_0(x). \tag{13}$$

Then, by differentiating (13) and recalling that $g_0(\bar{x})=0$ and $\lambda_\nu(\bar{x})=0$, we obtain (11). □

The following lemma is needed in the proof of the next theorem:

LEMMA 5. *Let $P(x)$, $Q(x)$ and $R(x)$ be three quadratic forms on R^n such that*

(i) $Q(x)\geqq 0$,

(ii) $Q(x)=0$ *and* $P(x)\leqq 0$ *imply* $x=0$.

Then there exists an $\bar{\varepsilon}>0$ such that for all $\varepsilon\in(0,\bar{\varepsilon}]$:

$$P(x)+\frac{1}{\varepsilon}Q(x)+\varepsilon R(x)>0 \quad \textit{for all } x\neq 0.$$

Proof. The proof follows easily from a known result on pairs of quadratic forms [18]. □

Then we have:

THEOREM 6. *Let $(\bar{x},\bar{\lambda})$ be a* K–T *pair for problem* P *and assume that*

(i) *strict complementarity holds at $(\bar{x},\bar{\lambda})$;*

(ii) *$\bar{x}$ is an isolated local minimum point for problem* P *satisfying the second order sufficiency condition*:

$$x'\nabla_x^2 L(\bar{x},\bar{\lambda})x>0 \quad \textit{for all } x\text{:}\quad \frac{\partial g_0(\bar{x})}{\partial x}x=0,\quad x\neq 0.$$

Then there exists an $\bar{\varepsilon}>0$ such that for all $\varepsilon\in(0,\bar{\varepsilon}]$, $\bar{x}$ is an isolated local minimum point for $U(x;\varepsilon)$ and the Hessian matrix $\nabla^2 U(\bar{x};\varepsilon)$ is positive definite.

Proof. By Theorem 1, $\bar{x}$ is a stationary point of $U(x;\varepsilon)$ for any $\varepsilon>0$ and by Proposition 5, $\nabla^2 U(\bar{x};\varepsilon)$ exists and it is given by (11).

Consider now the quadratic forms:

$$P(x)\triangleq x'\nabla_x^2 L(\bar{x},\lambda(\bar{x}))x+2x'\frac{\partial \lambda_0(\bar{x})'}{\partial x}\frac{\partial g_0(\bar{x})}{\partial x}x,$$

$$Q(x)\triangleq 2\left\|\frac{\partial g_0(\bar{x})}{\partial x}x\right\|^2,\qquad R(x)\triangleq -\frac{1}{2}\left\|\frac{\partial \lambda_\nu(\bar{x})}{\partial x}x\right\|^2,$$

so that $x'\nabla^2 U(\bar{x};\varepsilon)x=P(x)+(1/\varepsilon)Q(x)+\varepsilon R(x)$.

Then it can be easily verified that, by (ii) the quadratic forms $P(x)$ and $Q(x)$ satisfy the assumptions of Lemma 5, so that $\nabla^2 U(\bar{x};\varepsilon)$ is positive definite for $\varepsilon\in(0,\bar{\varepsilon}]$, where $\bar{\varepsilon}>0$ is the number considered in Lemma 5. □

Finally, we have a converse result.

THEOREM 7. *Let X be a compact subset of R^n; suppose that Assumption* B *holds on X and that strict complementarity holds at any* K–T *pair $(\bar{x},\bar{\lambda})$ with $\bar{x}\in X$.*

Then there exists an $\bar{\varepsilon}>0$ such that, for all $\varepsilon\in(0,\bar{\varepsilon}]$, if $\bar{x}\in X$ is a local unconstrained minimum point of $U(x;\varepsilon)$ with positive definite Hessian $\nabla^2 U(\bar{x};\varepsilon)$, $\bar{x}$ is an isolated local minimum point of problem P, *satisfying the second order sufficiency conditions.*

Proof. By Theorem 2, there exists $\bar{\varepsilon}>0$ such that for all $\varepsilon\in(0,\bar{\varepsilon}]$, if $\bar{x}$ is a local unconstrained minimum point of $U(x;\varepsilon)$, the pair $(\bar{x},\lambda(\bar{x}))$ satisfies the K–T conditions

for problem P. Let $\varepsilon \in (0, \bar{\varepsilon}]$; then by Proposition 5, $\nabla^2 U(\bar{x}; \varepsilon)$ exists and it is given by (11). Therefore, by assumption

$$x' \nabla^2 U(\bar{x}; \varepsilon) x > 0 \quad \text{for } x \in R^n, \quad x \neq 0.$$

In particular, recalling (11), we have:

$$x' \nabla^2 L(\bar{x}, \lambda(\bar{x})) x > 0 \quad \text{for all } x \in R^n: \quad \frac{\partial g_0(\bar{x})}{\partial x} x = 0, \quad x \neq 0. \qquad \square$$

6. Conclusions. In this paper we have shown that a continuously differentiable exact penalty function can be constructed for inequality constrained problems. The properties of this function have been investigated from the theoretical point of view.

Computational aspects are beyond the scope of the present paper and have been considered in [19]. We observe that, from a computational point of view, the main difficulty in the minimization of the function U lies in the matrix inversion required for the evaluation of the multiplier function. For problems with a large number of constraints the exact augmented Lagrangian approach proposed in [11] could be more advantageous.

Another point which deserves attention is the choice of the penalty coefficient. In this regard we mention the possibility of employing the procedure for the automatic selection of the penalty coefficient proposed in [17].

Finally we remark that the results established in § 4 could be of interest in the context of constrained global optimization, since they enable, in principle, the replacement of the constrained problem with a problem whose global solutions can be located in the interior of a set defined by simple bounds on the variables.

REFERENCES

[1] W. I. ZANGWILL, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

[2] T. PIETRZYKOWSKI, *An exact potential method for constrained maxima*, SIAM J. Numer. Anal., 6 (1969), pp. 299–304.

[3] A. R. CONN, *Constrained optimization using a nondifferentiable penalty function*, SIAM J. Numer. Anal., 10 (1973), pp. 760–784.

[4] S. P. HAN AND O. L. MANGASARIAN, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251–269.

[5] D. Q. MAYNE AND N. MARATOS, *A first order, exact penalty function algorithm for equality constrained optimization problems*, Math. Programming, 16 (1979), pp. 303–324.

[6] R. FLETCHER, *A class of methods for nonlinear programming with termination and convergence properties*, in Integer and Nonlinear Programming, J. Abadie, ed., North-Holland, Amsterdam, 1970, pp. 157–175.

[7] H. MUKAI AND E. POLAK, *A quadratically convergent primal-dual algorithm with global convergence properties for solving optimization problems with equality constraints*, Math. Programming, 9 (1975), pp. 336–349.

[8] R. A. TAPIA, *Quasi-Newton methods for equality constrained optimization: equivalence of existing methods and a new implementation*, in Nonlinear Programming 3, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 124–164.

[9] R. FLETCHER, *An exact penalty function for nonlinear programming with inequalities*, Math. Programming, 5 (1973), pp. 129–150.

[10] G. DI PILLO AND L. GRIPPO, *A new class of augmented Lagrangians in nonlinear programming*, this Journal, 17 (1979), pp. 618–628.

[11] ———, *A new augmented Lagrangian function for inequality constraints in nonlinear programming*, J. Optim. Theory Appl., 36, n. 4 (1982), pp. 495–519.

[12] G. DI PILLO, L. GRIPPO AND F. LAMPARIELLO, *A method for solving equality constrained optimization problems by unconstrained minimization*, in Optimization Techniques, 9th IFIP Conference, K. Iracki, K. Malanowski and S. Walukiewicz, eds., Springer-Verlag, Berlin, 1980.

[13] G. DI PILLO, L. GRIPPO AND F. LAMPARIELLO, *A class of algorithms for the solution of optimization problems with inequalities*, in System Modeling and Optimization, 10th IFIP Conference, R. F. Drenick and F. Kozin, eds., Springer-Verlag, Berlin, 1982.

[14] L. C. W. DIXON, *Exact penalty function methods in nonlinear programming*, NOC, the Hatfield Polytechnic, Tech. Rep. 103, February, 1979.

[15] D. P. BERTSEKAS, *Enlarging the region of convergence of Newton's method for constrained optimization*, J. Optim. Theory Appl., 36, n. 2 (1982), pp. 221–252.

[16] ———, *Variable metric methods for constrained optimization using differentiable exact penalty functions*, Proc. Eighteenth Annual Allerton Conference on Communication, Control and Computing, Allerton Park, IL., Oct. 1980.

[17] T. GLAD AND E. POLAK, *A multiplier method with automatic limitation of penalty growth*, Math. Programming, 17 (1979), pp. 140–155.

[18] M. R. HESTENES, *Optimization Theory. The Finite Dimensional Case*, John Wiley, New York, 1975.

[19] G. DI PILLO AND L. GRIPPO, *A class of continuously differentiable exact penalty function algorithms for nonlinear programming problems*, 11th IFIP Conference on System Modeling and Optimization, Copenhagen, July, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

008

OPTIMALITY CONDITIONS FOR DISTRIBUTED CONTROL PROBLEMS WITH NONLINEAR STATE EQUATION*

D. TIBA†

Abstract. We develop an abstract approximating process for control problems with convex cost governed by evolution equations involving monotone operators. Applications are given concerning necessary conditions of optimality for distributed control problems with hyperbolic, parabolic and delay differential state systems. These conditions are expressed by means of the Clarke [11] generalized gradient.

Key words. necessary conditions, nonlinear evolution equations, distributed systems

1. Introduction.

We consider the following abstract control problem:

$$\text{(1.1)} \qquad \text{Minimize} \int_0^T L(Sy(t), u(t))\, dt$$

subject to:

$$\text{(1.2)} \qquad y'(t) + My(t) + \nu y(t) \ni Fu(t) \quad \text{a.e. } [0, T],$$

$$\text{(1.3)} \qquad y(0) = y_0.$$

Here $S: Z \to X$, $F: W \to X$ are linear continuous operators, $M: Z \to Z$ is a (multivalued) maximal monotone operator, $y_0 \in \text{dom}\,(M)$ (the domain of M). The mapping $L: X \times W \to]-\infty, +\infty]$ is convex, lower semicontinuous, proper, with finite Hamiltonian and $\nu \in R$. We assume that W, Z, X, Y are Hilbert spaces with $Y \subseteq Z$ algebraically and topologically.

By choosing in an appropriate manner the spaces and the operators, various variational inequalities or nonlinear equations of hyperbolic or parabolic type, nonlinear differential-delay equations, etc., can be written in the form (1.2), (1.3).

The question of the necessary optimality conditions in the case of control problems governed by variational inequalities was asked by F. Mignot in [15]. The elliptic state systems are considered by F. Mignot [15], F. Mignot and J. P. Puel [16], V. Barbu [5].

An extensive study of parabolic control problems, including free boundary problems, is due to V. Barbu [4], [6], [7], C. Saguez [21], Z. Meike and D. Tiba [17].

In § 3 we give an abstract approximating scheme for problem (1.1)–(1.3) which generalizes the one used by V. Barbu [4] for $M = \partial\varphi$, a subdifferential, $S = I$, the identity operator, and $Z = X = Y$. This enables us to discuss in §§ 4, 5 the optimization of systems governed by hyperbolic variational inequalities with unilateral conditions in the domain or on the boundary. Other applications are given in § 6 for certain nonlinear parabolic equations and in the last section for a nonlinear hereditary system.

Our methods are closely related to those of V. Barbu. They consist in approximating the given control problem by a family of smooth problems and afterwards in passing to the limit in the approximate optimality equations. As a major difficulty in the limiting argument and a main difference from the techniques used in the parabolic case, in the approximate adjoint equation we have to consider the limit of the product of two weakly convergent sequences. The problem is very much discussed in the literature (see for instance the recent survey of B. Dacorogna [12]). In order to identify the limit, we use a trick based on the subdifferential of a saddle function. The core of the

* Received by the editors December 4, 1981, and in revised form December 12, 1982.
† Department of Mathematics, INCREST, Bucuresti 79622, Bd.Pacii 220, Romania.

paper consists of Theorems 3.5, 4.7, 4.8, 5.8, 6.6, 7.1 (necessary conditions) and Theorems 5.1, 6.2 (existence and continuous dependence for the state equation).

2. Some remarks on notation. All spaces are real. If E is a Banach space with norm $|\cdot|_E$, then $L^p(0, T; E)$, $1 \leqq p \leqq \infty$, is the space of p-integrable, E-valued functions, $C(0, T; E)$ is the Banach space of continuous, E-valued functions and

$$W^{1,p}(0, T; E) = \{y \in L^p(0, T; E);\ y' \in L^p(0, T; E)\}$$

where y' is the derivative of y in the sense of E-valued distributions on $]0, T[$.

Let $\varphi: E \to]-\infty, +\infty]$ be a convex, lower-semicontinuous function. We denote by $\partial\varphi(x) \subset E^*$ (E^* – the dual space) the set of all subgradients of φ at x:

$$\partial\varphi(x) = \{x^* \in E^*;\ \varphi(x) \leqq \varphi(y) + (x^*, x - y),\ \forall y \in E\}.$$

When φ is Gâteaux differentiable, then $\partial\varphi(x)$ is single valued, $\partial\varphi(x) = \nabla\varphi(x)$, the Gâteaux differential. The regularization of φ is defined by:

$$\varphi_\varepsilon(x) = \inf\left\{\frac{|x - z|_E^2}{2\varepsilon} + \varphi(z);\ z \in E\right\}$$

and it is Gâteaux differentiable.

Consider G another Banach space and $K: E \times G \to [-\infty, +\infty]$ a closed, proper, saddle function (see Rockafellar [18]). The subdifferential of K is defined by:

$$\partial K(e, g) = \langle -\partial_e K(e, g), \partial_g K(e, g)\rangle,$$

$$\partial_e K(e, g) = \{e^* \in E^*;\ K(u, g) \leqq K(e, g) + (u - e, e^*),\ \forall u\},$$

$$\partial_g K(e, g) = \{g^* \in G^*;\ K(e, g) \leqq K(e, v) + (g - v, g^*),\ \forall v\}$$

and it is a maximal monotone operator.

Here $\langle \cdot, \cdot \rangle$ is an ordered pair and $(\cdot, \cdot)$ is the pairing between Banach spaces and their duals.

If L is a convex function on a product space $X \times W$, then the Hamiltonian of L is given by

$$H(x, p) = \sup\{(p, w) - L(x, w);\ w \in W\}$$

and it is a saddle function.

For a general background on convex analysis, see Rockafellar [19], and Barbu–Precupanu [8].

Let $\psi: R^m \to R$ be locally Lipschitz. We associate with ψ the generalized gradient of Clarke [11], denoted also by $\partial\psi$:

$$\partial\psi(y) = \mathrm{Conv}\{w \in R^m;\ w = \lim_{y_n \to y} \nabla\psi(y_n)\}.$$

When ψ is convex, $\partial\psi$ is just the subdifferential of ψ. For generalized gradients see also the survey of Rockafellar [20].

Finally, we denote by $H^K(\Omega)$, $H_0^K(\Omega)$, $W^{K,P}(\Omega)$, $W_0^{K,P}(\Omega)$, $H^s(\Gamma)$ usual Sobolev spaces, where Ω is a bounded domain with sufficiently smooth boundary Γ of the Euclidean space R^N.

3. The abstract control problem. Let $M^\varepsilon: Z \to Z$, $\varepsilon > 0$, be a family of maximal monotone operators. We denote $\theta_\varepsilon: L^2(0, T; Y) \to L^2(0, T; Z)$ the correspondence

$f \to y$ given by:

$$\begin{aligned} &y'(t) + M^\varepsilon y(t) + \nu y(t) = f(t) \quad \text{a.e. } [0, T], \\ &y(0) = y_0 \end{aligned} \tag{3.1}$$

and similarly we denote by θ the corresponding map when M^ε is replaced by M.

Applications θ, θ_ε are well defined according to V. Barbu [3, Thm. 2.1].

Now, we list the main hypotheses:

(a) $\theta_\varepsilon \circ F: L^2(0, T; W) \to C(0, T; Z)$ is completely continuous, uniformly in ε, that is:

$w_n \to w$ weakly in $L^2(0, T; W)$ implies $(\theta_\varepsilon \circ F)(w_n) \to (\theta_\varepsilon \circ F)(w)$ strongly in $C(0, T; Z)$, uniformly with respect to ε.

(b) $S \circ \theta_\varepsilon : L^2(0, T; Y) \to L^2(0, T; X)$ is Gâteaux differentiable for every $\varepsilon > 0$.

(c) θ_ε approximates θ uniformly:

$$|\theta_\varepsilon(f)(t) - \theta(f)(t)|_Z \leqq c \cdot \delta(\varepsilon) \quad \forall t \in [0, T]$$

where $\delta(\varepsilon) \to 0$ when $\varepsilon \to 0$, $f \in L^2(0, T; Y)$ and c is a constant depending only on $|f|_{L^2(0,T;Y)}$.

Consider the following approximate control problem:

$$\text{Minimize} \left\{ \int_0^T L^\varepsilon(Sy(t), u(t))\, dt + \frac{1}{2} \int_0^T |u(t) - u^*(t)|_W^2\, dt \right\} \tag{3.2}$$

subject to:

$$\begin{aligned} &y'(t) + M^\varepsilon y(t) + \nu y(t) = Fu(t) \quad \text{a.e. } [0, T], \\ &y(0) = y_0, \end{aligned} \tag{3.3}$$

where $L^\varepsilon = L_{\delta(\varepsilon)}$ is the regularization of the convex function L and u^*, y^* are the optimal control and the optimal state in problem (1.1)–(1.3), which are assumed to exist.

Remark 1. If L satisfies a coercivity assumption with respect to u, it is easy to infer from properties (a), (c) the existence of u^*.

This is supplied, for instance, by quadratic cost functionals or by

$$L(y, u) = \begin{cases} L_1(y, u) & \text{if } |u|_W \leqq 1, \\ +\infty & \text{otherwise} \end{cases}$$

when constraints are imposed on the control. Here L_1 is another convex, lower semicontinuous proper function, with finite Hamiltonian.

However, our main goal is to obtain necessary optimality conditions and we shall assume the existence of the optimal pair $\langle u^*, y^* \rangle$.

LEMMA 3.1. *Problem* (3.2), (3.3) *has a solution* $\langle y_\varepsilon, u_\varepsilon \rangle \in L^2(0, T; Z) \times L^2(0, T; W)$ *with* $y_\varepsilon = \theta_\varepsilon(Fu_\varepsilon)$.

Proof. The functional (3.2) is coercive in u because L^ε can be bounded from below by an affine function and $y = \theta_\varepsilon(Fu)$ admits a good evaluation in (3.3). It also is weakly lower semicontinuous since $S \circ \theta_\varepsilon$ is completely continuous by (a) and L^ε is convex, lower semicontinuous.

LEMMA 3.2. *For every* $\varepsilon > 0$ *there exists* $p_\varepsilon \in L^2(0, T; Y^*)$ *such that*:

$$p_\varepsilon = -[\nabla(S \circ \theta_\varepsilon)(Fu_\varepsilon)]^* \partial_1 L^\varepsilon(Sy_\varepsilon, u_\varepsilon), \tag{3.4}$$

$$F^*_{p_\varepsilon} = \partial_2 L^\varepsilon(Sy_\varepsilon, u_\varepsilon) + u_\varepsilon - u^*. \tag{3.5}$$

Here F^* is the adjoint of F and $\partial_1 L^\varepsilon$, $\partial_2 L^\varepsilon$ are the components of the ordered pair $\nabla L^\varepsilon(Sy_\varepsilon, u_\varepsilon) \in X \times W$.

Proof. L^ε is Fréchet differentiable and $S \circ \theta_\varepsilon$ is Gâteaux differentiable. At the minimum point $\langle y_\varepsilon, u_\varepsilon \rangle$, the Gâteaux differential vanishes:

$$\int_0^T (\partial_1 L^\varepsilon(Sy_\varepsilon, u_\varepsilon), \nabla(S \circ \theta_\varepsilon)(Fu_\varepsilon)Fv)_X \, dt + \int_0^T (\partial_2 L^\varepsilon(Sy_\varepsilon, u_\varepsilon) + u_\varepsilon - u^*, v)_W \, dt = 0$$

for every $v \in L^2(0, T; W)$. Defining p_ε as in (3.4), one obtains at once (3.5).

LEMMA 3.3. *We have*

$$y_\varepsilon \to y^* \quad \textit{strongly in } C(0, T; Z), \tag{3.6}$$

$$u_\varepsilon \to u^* \quad \textit{strongly in } L^2(0, T; W) \tag{3.7}$$

as $\varepsilon \to 0$.

Proof. Since $\langle u_\varepsilon, y_\varepsilon \rangle$ is a solution of (3.2),

$$\int_0^T L^\varepsilon(Sy_\varepsilon, u_\varepsilon) \, dt + \frac{1}{2} \int_0^T |u_\varepsilon - u^*|_W^2 \, dt \leqq \int_0^T L^\varepsilon(S \circ \theta_\varepsilon(Fu^*), u^*) \, dt. \tag{3.8}$$

From the definition of the regularization L^ε it follows that

$$L^\varepsilon(S \circ \theta_\varepsilon(Fu^*)(t), u^*(t)) \leqq L(Sy^*(t), u^*(t)) + C \cdot \frac{|\theta_\varepsilon(Fu^*)(t) - \theta(Fu^*)(t)|^2}{2\delta(\varepsilon)}.$$

From condition (c), θ_ε approximates θ uniformly and we obtain

$$\limsup_{\varepsilon \to 0} \left\{ \int_0^T L^\varepsilon(Sy_\varepsilon, u_\varepsilon) \, dt + \frac{1}{2} \int_0^T |u^* - u_\varepsilon|_W^2 \, dt \right\} \leqq \int_0^T L(Sy^*, u^*) \, dt. \tag{3.9}$$

Since the cost functional is coercive uniformly in ε, we deduce that $\{u_\varepsilon\}$ is bounded in $L^2(0, T; W)$. Therefore for a convenient subsequence we get $u_\varepsilon \to u_0$ weakly.

From assumptions (a) and (c) we deduce the inequality

$$\begin{aligned} |\theta_\varepsilon(Fu_\varepsilon) - \theta(Fu_0)|_{C(0,T;Z)} &\leqq |\theta_\varepsilon(Fu_\varepsilon) - \theta_\varepsilon(Fu_0)|_{C(0,T;Z)} \\ &\quad + |\theta_\varepsilon(Fu_0) - \theta(Fu_0)|_{C(0,T;Z)} \\ &\leqq |\theta_\varepsilon(Fu_\varepsilon) - \theta_\varepsilon(Fu_0)|_{C(0,T;Z)} + \delta(\varepsilon), \end{aligned}$$

hence $y_\varepsilon \to \tilde{y} = \theta(Fu_0)$ strongly in $C(0, T; Z)$.

Let $\varepsilon \to 0$ in (3.8). The properties of the regularization imply

$$\liminf_{\varepsilon \to 0} \int_0^T L^\varepsilon(Sy_\varepsilon, u_\varepsilon) \, dt \geqq \int_0^T L(S\tilde{y}, u_0) \, dt.$$

Therefore, since the norm is weakly lower semicontinuous we obtain from (3.9),

$$\int_0^T L(S\tilde{y}, u_0) \, dt + \frac{1}{2} \int_0^T |u^* - u_0|_W^2 \, dt \leqq \int_0^T L(Sy^*, u^*) \, dt.$$

As $\langle y^*, u^* \rangle$ is an optimal pair, it follows that $u_0 = u^*$, $\tilde{y} = \theta(Fu^*) = y^*$.

Since any sequence contains a subsequence for which the limit exists and equals $\langle u^*, y^* \rangle$, the lemma is proved.

LEMMA 3.4. *The Gâteaux differential of* $S \circ \theta_\varepsilon$ *satisfies*

$$|[\nabla(S \circ \theta_\varepsilon)(w)v](t)|_X \leqq C \int_0^t |v(s)|_Y \, ds,$$

for every $w, v \in L^2(0, T; Y)$.

Proof. By definition

$$\nabla(S \circ \theta_\varepsilon)(w)v = \lim_{\lambda \to 0} \frac{Sy_\lambda - Sy}{\lambda}$$

strongly in $L^2(0, T; X)$, where $y_\lambda = \theta_\varepsilon(w + \lambda v)$ and $y = \theta_\varepsilon(w)$. By condition (b) this limit is assumed to exist. We have:

$$\frac{1}{2}|y_\lambda(t) - y(t)|_Z^2 + \nu \int_0^t |y_\lambda(s) - y(s)|_Z^2 \, ds \leqq \lambda \int_0^t (v(s), y_\lambda(s) - y(s))_Z \, ds. \tag{3.10}$$

Multiply by $e^{2\nu t}$ and integrate over $[0, t]$; after a short computation, we get

$$\int_0^t |z(s)|^2 \, ds \leqq \frac{\lambda}{\nu} \int_0^t (v(s), z(s)) \, ds - \frac{\lambda}{\nu} \int_0^t e^{2\nu(s-t)} \cdot (v(s), z(s)) \, ds \tag{3.11}$$

where $z = y_\lambda - y$.

Since the situation $\nu \geqq 0$ is straightforward, we assume $\nu < 0$, multiply by ν in (3.11) and combine with (3.10) to obtain

$$|z(t)|_Z^2 \leqq \lambda \cdot C \int_0^t |v(s)|_Y \cdot |z(s)|_Z \, ds. \tag{3.12}$$

The Brezis variant of the Gronwall lemma (see H. Brezis [10, p. 157]) ends the proof.

Remark 2. By Lemma 3.4, the operator $\nabla(S \circ \theta_\varepsilon)(w): L^2(0, T; Y) \to L^2(0, T; X)$ can be extended by continuity to the whole space $L^1(0, T; Y)$ and a similar relation can be written for the adjoint operator $\nabla(S \circ \theta_\varepsilon)(w)^*: L^2(0, T; X) \to L^\infty(0, T; Y^*)$. We have

$$\begin{aligned}
|\nabla(S \circ \theta_\varepsilon)(w)^* v|_{L^\infty(t,T;Y^*)} &= \sup_{|p|_{L^1(t,T;Y)} \leqq 1} \int_t^T (\nabla(S \circ \theta_\varepsilon)(w)^* v, p) \, ds \\
&= \sup_{|p|_{L^1(t,T;Y)} \leqq 1} \int_t^T (v, \nabla(S \circ \theta_\varepsilon)(w)p) \, ds \\
&\leqq \sup_{|p|_{L^1(t,T;Y)} \leqq 1} \int_t^T |v(s)|_X \cdot |\nabla(S \circ \theta_\varepsilon)(w)p|_X \, ds \\
&\leqq C \int_t^T |v(s)|_X \, ds.
\end{aligned}$$

Therefore:

$$|[\nabla(S \circ \theta_\varepsilon)(w)^* v](t)|_{Y^*} \leqq C \int_t^T |v(s)|_X \, ds \quad \text{a.e. } [0, T]. \tag{3.13}$$

THEOREM 3.5. *There exist* $p \in L^\infty(0, T; Y^*)$, $q \in L^2(0, T; X)$ *such that, for* $\varepsilon \to 0$:

$$p_\varepsilon \to p \quad \textit{weakly}^* \textit{ in } L^\infty(0, T; Y^*), \tag{3.14}$$

$$\partial_1 L^\varepsilon(Sy_\varepsilon, u_\varepsilon) \to q \quad \textit{weakly in } L^1(0, T; X), \tag{3.15}$$

$$\langle q(t), F^*p(t)\rangle \in \partial L(Sy^*(t), u^*(t)) \quad \textit{a.e. } [0, T]. \tag{3.16}$$

Proof. By the definition of the subdifferential, we get

$$(\partial_1 L^\varepsilon(Sy_\varepsilon, u_\varepsilon), Sy_\varepsilon - Sy^* - \rho w)_X + (\partial_2 L^\varepsilon(Sy_\varepsilon, u_\varepsilon), u_\varepsilon - v_0)_W$$
$$\geqq L^\varepsilon(Sy_\varepsilon, u_\varepsilon) - L^\varepsilon(Sy^* + \rho w, v_0), \qquad |w|_X = 1,$$

and by Lemma 3.3 and Lemma 3.2:

$$\frac{\rho}{2}|\partial L^\varepsilon(Sy_\varepsilon(t), u_\varepsilon(t))|_X \leqq (|F^*p_\varepsilon(t)|_W + |u_\varepsilon(t) - u^*(t)|_W) \tag{3.17}$$
$$\cdot (C + |u_\varepsilon(t)|_W) + C \quad \text{a.e. } [0, T].$$

Here we also use the assumption that L has a finite Hamiltonian: the Hamiltonian function and its subdifferential are locally bounded on $X \times W$. Therefore we can find a measurable selection $v_0(t)$ of $\partial_2 H(y^*(t) + \rho w, 0)$ such that for ρ sufficiently small and $|w|_X = 1$, we have $|v_0(t)| \leqq C$ and

$$L(y^*(t) + \rho w, v_0(t)) = -H(y^*(t) + \rho w, 0) \leqq C \quad \text{a.e. } t \in [0, T].$$

But from (3.4), (3.13) we can write

$$|p_\varepsilon(t)|_{Y^*} \leqq C \int_t^T |\partial_1 L^\varepsilon(Sy_\varepsilon, u_\varepsilon)|_X \, ds \quad \text{a.e. } [0, T].$$

The Gronwall lemma gives

$$|p_\varepsilon(t)|_{Y^*} \leqq C \quad \text{a.e. } [0, T]$$

and next

$$|\partial_1 L^\varepsilon(Sy_\varepsilon(t), u_\varepsilon(t))|_X \leqq C(1 + |u_\varepsilon(t) - u^*(t)|_W) \cdot (1 + |u^*(t)|_W).$$

The Dunford–Pettis theorem shows that $\partial_1 L^\varepsilon(Sy_\varepsilon, u_\varepsilon) \to q$ weakly in $L^1(0, T; X)$.

We also have $p_\varepsilon \to p$ weakly* in $L^\infty(0, T; Y^*)$. Relation (3.16) and $q \in L^2(0, T; X)$ can be derived by standard arguments because y_ε, u_ε converge strongly by Lemma 3.3.

Remark 3. Relation (3.16) stands for the so called "maximum principle".

Remark 4. The same treatment can be carried out in case u^* is only a local minimum for functional (1.1). We have to take instead of (3.2) the approximation

$$\int_0^T L^\varepsilon(Sy, u) \, dt + \eta \int_0^T |u - u^*|_W^2 \, dt, \tag{3.2}'$$

where η is a positive constant.

As $\eta \cdot \int_0^T |u_\varepsilon - u^*|_W^2 \, dt$ is bounded, when η is large enough we can deduce that $\{u_\varepsilon\}$ is in the neighborhood of u^* considered in the definition of the local minimum.

Remark 5. Taking $M = \partial\varphi$, a subdifferential, and $Z = X = Y$, $S = I$, the identity operator, one can obtain the abstract results of V. Barbu [4].

Remark 6. In the next sections we apply the above results to various problems. To do this we put them in the form (1.1)–(1.3), that is we define the spaces Z, X, Y, W and the operators M, S, F; next we choose an appropriate approximation M^ε, which will not coincide with the Yosida approximation of a maximal monotone

operator. Much effort is needed to check properties (a), (b), (c). For that several existence and continuous dependence results are obtained for certain hyperbolic and parabolic equations (see Theorems 5.1, 6.2).

In the last step, we pass to the limit in (3.4). This is not supplied by Theorem 3.5 because it involves the product of two weakly convergent sequences, and has to be done separately in each application.

4. Hyperbolic control problems. Unilateral conditions in the domain. We shall be concerned with the control problem:

$$\text{(4.1)} \qquad \text{Minimize} \int_0^T L(y, u)\, dt$$

subject to:

$$\text{(4.2)} \qquad y_{tt} - \Delta y + \beta(y_t) \ni Bu \quad \text{a.e. } \Omega \times]0, T[,$$

$$\text{(4.3)} \qquad y(0, x) = y_0(x), \quad y_t(0, x) = v_0(x) \quad \text{a.e. } \Omega,$$

$$\text{(4.4)} \qquad y(t, x) = 0 \quad \text{a.e. } \Gamma \times]0, T[,$$

where $\beta \subseteq R \times R$ is a maximal monotone graph.

The case $\beta(y) = |y|^{p-1} \cdot y$ is briefly discussed by J. L. Lions [14, p. 344].

We denote $V = H_0^1(\Omega)$, $H = L^2(\Omega)$ and we take $W = \{0\} \times U$, U a Hilbert space of control, $Z = V \times H$ where V has the inner product induced by $-\Delta$ such that Z can be identified with its dual. We also define $X = H$, $Y = \{0\} \times V$.

The corresponding operators are $F = \langle 0, B \rangle$ with $B: U \to V$ linear, continuous and $S: V \times H \to H$ is given by $S(y, v) = y$. Operator $M + \nu I: Z \to Z$ is

$$M + \nu I = \begin{bmatrix} 0 & -1 \\ -\Delta & \beta \end{bmatrix}$$

as (4.2) can be put in the form

$$\frac{d}{dt}\begin{bmatrix} y \\ v \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ -\Delta & \beta \end{bmatrix}\begin{bmatrix} y \\ v \end{bmatrix} \ni \begin{bmatrix} 0 \\ Bu \end{bmatrix} \quad \text{a.e. } [0, T].$$

We write M instead of $M + \nu I$ and the approximants M^ε are obtained replacing in M, the graph β by β^ε, where

$$\text{(4.5)} \qquad \beta^\varepsilon(r) = \int_{-\infty}^{\infty} \beta_\varepsilon(r - \varepsilon\theta)\rho(\theta)\, d\theta$$

and $\beta_\varepsilon(r) = \beta((I + \varepsilon\beta)^{-1}(r))$ is the Yosida regularization of β. Function ρ is a Friedrichs mollifier i.e. $\rho \in C_0^\infty(R)$, $\rho(r) \geqq 0$, $\rho(-r) = \rho(r)$, $\rho(r) = 0$ for $|r| \geqq 1$ and $\int_{-\infty}^{\infty} \rho(r)\, dr = 1$.

Equation (3.3) becomes

$$\text{(4.6)} \qquad y_{tt} - \Delta y + \beta^\varepsilon(y_t) = Bu \quad \text{a.e. } \Omega \times]0, T[.$$

To begin with, we discuss the existence in the state equation.

THEOREM 4.1. *Assume that* $f \in L^2(0, T; H_0^1(\Omega))$, $y_0 \in H_0^1(\Omega)$, $\Delta y_0 \in L^2(\Omega)$, $v_0(x) \in$ dom (β) a.e. Ω. *Then there exists a unique solution* y *to*

$$\text{(4.7)} \qquad y_{tt} - \Delta y + \beta(y_t) \ni f \quad \text{a.e. } \Omega \times]0, T[$$

and (4.3), (4.4) *such that* $y \in C(0, T; H_0^1(\Omega))$, $y_t \in C(0, T; L^2(\Omega)) \cap L^\infty(0, T; H_0^1(\Omega))$ *and* $y_{tt} \in L^2(0, T; L^2(\Omega))$.

This result can be found in the book of V. Barbu [3, p. 279]. Since some of the techniques will be useful in the sequel, we outline the proof of the following corollary:

COROLLARY 4.2. *Let* $f_n \to f$ *weakly in* $L^2(0, T; H_0^1(\Omega))$. *Then* $y^n \to y$ *in* $C(0, T; H_0^1(\Omega))$ *and* $y_t^n \to y_t$ *in* $C(0, T; L^2(\Omega))$ *strongly where* y^n *denotes the solution of* (4.7), (4.3), (4.4) *corresponding to* f_n.

Proof. Multiplying by $y_t^n - v_0$ and integrating over $[0, t]$, we get $\{y_t^n\}$, $\{y^n\}$ bounded in $L^\infty(0, T; H)$, $L^\infty(0, T; V)$.

Let A_H denote the realization of $-\Delta$ in H, which is maximal monotone, and A_λ the Yosida approximation of A_H.

Multiply by $A_\lambda y_t^n$. Then $(A_\lambda y_t^n, \beta(y_t^n)) \geqq 0$ since $(A_\lambda z, \beta(z)) \geqq 0$; for all $z \in \text{dom}(\beta)$ is equivalent with $(A_H z, \beta_\varepsilon(z)) \geqq 0$ for $z \in \text{dom}(A_H)$ which is clearly true by:

$$\int_\Omega -\Delta z \beta_\varepsilon(z)\, dx = \int_\Omega \text{grad}\, z \cdot \text{grad}\, \beta_\varepsilon(z)\, dx \geqq 0$$

from the monotonicity of β_ε (see V. Barbu [3], p. 183).

We infer

$$\frac{1}{2}(A_\lambda y_t^n(t), y_t^n(t)) - \frac{1}{2}(A_\lambda v_0, v_0) + \frac{1}{2}(A_H y^n(t), A_\lambda y^n(t))$$

$$-\frac{1}{2}(A_H y_0, A_\lambda y_0) \leqq \int_0^t (f_n, A_\lambda y_t^n)\, ds.$$

Since f_n is bounded in $L^2(0, T; V)$, a standard argument gives $\{A_\lambda y^n\}$ bounded in $L^\infty(0, T; H)$ and $\{(I + \lambda A_H)^{-1} y_t^n\}$ bounded in $L^\infty(0, T; V)$, with respect to λ, n.

The first yields $(I + \lambda A_H)^{-1} y^n \to y^n$ strongly in $C(0, T; H)$ as $\lambda \to 0$. As the operator is linear, we infer the same for the derivative $(I + \lambda A_H)^{-1} y_t^n \to y_t^n$ weakly* in $L^\infty(0, T; V)$.

This allows us to make $\lambda \to 0$ and to get $\{y_t^n\}$, $\{A_H y_t^n\}$ bounded in $L^\infty(0, T; V)$ and $L^\infty(0, T; H)$ respectively.

Multiplying by y_{tt}^n and integrating over $[0, T]$, it is easy to see $\{y_{tt}^n\}$ bounded in $L^2(0, T; H)$. From (4.7) we obtain $\{\beta(y_t^n)\}$ bounded in $L^2(0, T; H)$ too.

Since $\text{dom}(A_H) \subset V$ is compact, from the above boundedness we get $y^n \to y$ in $C(0, T; V)$ on a convenient subsequence, by the Ascoli–Arzèla theorem. Then

$$y_t^n \to y_t \quad \text{strongly in } C(0, T; H),$$

$$y_{tt}^n \to y_{tt} \quad \text{weakly in } L^2(0, T; H),$$

$$Ay^n \to Ay \quad \text{weakly* in } L^\infty(0, T; H),$$

$$\beta(y_t^n) \to \beta(y_t) \quad \text{weakly in } L^2(0, T; H)$$

by the demiclosedness of β and y is the unique solution of (4.7), (4.3), (4.4). Therefore the convergence is true on the initial sequence and the proof is finished.

We start to check conditions (a), (b), (c).

LEMMA 4.3. *Let* $f_\varepsilon \to f$ *weakly in* $L^2(0, T; H_0^1(\Omega))$. *Then* $y^\varepsilon = \theta_\varepsilon(f_\varepsilon) \to y = \theta(f)$ *strongly in* $C(0, T; H_0^1(\Omega))$ *and* $y_t^\varepsilon \to y_t$ *strongly in* $C(0, T; L^2(\Omega))$.

This is a variant of the above corollary.

LEMMA 4.4. *The operator* $S \circ \theta_\varepsilon : L^2(0, T; H_0^1(\Omega)) \to L^2(0, T; L^2(\Omega))$ *is Gâteaux differentiable and* $r = \nabla(S \circ \theta_\varepsilon)(f)g$ *satisfies in a weak sense*:

$$r_{tt} - \Delta r + \nabla \beta^\varepsilon((S \circ \theta_\varepsilon)(f)t) \cdot r_t = g \quad \text{a.e. } \Omega \times]0, T[, \tag{4.8}$$

$$r(0, x) = r_t(0, x) = 0 \quad \text{a.e. } \Omega, \tag{4.9}$$

$$r(t, x) = 0 \quad \text{a.e. } \Gamma \times]0, T[\tag{4.10}$$

for $f, g \in L^2(0, T; H_0^1(\Omega))$. *Moreover* $r \in L^\infty(0, T; H_0^1(\Omega))$ *and* $r_t \in L^\infty(0, T; L^2(\Omega))$, $r_{tt} \in L^\infty(0, T; H^{-1}(\Omega))$. *The solution to problem* (4.8)–(4.10) *is unique.*

Proof. Denote $y^\lambda = \theta_\varepsilon(f + \lambda g)$, $y = \theta_\varepsilon(f)$, $z^\lambda = (y^\lambda - y)/\lambda$.

Then, by the definition of θ_ε, we get:

$$\frac{1}{2}\frac{d}{dt}|z_t^\lambda|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}|\nabla z^\lambda|^2_{L^2(\Omega)} \leqq \int_\Omega g \cdot z_t^\lambda \, dx.$$

We infer $\{z^\lambda\}$ bounded in $L^\infty(0, T; H_0^1(\Omega))$ and $\{z_t^\lambda\}$ bounded in $L^\infty(0, T; L^2(\Omega))$. Besides:

$$y^\lambda \to y \quad \text{strongly in } C(0, T; H_0^1(\Omega)),$$

$$y_t^\lambda \to y_t \quad \text{strongly in } C(0, T; L^2(\Omega)).$$

As $H_0^1(\Omega) \subset L^2(\Omega)$ is compact, we have:

$$z^\lambda \to r \quad \text{strongly in } C(0, T; L^2(\Omega)),$$

$$z_t^\lambda \to r_t \quad \text{weakly* in } L^\infty(0, T; L^2(\Omega)),$$

$$z_{tt}^\lambda \to r_{tt} \quad \text{weakly* in } L^\infty(0, T; H^{-1}(\Omega)).$$

The nonlinear term $(\beta^\varepsilon(y_t^\lambda) - \beta^\varepsilon(y_t))/\lambda$ is bounded in $L^\infty(0, T; L^2(\Omega))$ since β^ε is Lipschitz of constant $1/\varepsilon$. We write it in the form:

$$\frac{\beta^\varepsilon(y_t^\lambda(t, x)) - \beta^\varepsilon(y_t(t, x))}{\lambda} = \frac{\beta^\varepsilon(y_t^\lambda(t, x)) - \beta^\varepsilon(y_t(t, x))}{y_t^\lambda(t, x) - y_t(t, x)} \cdot z_t^\lambda(t, x).$$

As β^ε is differentiable we get:

$$\frac{\beta(y_t^\lambda(t, x)) - \beta^\varepsilon(y_t(t, x))}{y_t^\lambda(t, x) - y_t(t, x)} \to \nabla\beta^\varepsilon(y_t(t, x)) \quad \text{a.e. } \Omega \times]0, T[,$$

since $y_t^\lambda(t, x) \to y_t(t, x)$ a.e. $\Omega \times]0, T[$. Then we can deduce:

$$\frac{\beta^\varepsilon(y_t^\lambda) - \beta^\varepsilon(y_t)}{\lambda} \to \nabla\beta^\varepsilon(y_t) \cdot r_t \quad \text{weakly in } L^2(Q).$$

We conclude that (4.8)–(4.10) are satisfied and the proof is finished.

For further needs we explain the adjoint operator $\nabla(S \circ \theta_\varepsilon)(f)^*$. Denote $\nabla(S \circ \theta_\varepsilon)(f)^* q = p$, where $f \in L^2(0, T; H_0^1(\Omega))$, $q \in L^2(0, T; L^2(\Omega))$ and $p \in L^2(0, T; H^{-1}(\Omega))$.

We define the adjoint state p by:

$$p = -m_t, \tag{4.11}$$

$$m_{tt} - \Delta m - \nabla\beta^\varepsilon(y_t) \cdot m_t = \int_t^T q \quad \text{a.e. } Q, \tag{4.12}$$

$$m(T, x) = m_t(T, x) = 0 \quad \text{a.e. } \Omega, \tag{4.13}$$

$$m(t, x) = 0 \quad \text{a.e. } \Gamma \times]0, T[. \tag{4.14}$$

Remark 1. The existence and uniqueness in (4.12)–(4.14) can be established, for instance, by approximating function $\nabla\beta^\varepsilon(y_t)$ by a $C^\infty \cap L^\infty$ mapping.

It is easy to obtain a priori estimates and to pass to the limit. We shall give a more detailed motivation, although by another method, in the next section.

We pass to condition (c) now.

LEMMA 4.5. *Let* $y^\varepsilon = \theta_\varepsilon(f)$ *and* $y = \theta(f)$. *Then*

$$|y^\varepsilon(t) - y(t)|_V + |y_t^\varepsilon(t) - y_t(t)|_H \leqq C \cdot \varepsilon^{1/2}. \tag{4.15}$$

Proof. Denote $y^\lambda = \theta_\lambda(f)$. The following equality is true:

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}|y_t^\varepsilon - y_t^\lambda|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}|\nabla(y^\varepsilon - y^\lambda)|^2_{L^2(\Omega)} \\ &\quad + \int_\Omega (\beta^\varepsilon(y_t^\varepsilon) - \beta^\lambda(y_t^\lambda))(y_t^\varepsilon - y_t^\lambda)\, dx = 0. \end{aligned} \tag{4.16}$$

We recall the inequality:

$$\begin{aligned} &(\beta^\varepsilon(u) - \beta^\varepsilon(v))(u - v) \\ &\geqq \int_{-\infty}^{\infty} (\beta_\varepsilon(u - \varepsilon\tau) - \beta_\lambda(v - \lambda\tau)) \cdot (\varepsilon\beta_\varepsilon(u - \varepsilon\tau) - \lambda\beta_\lambda(v - \lambda\tau))\rho(\tau)\, d\tau \\ &\quad + (\varepsilon - \lambda)\int_{-\infty}^{\infty} (\beta_\varepsilon(u - \varepsilon\tau) - \beta_\lambda(v - \lambda\tau))\tau\rho(\tau)\, d\tau. \end{aligned} \tag{4.17}$$

Combining (4.16), (4.17) and integrating over $[0, t]$, as $\{\beta^\varepsilon(y_t^\varepsilon)\}$ is bounded in $L^2(Q)$ (see the proof of Corollary 4.2), we get

$$\tfrac{1}{2}|y_t^\varepsilon(t) - y_t^\lambda(t)|^2_H + \tfrac{1}{2}|y^\varepsilon(t) - y^\lambda(t)|^2_V \leqq C(\varepsilon + \lambda). \tag{4.18}$$

Then $y^\varepsilon \to y$ strongly in $C(0, T; H_0^1(\Omega))$ and $y_t^\varepsilon \to y_t$ strongly in $C(0, T; L^2(\Omega))$ and it can be deduced easily $y = \theta(f)$. So, for $\lambda \to 0$ in (4.18), we obtain (4.15).

Now, we are able to use the abstract results of § 3. We write the approximate control problem:

$$\text{Minimize} \int_0^T L^\varepsilon(y(t), u(t))\, dt + \frac{1}{2}\int_0^T |u(t) - u^*(t)|^2_U\, dt \tag{4.19}$$

with state equation:

$$y_{tt} - \Delta y + \beta^\varepsilon(y_t) = Bu \quad \text{a.e. } \Omega \times]0, T[, \tag{4.20}$$

$$y(0, x) = y_0(x), \quad y_t(0, x) = v_0(x) \quad \text{a.e. } \Omega, \tag{4.21}$$

$$y(t, x) = 0 \quad \text{a.e. } \Gamma \times]0, T[. \tag{4.22}$$

PROPOSITION 4.6. *Problem* (4.19)–(4.22) *has solution* $\langle y^\varepsilon, u_\varepsilon \rangle$ *in* $W^{2,2}(0, T; L^2(\Omega)) \times L^2(0, T; U)$ *and there exists* $m^\varepsilon \in C(0, T; L^2(\Omega))$ *such that the approximating optimality system*

$$y_{tt}^\varepsilon - \Delta y^\varepsilon + \beta^\varepsilon(y_t^\varepsilon) = Bu_\varepsilon \quad \text{a.e. } \Omega \times]0, T[, \tag{4.23}$$

$$m_{tt}^\varepsilon - \Delta m^\varepsilon - \nabla\beta^\varepsilon(y_t^\varepsilon) \cdot m_t^\varepsilon = -\int_t^T q_\varepsilon \quad \text{a.e. } \Omega \times]0, T[, \tag{4.24}$$

$$y^\varepsilon(0, x) = y_0(x), \quad y_t^\varepsilon(0, x) = v_0(x) \quad \text{a.e. } \Omega, \tag{4.25}$$

$$m^\varepsilon(T, x) = 0, \quad m_t^\varepsilon(T, x) = 0 \quad \text{a.e. } \Omega, \tag{4.26}$$

$$y^\varepsilon(t, x) = 0, \quad m^\varepsilon(t, x) = 0 \quad \text{a.e. } \Gamma \times]0, T[, \tag{4.27}$$

$$\langle q_\varepsilon(t), -B^*m_t^\varepsilon(t)+u_\varepsilon(t)-u^*(t)\rangle \in \partial L^\varepsilon(y^\varepsilon(t), u_\varepsilon(t)) \quad \text{a.e. } [0, T]. \tag{4.28}$$

is satisfied. Moreover, we have $y^\varepsilon \to y^*$ *strongly in* $C(0, T; H_0^1(\Omega))$ $y_t^\varepsilon \to y_t^*$ *strongly in* $C(0, T; L^2(\Omega))$, $u_\varepsilon \to u^*$ *strongly in* $L^2(0, T; U)$, $p^\varepsilon = -m_t^\varepsilon \to p$ *weakly* in* $L^\infty(0, T; L^2(\Omega))$ *and* $q_\varepsilon \to q$ *weakly in* $L^1(0, T; L^2(\Omega))$ *where*

$$\langle q(t), B^*p(t)\rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T]. \tag{4.29}$$

To pass to the limit in (4.24) we make the additional assumption that β is locally Lipschitz and

$$|\nabla\beta(y)\cdot y| \leqq C(|\beta(y)|+y^2+1) \quad \text{a.e. } R. \tag{4.30}$$

From (4.30) it is easy to obtain:

$$|\nabla\beta^\varepsilon(y)\cdot y| \leqq C(|\beta^\varepsilon(y)|+y^2+1) \quad \forall y \tag{4.31}$$

with C a positive constant, independent of ε.

We have:

THEOREM 4.7. *Let* $\langle y^*, u^*\rangle \in W^{2,2}(0, T; L^2(\Omega)) \times L^2(0, T; U)$ *be an optimal pair for problem* (4.1)–(4.4) *and* β *satisfy* (4.30). *There exist functions* m *in* $L^\infty(0, T; H_0^1(\Omega)) \cap W^{1,\infty}(0, T; L^2(\Omega))$, $q \in L^2(Q)$ *and* $h \in L^1(Q)$ *which satisfy*:

$$y_{tt}^* - \Delta y^* + \beta(y_t^*) \ni Bu^* \quad \text{a.e. } \Omega\times]0, T[,$$

$$m_{tt} - \Delta m - h = -\int_t^T q \quad \text{a.e. } \Omega\times]0, T[,$$

$$y^*(0, x) = y_0(x), \quad y_t^*(0, x) = v_0(x) \quad \text{a.e. } \Omega,$$

$$m(T, x) = 0, \quad m_t(T, x) = 0 \quad \text{a.e. } \Omega,$$

$$\langle q(t), -B^*m_t(t)\rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T].$$

Proof. Multiplying (4.24) by m_t^ε, we get $\{m^\varepsilon\}$ bounded in $L^\infty(0, T; H_0^1(\Omega))$ and $\{m_t^\varepsilon\}$ bounded in $L^\infty(0, T; L^2(\Omega))$.

Fix n a natural number and consider $E_n^\varepsilon = \{(x, t) \in Q; |y_t^\varepsilon(x, t)| \leqq n\}$, $\varepsilon > 0$.

Then:

$$|\nabla\beta^\varepsilon(y_t^\varepsilon(t, x))| \leqq C_n \quad \text{a.e. } (t, x) \in E_n^\varepsilon$$

with C_n independent of ε as β is locally Lipschitz. Let E be a measurable subset of $Q = \Omega\times]0, T[$.

$$\left|\int_E \nabla\beta^\varepsilon(y_t^\varepsilon)\cdot m_t^\varepsilon \, dx\, dt\right| \leqq \int_{E\cap E_n^\varepsilon} \nabla\beta^\varepsilon(y_t^\varepsilon)\cdot|m_t^\varepsilon| \, dx\, dt + \int_{E-E_n^\varepsilon} \nabla\beta^\varepsilon(y_t^\varepsilon)|m_t^\varepsilon| \, dx\, dt$$

$$\leqq C\cdot\frac{1}{n}\int_{E-E_n^\varepsilon} |\beta^\varepsilon(y_t^\varepsilon)|\cdot|m_t^\varepsilon| \, dx\, dt + C\cdot\frac{1}{n} + C_n\int_E |m_t^\varepsilon| \, dx\, dt$$

$$+ C\int_{E-E_n^\varepsilon} |y_t^\varepsilon|\cdot|m_t^\varepsilon| \, dx\, dt.$$

Taking into account that $\{m_t^\varepsilon\}$ is bounded in $L^\infty(0, T; L^2(\Omega))$ and $\beta^\varepsilon(y_t^\varepsilon)$ bounded in $L^2(Q)$, we obtain:

$$\left|\int_E \nabla\beta^\varepsilon(y_t^\varepsilon)\cdot m_t^\varepsilon \, dx\, dt\right| \leqq C\cdot\mu(E)^{1/2}C_n + C\cdot\frac{1}{n} + C\left(\int_{E-E_n^\varepsilon} |y_t^\varepsilon|^2\right)^{1/2}.$$

Since $\{y_t^\varepsilon\}$ is bounded in $L^\infty(0, T; H_0^1(\Omega))$, by the Sobolev embedding theorem it yields $\{y_t^\varepsilon\}$ bounded in $L^s(Q)$ with some $s>2$. Then the last term of the sum is equicontinuous too. The Dunford–Pettis criterion gives:

$$\nabla\beta^\varepsilon(y_t^\varepsilon)\cdot m_t^\varepsilon \to h \quad \text{weakly in } L^1(Q).$$

Remark 2. In some important cases more information concerning the function h can be deduced.

Example 1. *The convex case.* We shall suppose that

$$\beta = \xi - \iota \tag{4.32}$$

where ξ, ι are convex, real functions on R.

THEOREM 4.8. *Under the assumptions of Theorem* 4.7 *and also* (4.32), *there exist functions* $m \in L^\infty(0, T; H_0^1(\Omega)) \cap W^{1,\infty}(0, T; L^2(\Omega))$ *and* $q \in L^2(Q)$ *which satisfy*:

$$y_{tt}^* - \Delta y^* + \beta(y_t^*) \ni Bu^* \quad \text{a.e. } Q,$$

$$m_{tt} - \Delta m - \partial\beta(y_t^*)m_t \ni -\int_t^T q \quad \text{a.e. } Q,$$

$$y^*(0, x) = y_0(x), \quad y_t^*(0, x) = v_0(x) \quad \text{a.e. } \Omega,$$

$$m(T, x) = m_t(T, x) = 0 \quad \text{a.e. } \Omega,$$

$$\langle q(t), -B^* m_t(t)\rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T],$$

in a weak sense. Here $\partial\beta$ *is the generalized gradient of the locally Lipschitz mapping* β.

Proof. For the sake of simplicity we take β convex.

Write $m_t^\varepsilon = m_+^\varepsilon - m_-^\varepsilon$, where m_+^ε, m_-^ε denote the positive and the negative part of m_t^ε. Extracting more subsequences we can obtain:

$$m_+^\varepsilon \to v^+, \; m_-^\varepsilon \to v^- \quad \text{weakly in } L^2(Q),$$

$$m_t = v^+ - v^- \tag{4.33}$$

and (adding a constant if necessary) m_+^ε, m_-^ε are strict positive a.e.

We give now a more precise calculation of $\nabla\beta^\varepsilon(y)$, $y \in R$, available for locally Lipschitz functions β. As β^ε can be written in the form

$$\beta^\varepsilon(y) = \int_{-\infty}^{\infty} \frac{I - (I+\varepsilon\beta)^{-1}(y - \varepsilon\theta)}{\varepsilon}\rho(\theta)\, d\theta,$$

we infer

$$\nabla\beta^\varepsilon(y) = \int_{-\infty}^{\infty} \frac{\nabla\beta((I+\varepsilon\beta)^{-1}(y-\varepsilon\theta))}{1+\varepsilon\nabla\beta((I+\varepsilon\beta)^{-1}(y-\varepsilon\theta))}\rho(\theta)\, d\theta. \tag{4.34}$$

From the Egorov theorem, for every $\eta > 0$, there is $Q_\eta \subset Q$ with mes $(Q - Q_\eta) < \eta$ and $y_t^\varepsilon \to y_t^*$ uniformly on Q_η. Then $\nabla\beta^\varepsilon(y_t^\varepsilon) \to g$ weakly* in $L^\infty(Q_\eta)$ and by a lemma due to Barbu [6], we have

$$g(t, x) \in \partial\beta(y_t^*(t, x)) \quad \text{a.e. } Q_\eta. \tag{4.35}$$

We are interested in the weak convergence of $\nabla\beta^\varepsilon(y_t^\varepsilon)\cdot m_+^\varepsilon$ in $L^2(Q_\eta)$. Consider any f in $L^2(Q_\eta)$:

$$\int_{Q_\eta} \nabla\beta^\varepsilon(y_t^\varepsilon)\cdot m_+^\varepsilon \cdot f\, dx\, dt = \int_{-1}^{1} \rho(\tau)\, d\tau \int_{Q_\eta} m_+^\varepsilon \cdot f \cdot \partial\beta(\,\cdot\,)\frac{1}{1+\varepsilon\partial\beta(\cdot)}\, dx\, dt.$$

We omit $(I+\varepsilon\beta)^{-1}$ in order to shorten the notation.

On Q_η $1+\varepsilon\partial\beta(\cdot)\to 1$ uniformly for $\varepsilon\to 0$, so we have to study only the integral:

$$\int_{Q_\eta} m_+^\varepsilon\cdot\partial\beta(\cdot)\cdot f\,dx\,dt, \qquad \theta\in[-1,1]\text{ fixed.} \tag{4.36}$$

Consider the saddle function:

$$K(m,y)=\begin{cases} m\cdot\beta(y), & m\geqq 0,\\ -\infty, & m<0,\end{cases} \tag{4.37}$$

which is proper, closed. The maximal monotone operator ∂K in $R^2\times R^2$ is given by

$$\partial K(m,y)=\langle -\beta(y), m\partial\beta(y)\rangle. \tag{4.38}$$

Take the realization of ∂K in $L^2(Q_\eta)\times L^2(Q_\eta)$:

$$\partial\tilde{K}(m,y)(t,x)=\partial K(m(t,x),y(t,x)) \quad \text{a.e. } Q_\eta \tag{4.39}$$

for every $m, y\in L^2(Q_\eta)$, $m\geqq 0$ a.e. Q_η.

Operator $\partial\tilde{K}$ is maximal monotone. We have:

$$\langle -\beta(\cdot), m_+^\varepsilon\partial\beta(\cdot)\rangle\in\partial\tilde{K}(m_+^\varepsilon,(I+\varepsilon\beta)^{-1}(y_t^\varepsilon-\varepsilon\theta)). \tag{4.40}$$

We remark that

$$\langle -\beta(\cdot), m_+^\varepsilon\cdot\partial\beta(\cdot)\rangle\to\langle -\beta(y_t^*),\tilde{h}\rangle \quad \text{weakly in } L^2(Q_\eta)\times L^2(Q_\eta), \tag{4.41}$$

$$\langle m_+^\varepsilon,(I+\varepsilon\beta)^{-1}(y_t^\varepsilon-\varepsilon\theta)\rangle\to\langle v^+,y_t^*\rangle \tag{4.42}$$

weakly in $L^2(Q_\eta)\times L^2(Q_\eta)$ and

$$\lim_{\lambda,\varepsilon\to 0}(\langle m_+^\varepsilon,(I+\varepsilon\beta)^{-1}(y_t^\varepsilon-\varepsilon\theta)\rangle-\langle m_+^\lambda,(I+\lambda\beta)^{-1}(y_t^\lambda-\lambda\theta)\rangle \cdot\langle -\beta(\cdot), m_+^\varepsilon\partial\beta(\cdot)\rangle-\langle -\beta(\cdot), m_+^\lambda\partial\beta(\cdot)\rangle)_{L^2(Q_\eta)\times L^2(Q_\eta)}=0 \tag{4.43}$$

since $\beta((I+\varepsilon\beta)^{-1}(y_t^\varepsilon-\varepsilon\theta))\to\beta(y_t^*)$ and $(I+\varepsilon\beta)^{-1}(y_t^\varepsilon-\varepsilon\theta)\to y_t^*$ uniformly on Q_η.

Applying a well-known property of monotone operators (Barbu [3, p. 42]) we get:

$$\begin{aligned}&\langle -\beta(y_t^*),\tilde{h}\rangle\in\partial\tilde{K}(v^+,y_t^*) \quad \text{so}\\ &\tilde{h}(t,x)\in v^+(t,x)\cdot\partial\beta(y_t^*(t,x)) \quad \text{a.e. } Q.\end{aligned} \tag{4.44}$$

In a similar way, we derive:

$$\begin{aligned}&\lim_{\varepsilon\to 0} m_-^\varepsilon\cdot\partial\beta(\cdot)=\underset{\sim}{h} \quad \text{weakly in } L^2(Q_\eta) \quad \text{and}\\ &\underset{\sim}{h}(t,x)\in v^-(t,x)\cdot\partial\beta(y_t^*(t,x)) \quad \text{a.e. } Q.\end{aligned} \tag{4.45}$$

We have:

$$h(t,x)=\tilde{h}(t,x)-\underset{\sim}{h}(t,x)\in v^+(t,x)\cdot\partial\beta(y_t^*(t,x))-v^-(t,x)\cdot\partial\beta(y_t^*(t,x)).$$

We write that $h(t,x)=m_t\cdot\partial\beta(y_t^*(t,x))$ a.e. Q by convention since it is possible that $\tilde{h}$, $\underset{\sim}{h}$ result by using different selections of the (multivalued) subdifferential $\partial\beta(y_t^*)$.

When (4.32) is satisfied, we can prove in the same conventional sense $h(t,x)\in m_t\cdot\partial\beta(y_t^*(t,x))$ since $\beta^\varepsilon=\xi^\varepsilon-\iota^\varepsilon$ and by Barbu [6, Lemma 3] we have

$$\nabla\beta^\varepsilon(y_t^\varepsilon)\to\partial\beta(y_t^*) \quad \text{weakly in } L^2(Q_\eta)$$

where $\partial\beta$ is the generalized gradient of β.

Example 2. *The differentiable case.* Here we assume that β is a differentiable function. Then

$$\nabla\beta^\varepsilon(y_t^\varepsilon(t,x)) \to \nabla\beta(y_t^*(t,x)) \quad \text{a.e. } Q_\eta,$$

$$\nabla\beta^\varepsilon(y_t^\varepsilon) \to \nabla\beta(y_t^*) \quad \text{weakly* in } L^\infty(Q_\eta)$$

with Q_η defined in (4.35). It follows $\nabla\beta^\varepsilon(y_t^\varepsilon) \to \nabla\beta(y_t^*)$ strongly in $L^2(Q_\eta)$ by the Lebesgue theorem. In this case we have

$$\nabla\beta^\varepsilon(y_t^\varepsilon)\cdot m_t^\varepsilon \to \nabla\beta(y_t^*)\cdot m_t \quad \text{weakly in } L^1(Q_\eta)$$

and a result similar to Theorem 4.8 can be stated.

Of course, a more direct approach could be carried out in this situation.

5. Hyperbolic control problems. Unilateral conditions on the boundary. Now, we consider the problem

$$\text{(5.1)} \qquad \text{Minimize} \int_0^T L(y,u)\,dt$$

subject to

$$\text{(5.2)} \qquad y_{tt} - \Delta y = Bu \quad \text{a.e. } \Omega\times]0,T[,$$

$$\text{(5.3)} \qquad y(0,x) = y_0(x), \qquad y_t(0,x) = v_0(x) \quad \text{a.e. } \Omega,$$

$$\text{(5.4)} \qquad -\frac{\partial y}{\partial n} \in \beta(y_t) \quad \text{a.e. } \Gamma\times]0,T[.$$

First we deal with the existence in equation:

$$\text{(5.5)} \qquad y_{tt} - \Delta y = f \quad \text{a.e. } \Omega\times]0,T[$$

and (5.3), (5.4), where $f \in L^2(0,T;H^1(\Omega))$. In Brezis [9] existence of strong solutions is established for $f \in W^{1,1}(0,T;L^2(\Omega))$. The following theorem is a variant of his result and can be compared with Theorem 4.1.

THEOREM 5.1. *Assume that* $y_0 \in H^2(\Omega)$, $v_0 \in H^1(\Omega)$, $-\partial y_0/\partial n \in \beta(v_0)$ *and* $f \in L^2(0,T;H^1(\Omega))$. *Then there exists a unique solution to* (5.3)–(5.5) *which satisfies* $y \in L^\infty(0,T;H^2(\Omega)) \cap C(0,T;H^1(\Omega))$, $y_t \in L^\infty(0,T;H^1(\Omega)) \cap C(0,T;L^2(\Omega))$ *and* $y_{tt} \in L^2(0,T;L^2(\Omega))$.

Proof. Let $f_n \in W^{1,1}(0,T;L^2(\Omega))$ and $f_n \to f$ strongly in $L^2(0,T;H^1(\Omega))$. The approximate problem:

$$\text{(5.6)} \qquad y_{tt}^n - \Delta y^n = f_n \quad \text{a.e. } \Omega\times]0,T[$$

and (5.3)–(5.4) has a unique solution by Brezis [9, Thm. III.5] which verifies the above regularity.

Multiply (5.6) by y_t^n, use Green's formula and (5.4) to obtain $\{y^n\}$ bounded in $L^\infty(0,T;H^1(\Omega))$ and $\{y_t^n\}$ bounded in $L^\infty(0,T;L^2(\Omega))$. Next integrate over $[t,t+h]\subset[0,T]$, multiply by $-\Delta[y_n(t+h)-y_n(t)]$ and divide by h^2; then from (5.4) we get:

$$\frac{1}{2}\frac{d}{dt}\left|\operatorname{grad}\frac{y^n(t+h)-y^n(t)}{h}\right|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}\left|\Delta\frac{1}{h}\int_t^{t+h} y^n\right|^2_{L^2(\Omega)}$$

$$\leqq -\int_\Omega \frac{1}{h}\int_t^{t+h} f_n\,d\tau\cdot\Delta\frac{y^n(t+h)-y^n(t)}{h}\,dx.$$

We integrate over $[0, t]$ to get

$$\frac{1}{2}\left|\nabla\frac{y^n(t+h)-y^n(t)}{h}\right|^2_{L^2(\Omega)}+\frac{1}{2}\left|\Delta\frac{1}{h}\int_t^{t+h}y^n\right|^2_{L^2(\Omega)}$$

$$\leqq\frac{1}{2}\left|\nabla\frac{y^n(h)-y^n(0)}{h}\right|^2_{L^2(\Omega)}+\frac{1}{2}\left|\Delta\frac{1}{h}\int_0^{h}y^n\right|^2_{L^2(\Omega)}$$

$$-\int_0^t\int_\Omega\left(\frac{1}{h}\int_\tau^{\tau+h}f_n\,ds\right)\cdot\Delta\frac{y^n(\tau+h)-y^n(\tau)}{h}\,dx\,d\tau.$$

Letting $h\to 0$, which is possible in our assumptions, we get:

$$\tfrac{1}{2}|\nabla y_t^n|^2_{L^2(\Omega)}+\tfrac{1}{2}|\Delta y^n|^2_{L^2(\Omega)}\leqq\tfrac{1}{2}|\nabla v_0|^2_{L^2(\Omega)}+\tfrac{1}{2}|\Delta y_0|^2_{L^2(\Omega)}-\int_0^t\int_\Omega f_n\cdot\Delta y_t^n\,dx\,d\tau.$$

The integral in the right-hand side is in fact the duality between $H^1(\Omega)$ and $(H^1(\Omega))^*$.

As $\Delta: H^1(\Omega)\to(H^1(\Omega))^*$ is linear continuous and $\{y_t^n\}$ is bounded in $L^\infty(0, T; L^2(\Omega))$ we can estimate:

$$\tfrac{1}{2}|\nabla y_t^n|^2_{L^2(\Omega)}+\tfrac{1}{2}|\Delta y^n|^2_{L^2(\Omega)}$$

$$\leqq\tfrac{1}{2}|\nabla v_0|^2_{L^2(\Omega)}+\tfrac{1}{2}|\Delta u_0|^2_{L^2(\Omega)}+C\int_0^t|y_t^n|_{H^1(\Omega)}\,d\tau\leqq C+\int_0^t|\nabla y_t^n|_{L^2(\Omega)}\,d\tau.$$

By a variant of Gronwall's lemma we have that $\{y_t^n\}$ is bounded in $L^\infty(0, T; H^1(\Omega))$, $\{\Delta y^n\}$ is bounded in $L^\infty(0, T; L^2(\Omega))$ and $\{y_{tt}^n\}$ is bounded in $L^2(0, T; L^2(\Omega))$.

To prove the regularity in x of the solution we use the method of translations parallel to the boundary due to Agmon, Douglis, Nirenberg [2]. For the sake of simplicity we follow Brezis [9] and suppose $\Omega=\{x=(x', x_N);\ x_N>0\}$. We denote

$$\delta y(x)=\frac{y(x+he_j)-y(x)}{h},\qquad \delta^* y(x)=\frac{y(x-he_j)-y(x)}{h}$$

where $1\leqq j\leqq N-1$.

Multiply (3.7) by $\delta^*\delta(y_t^n)$ and use Green's formula to obtain

$$\frac{1}{2}\frac{d}{dt}\int_\Omega|\delta y_t^n|^2\,dx+\frac{1}{2}\frac{d}{dt}\int_\Omega|\nabla\delta y^n|^2\,dx\leqq\int_\Omega f_n\cdot\delta^*\delta(y_t^n)\,dx.$$

Now integrate over $[0, t]$ and let $h\to 0$ to infer

$$\frac{1}{2}\int_\Omega\left|\frac{\partial y_t^n}{\partial x_j}\right|^2dx-\frac{1}{2}\int_\Omega\left|\frac{\partial v_0}{\partial x_j}\right|^2dx+\frac{1}{2}\int_\Omega\sum_{i=1}^N\left|\frac{\partial y^n}{\partial x_i\,\partial x_j}\right|^2dx$$

$$-\frac{1}{2}\int_\Omega\sum_{i=1}^N\left|\frac{\partial y_0}{\partial x_i\,\partial x_j}\right|^2dx\leqq\int_0^t\int_\Omega\frac{\partial f_n}{\partial x_j}\cdot\frac{\partial y_t^n}{\partial x_j}\,dx\,d\tau.$$

As $\{y_t^n\}$, $\{y_{tt}^n\}$ are bounded, we get $\{y^n\}$ bounded in $L^2(0, T; H^2(\Omega))$.

By the Aubin [1] compactness result, as $H^2(\Omega)$ is compactly embedded in $H^{5/3}(\Omega)$, by taking a convenient subsequence if necessary, we get

$$y^n\to y\quad\text{strongly in } L^2(0, T; H^{5/3}(\Omega)).$$

Then the trace theorem gives

$$\frac{\partial y^n}{\partial n} \to \frac{\partial y}{\partial n} \quad \text{strongly in } L^2(0, T; L^2(\Gamma)).$$

Similarly, we have

$$y_t^n \to y_t \quad \text{strongly in } L^2(0, T, H^{2/3}(\Omega)),$$

$$y_t^n|_\Gamma \to y_t|_\Gamma \quad \text{strongly in } L^2(0, T; L^2(\Gamma)).$$

A well-known property of maximal monotone operators shows that $-\partial y/\partial n \in \beta(y_t)$ a.e. $\Gamma \times]0, T[$. Then obviously y satisfies (5.3)–(5.5). Uniqueness is easy to prove and it implies that $y^n \to y$ without extracting subsequences.

COROLLARY 5.2. *Let $f_n \to f$ weakly in $L^2(0, T; H^1(\Omega))$. Then $y^n \to y$ strongly in $C(0, T; H^1(\Omega))$ and $y_t^n \to y_t$ strongly in $C(0, T; L^2(\Omega))$.*

Now we are in position to apply the abstract scheme developed in § 3. We take the spaces $W = \{0\} \times U$, $Z = H^1(\Omega) \times L^2(\Omega)$, $X = L^2(\Omega)$, $Y = H^1(\Omega)$. The corresponding operators are $F = \langle 0, B \rangle$, $B: U \to H^1(\Omega)$ linear, continuous, $S: H^1(\Omega) \times L^2(\Omega) \to L^2(\Omega)$ is given by $S(y, v) = y$ also linear, continuous. The nonlinear operator M is defined by

$$M = \begin{bmatrix} 0 & -1 \\ -\Delta & \partial\varphi \end{bmatrix}$$

where $\varphi: H^1(\Omega) \to]-\infty, +\infty]$ is the convex, lower semicontinuous function:

$$\varphi(u) = \int_\Gamma j(u)\, d\Gamma \tag{5.7}$$

with $j: R \to]-\infty, +\infty]$ being the convex, lower semicontinuous function such that $\beta = \partial j$.

Throughout this section β will be assumed strongly monotone:

$$\beta = \gamma + \eta I \tag{5.8}$$

with γ maximal monotone in $R \times R$ and $\eta > 0$.

Operator M^ε is obtained replacing β by $\beta^\varepsilon = \gamma^\varepsilon + \eta I$ where γ^ε is given by (4.5).

Let us verify conditions (a), (b), (c) of § 3.

LEMMA 5.3. *Let $f_\varepsilon \to f$ weakly in $L^2(0, T; H^1(\Omega))$. Then $y^\varepsilon = \theta_\varepsilon(f_\varepsilon) \to y = \theta(f)$ strongly in $C(0, T; H^1(\Omega))$ and $y_t^\varepsilon \to y_t$ strongly in $C(0, T; L^2(\Omega))$.*

This is a variant of Corollary 5.2.

LEMMA 5.4. *The operator $S \circ \theta_\varepsilon: L^2(0, T; H^1(\Omega)) \to L^2(0, T; L^2(\Omega))$ is Gâteaux differentiable and $r = \nabla(S \circ \theta_\varepsilon)(f)g$ satisfies in a weak sense:*

$$r_{tt} - \Delta r = g \quad \text{a.e. } \Omega \times]0, T[, \tag{5.9}$$

$$r(0, x) = r_t(0, x) = 0 \quad \text{a.e. } \Omega, \tag{5.10}$$

$$-\frac{\partial r}{\partial n} = \nabla\beta^\varepsilon(y_t) \cdot r_t \quad \text{a.e. } \Gamma \times]0, T[\tag{5.11}$$

where $y = \theta_\varepsilon(f)$. Moreover $r \in L^\infty(0, T; H^1(\Omega))$ and $r_t \in L^\infty(0, T; L^2(\Omega))$, $r_{tt} \in L^\infty(0, T; H^1(\Omega)^*)$. The solution to problem (5.9)–(5.11) is unique.

The proof follows the same steps as for Lemma 4.4 (see also the next proposition).

We describe some facts about the adjoint $\nabla(S \circ \theta_\varepsilon)(f)^*: L^2(0, T; L^2(\Omega)) \to L^2(0, T; (H^1(\Omega))^*)$. Denote $\nabla(S \circ \theta_\varepsilon)(f)^* q = p$ where $f \in L^2(0, T; H^1(\Omega))$, $q \in$

$L^2(0, T; L^2(\Omega))$ and $p \in L^2(0, T; H^1(\Omega)^*)$. We have $p = -m_t$, where

$$m_{tt} - \Delta m = \int_t^T q \quad \text{a.e. } \Omega \times]0, T[, \tag{5.12}$$

$$m(T, x) = m_t(T, x) = 0 \quad \text{a.e. } \Omega, \tag{5.13}$$

$$\frac{\partial m}{\partial n} = \nabla \beta^\varepsilon(y_t) \cdot m_t \quad \text{a.e. } \Gamma \times]0, T[. \tag{5.14}$$

For the existence and uniqueness in (5.12)–(5.14) we state:

PROPOSITION 5.5. *Assume β to be a strongly monotone graph in $R \times R$. Then the problem*

$$m_{tt} - \Delta m = w \quad \text{a.e. } \Omega \times]0, T[, \tag{5.12'}$$

$$m(0, x) = m_t(0, x) = 0 \quad \text{a.e. } \Omega, \tag{5.13'}$$

$$-\frac{\partial m}{\partial n} = \nabla \beta^\varepsilon(y_t) \cdot m_t \quad \text{a.e. } \Gamma \times]0, T[, \tag{5.14'}$$

where $w \in L^2(0, T; L^2(\Omega))$, has a unique weak solution $m \in L^\infty(0, T; H^1(\Omega))$ with $m_t \in L^\infty(0, T; L^2(\Omega))$.

Proof. We point out only that the weak solution is well defined as the limit of strong solutions when function $\nabla \beta^\varepsilon(y_t)$ is replaced by more regular mappings such that the Kato-Tanabe [13] existence results for time-dependent evolution problems hold.

Let $f_n, g_n \in C^\infty(Q) \cap L^\infty(Q)$ be such that

$$g_n \to \nabla \beta^\varepsilon(y_t) \quad \text{strongly in } L^\infty(0, T; L^\infty(\Gamma)),$$

$$f_n \to \nabla \beta^\varepsilon(y_t) \quad \text{strongly in } L^\infty(0, T; L^\infty(\Gamma)).$$

We denote the corresponding solutions of (5.12′)–(5.14′) by u^n, v^n. Then

$$\frac{1}{2} \frac{d}{dt} |u_t^n - v_t^n|^2_{L^2(\Omega)} - \int_\Omega \Delta(u^n - v^n)(u_t^n - v_t^n)\, dx = 0;$$

hence, integrating over $[0, t]$ and using Green's formula, we obtain

$$\begin{aligned} &\frac{1}{2} |u_t^n(t) - v_t^n(t)|^2_{L^2(\Omega)} + \frac{1}{2} |\nabla(u^n(t) - v^n(t))|^2_{L^2(\Omega)} \\ &+ \int_0^t \int_\Gamma (g_n u_t^n - f_n v_t^n) \cdot (u_t^n - v_t^n)\, d\Gamma\, d\tau = 0. \end{aligned} \tag{5.15}$$

It is easy to verify that $\{u^n\}$, $\{v^n\}$ are bounded in $L^\infty(0, T; H^1(\Omega))$, $\{u_t^n\}$, $\{v_t^n\}$ are bounded in $L^\infty(0, T; L^2(\Omega))$ and $\{u_t^n|_\Gamma\}$, $\{v_t^n|_\Gamma\}$ are bounded in $L^\infty(0, T; L^2(\Gamma))$. By taking convenient subsequences we have weak convergence towards u (respectively v) in the above spaces. By (4.5) we see that β^ε are strongly monotone, uniformly in ε; therefore one can assume that f_n, g_n are bounded from below by a positive constant η.

The last integral in (5.15) can be written as

$$\int_0^t \int_\Gamma (g_n \cdot u_t^n - f_n \cdot v_t^n)(u_t^n - v_t^n)\, d\Gamma\, d\tau$$

$$= \int_0^t \int_\Gamma g_n(u_t^n - v_t^n)^2 \, d\Gamma \, d\tau + \int_0^t \int_\Gamma (g_n - f_n) v_t^n (u_t^n - v_t^n) \, d\Gamma \, d\tau$$

$$\geqq \int_0^t \int_\Gamma \eta(u_t^n - v_t^n)^2 \, d\Gamma \, d\tau + \int_0^t \int_\Gamma (g_n - f_n) v_t^n (u_t^n - v_t^n) \, d\Gamma \, d\tau.$$

We have also

$$\left| \int_0^t \int_\Gamma (g_n - f_n) v_t^n (u_t^n - v_t^n) \, d\Gamma \, d\tau \right|$$

$$\leqq |g_n - f_n|_{L^\infty} \cdot |v_t^n|_{L^2} \cdot |u_t^n - v_t^n|_{L^2} \leqq C \cdot |g_n - f_n|_{L^\infty}. \tag{5.16}$$

Now, by (5.15), (5.16) and the weak lower semicontinuity of the norm, we conclude $u = v$ and this ends the proof.

LEMMA 5.6. *Denote* $y^\varepsilon = \theta_\varepsilon(f)$ *and* $y = \theta(f)$. *Then*

$$|y^\varepsilon(t) - y(t)|_{H^1(\Omega)} + |y_t^\varepsilon(t) - y_t(t)|_{L^2(\Omega)} \leqq C \cdot \varepsilon^{1/2}.$$

The proof uses estimates similar to (4.17), (4.16).

We can write the system (3.4), (3.5) for the approximate control problem

$$\text{Minimize} \left\{ \int_0^T L^\varepsilon(y, u) + \frac{1}{2} \int_0^T |u - u^*|_U^2 \, dt \right\} \tag{5.17}$$

subject to

$$y_{tt} - \Delta y = Bu \quad \text{a.e. } \Omega \times]0, T[, \tag{5.18}$$

$$y(0, x) = y_0(x), \quad y_t(0, x) = v_0(x) \quad \text{a.e. } \Omega, \tag{5.19}$$

$$-\frac{\partial y}{\partial n} = \beta^\varepsilon(y_t) \quad \text{a.e. } \Gamma \times]0, T[. \tag{5.20}$$

PROPOSITION 5.7. *Let* β *be strongly maximal monotone. The problem* (5.17)–(5.20) *has a solution* $\langle y^\varepsilon, u_\varepsilon \rangle$ *in* $W^{2,2}(0, T; L^2(\Omega)) \times L^2(0, T; U)$ *and there exists* $m^\varepsilon \in C(0, T; L^2(\Omega))$ *such that*:

$$y_{tt}^\varepsilon - \Delta y^\varepsilon = Bu_\varepsilon \quad \text{a.e. } \Omega \times]0, T[, \tag{5.21}$$

$$m_{tt}^\varepsilon - \Delta m^\varepsilon = -\int_t^T q_\varepsilon \quad \text{a.e. } \Omega \times]0, T[, \tag{5.22}$$

$$y^\varepsilon(0, x) = y_0(x), \quad y_t^\varepsilon(0, x) = v_0(x) \quad \text{a.e. } \Omega, \tag{5.23}$$

$$m^\varepsilon(T, x) = m_t^\varepsilon(T, x) = 0 \quad \text{a.e. } \Omega, \tag{5.24}$$

$$-\frac{\partial y^\varepsilon}{\partial n} = \beta^\varepsilon(y_t^\varepsilon) \quad \text{a.e. } \Gamma \times]0, T[, \tag{5.25}$$

$$\frac{\partial m^\varepsilon}{\partial n} = \nabla \beta^\varepsilon(y_t^\varepsilon) \cdot m_t^\varepsilon \quad \text{a.e. } \Gamma \times]0, T[, \tag{5.26}$$

$$\langle q_\varepsilon(t), -B^* m_t^\varepsilon(t) + u_\varepsilon(t) - u^*(t) \rangle = \partial L^\varepsilon(y^\varepsilon(t), u_\varepsilon(t)) \quad \text{a.e. } [0, T]. \tag{5.27}$$

Moreover, we have $y^\varepsilon \to y^*$ *strongly in* $C(0, T; H^1(\Omega))$, $y_t^\varepsilon \to y_t^*$ *strongly in* $C(0, T; L^2(\Omega))$, $u_\varepsilon \to u^*$ *strongly in* $L^2(0, T; U)$, $p^\varepsilon = -m_t^\varepsilon \to p$ *weakly* in* $L^\infty(0, T; L^2(\Omega))$ *and* $q_\varepsilon \to q$ *weakly in* $L^1(0, T; L^2(\Omega))$ *where*:

$$\langle q(t), B^* p(t) \rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T]. \tag{5.28}$$

To pass to the limit in (5.22), (5.26), (5.24) we again have to make the additional assumptions (4.30), (4.32).

THEOREM 5.8. *Let* $\langle y^*, u^*\rangle \in W^{2,2}(0, T; L^2(\Omega)) \times L^2(0, T; U)$ *be an optimal pair for problem* (5.1)–(5.4) *where* β *is a strongly maximal monotone graph in* $R \times R$ *satisfying* (4.30) *and* (4.32). *Then there exist functions* $m \in L^\infty(0, T; H^1(\Omega)) \cap W^{1,\infty}(0, T; L^2(\Omega))$, $q \in L^2(0, T; L^2(\Omega))$ *which satisfy in a weak sense*

$$y^*_{tt} - \Delta y^* = Bu^* \quad \text{a.e. } Q,$$

$$m_{tt} - \Delta m = -\int_t^T q \quad \text{a.e. } Q,$$

$$y^*(0, x) = y_0(x), \qquad y^*_t(0, x) = v_0(x) \quad \text{a.e. } \Omega,$$

$$m(T, x) = m_t(T, x) = 0 \quad \text{a.e. } \Omega,$$

$$-\frac{\partial y^*}{\partial n} \in \beta(y^*_t) \quad \text{a.e. } \Gamma \times]0, T[,$$

$$\frac{\partial m}{\partial n} \in \partial\beta(y^*_t) \cdot m_t \quad \text{a.e. } \Gamma \times]0, T[,$$

$$\langle q(t), -B^* m_t(t)\rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T].$$

Proof. Multiply (5.22) by m^ε_t and use Gronwall's lemma to deduce that $\{m^\varepsilon\}$ is bounded in $L^\infty(0, T; H^1(\Omega))$ and $\{m^\varepsilon_t\}$ is bounded in $L^\infty(0, T; L^2(\Omega))$ as $\{q_\varepsilon\}$ is bounded in $L^1(0, T; L^2(\Omega))$. We also get that $\nabla\beta^\varepsilon(y^\varepsilon_t) \cdot |m^\varepsilon_t|^2$ is bounded in $L^\infty(0, T; L^1(\Gamma))$.

From (4.30) by taking into account the cases $|y^\varepsilon_t(t, x)| \leqq \text{ct.}$, $|y^\varepsilon_t(t, x)| \geqq \text{ct.}$, it is easy to infer that $\{\nabla\beta^\varepsilon(y^\varepsilon_t)\}$ is bounded in $L^2(0, T; L^2(\Gamma))$. A standard procedure involving Young's inequality gives that $\{\nabla\beta^\varepsilon(y^\varepsilon_t) \cdot m^\varepsilon_t\}$ is bounded in $L^\alpha(0, T; L^\alpha(\Gamma))$ for some $\alpha > 1$ (see Theorem 6.2). Therefore $\nabla\beta^\varepsilon(y^\varepsilon_t) \cdot m^\varepsilon_t \to z$ weakly in $L^\alpha(0, T; L^\alpha(\Gamma))$, $\alpha > 1$.

A variant of [6, Lemma 3] shows that

$$\nabla\beta^\varepsilon(y^\varepsilon_t) \to \partial\beta(y^*_t) \quad \text{weakly* in } L^2(0, T; L^2(\Gamma)). \tag{5.29}$$

By (4.34) for every $\eta > 0$ we find $\Gamma_\eta \subset \Gamma$ with mes $(\Gamma - \Gamma\eta) < \eta$ such that $\nabla\beta^\varepsilon(y^\varepsilon_t) \geqq C_\eta > 0$ on Γ_η as β is strongly monotone. Then we obtain that $\{m^\varepsilon_t\}$ is bounded in $L^2(\Gamma_\eta)$.

Next, by using a concave-convex function, we can prove as in § 4 that in a generalized sense:

$$\nabla\beta^\varepsilon(y^\varepsilon_t) \cdot m^\varepsilon_t \to \partial\beta(y^*_t) \cdot m_t$$

weakly in $L^\infty(0, T; L^2(\Gamma_\eta))$, that is,

$$z(t, x) \in \partial\beta(y^*_t(t, x)) \cdot m_t(t, x) \quad \text{a.e. } \Gamma \times]0, T[$$

and the proof is finished.

Remark. A similar result can be stated when β is a differentiable function.

6. Parabolic control problems. We consider the problem

$$\text{Minimize} \int_0^T L(y(t), u(t))\, dt \tag{6.1}$$

subject to

$$y_t - \sum_{i=1}^{N} (a_i(y_{x_i}))_{x_i} = Bu \quad \text{a.e. } \Omega\times]0, T[, \tag{6.2}$$

$$y(0, x) = y_0(x) \quad \text{a.e. } \Omega, \qquad y(t, x) = 0 \quad \text{a.e. } \Gamma\times]0, T[. \tag{6.3}$$

For the functions a_i we assume

$$a_i(y)\cdot y \geqq w|y|^p + C, \qquad w>0, \quad p\geqq 2, \tag{6.4}$$

$$(a_i(y)-a_i(z))(y-z)\geqq \eta|y-z|^2, \qquad \eta>0. \tag{6.5}$$

$$a_i \text{ are locally Lipschitz and } \nabla a_i(y)\leqq C_1|y|^{p-2}+C_2 \quad \text{a.e. } R. \tag{6.6}$$

Remark 1. By (6.6) it is obvious that

$$|a_i(y)|\leqq C_1|y|^{p-1}+C,$$

where C_i denote different constants. Then using (6.4) we obtain existence and uniqueness in problem (6.2), (6.3) when the right-hand side is in $L^{p'}(0, T; W^{-1,p'}(\Omega))$ (see Barbu [3]).

To apply the abstract scheme of § 3 we take the spaces $W = U$; $Z, X, Y = L^2(\Omega)$, $S: L^2(\Omega)\to L^2(\Omega)$ is the identity and $F = B$. The operator $M: L^2(\Omega)\to L^2(\Omega)$ is the realization in $L^2(\Omega)$ of the generalized divergence operator induced by the functions a_i in $W_0^{1,p}(\Omega)\times W^{-1,p'}(\Omega)$. The family of operators M^ε is obtained similarly, replacing a_i by their regularizations:

$$a_i^\varepsilon(y) = \int_{-\infty}^{\infty} a_i(y-\varepsilon\theta)\rho(\theta)\, d\theta \tag{6.7}$$

with $\rho\in C_0^\infty(R)$, $\rho(-\theta)=\rho(\theta)$, $\rho(\theta)\geqq 0$, supp $\rho\subset[-1, 1]$ and $\int_{-\infty}^{\infty}\rho(\theta)\, d\theta = 1$.

An elementary calculation shows that functions a_i^ε verify (6.4)–(6.6) with modified constants, uniformly in ε in a neighbourhood of 0.

We check conditions (a), (b), (c).

LEMMA 6.1. *Let $f_\varepsilon\to f$ weakly in $L^2(0, T; L^2(\Omega))$. Then $\theta_\varepsilon(f_\varepsilon) = y^\varepsilon\to y = \theta(f)$ strongly in $C(0, T; L^2(\Omega))$ as $\varepsilon\to 0$.*

Proof. We have, after multiplication by y^ε:

$$\frac{1}{2}|y^\varepsilon(t)|^2_{L^2(\Omega)} + \sum_{i=1}^{N}\int_0^t\int_\Omega a_i(y_{x_i}^\varepsilon)\cdot y_{x_i}^\varepsilon\, dx\, dt = \frac{1}{2}|y_0|^2_{L^2(\Omega)} + \int_0^t\int_\Omega f_\varepsilon\cdot y^\varepsilon\, dx\, dt.$$

By (6.4) we get that $\{y^\varepsilon\}$ is bounded in $L^\infty(0, T; L^2(\Omega))\cap L^p(0, T; W_0^{1,p}(\Omega))$. From (6.6) we conclude that $\{a_i^\varepsilon(y_{x_i}^\varepsilon)\}$ is bounded in $L^{p'}(0, T; L^{p'}(\Omega))$, that is from (6.2) that $\{y_t^\varepsilon\}$ is bounded in $L^{p'}(0, T; W^{-1,p'}(\Omega))$.

The Aubin theorem gives $y^\varepsilon\to y$ strongly in $L^p(0, T; L^2(\Omega))$.

Now, a standard argument (see Barbu [3, p. 140]) gives $y=\theta(f)$ and $y^\varepsilon\to y$ strongly in $C(0, T; L^2(\Omega))\cap L^2(0, T; H_0^1(\Omega))$.

In order to make condition (b) clearer, we give the following existence result for a parabolic linear partial differential equation:

THEOREM 6.2. *The problem*

$$p_t - \sum_{i=1}^{N}(a_i(x, t)p_{x_i})_{x_i} = q \quad \text{a.e. } \Omega\times]0, T[, \tag{6.8}$$

$$p(t, x) = 0 \quad \text{a.e. } \Gamma\times]0, T[, \tag{6.9}$$

$$p(0, x) = p_0(x) \quad \text{a.e. } \Omega, \tag{6.10}$$

where

$$a_i \in L^\alpha(Q), \qquad \alpha > 1, \tag{6.11}$$

$$a_i \geqq \eta > 0, \quad \text{a.e. } Q \tag{6.12}$$

and $q \in L^2(Q)$, $p_0 \in L^2(\Omega)$ *has a unique solution* $p \in L^2(0, T; H_0^1(\Omega))$ *with* $p_t \in L^\tau(0, T; W^{-1,\tau}(\Omega))$ *with some* $\tau > 1$.

Proof. We take the approximating equation

$$p_t^\varepsilon - \sum_{i=1}^{N} (a_i^\varepsilon(x, t) \cdot p_{x_i}^\varepsilon)_{x_i} = q \quad \text{a.e. } Q, \tag{6.13}$$

where the functions a_i^ε are determined by:

$$a_i^\varepsilon(t, x) = \int_{R^{N+1}} a_i(t - \varepsilon\theta, x - \varepsilon y)\rho(\theta, y)\, d\theta\, dy \tag{6.14}$$

and ρ is a Friedrichs mollifier with support in the unit ball. The functions a_i^ε satisfy (6.12) and

$$a_i^\varepsilon \to a_i \quad \text{as } \varepsilon \to 0, \quad \text{strongly in } L^\alpha(Q), \tag{6.15}$$

$$a_i^\varepsilon \in L^\infty(Q) \cap C^\infty(R^{N+1}). \tag{6.16}$$

It is well known that problem (6.13), (6.9), (6.10) has a unique solution. We multiply (6.13) by p^ε and use Green's formula:

$$\frac{1}{2}\frac{d}{dt}|p^\varepsilon|^2_{L^2(\Omega)} + \sum_{i=1}^{N} \int_\Omega a_i^\varepsilon |p_{x_i}^\varepsilon|^2\, dx = \int_\Omega q \cdot p^\varepsilon\, dx. \tag{6.17}$$

From (6.12), (6.17), by Gronwall's lemma, we infer that $\{p^\varepsilon\}$ is bounded in $L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$, and $\{a_i^\varepsilon \cdot |p_{x_i}^\varepsilon|^2\}$ is bounded in $L^1(Q)$.

We use Young's inequality to estimate $a_i^\varepsilon \cdot p_{x_i}^\varepsilon$:

$$|a_i^\varepsilon \cdot p_{x_i}^\varepsilon| \leqq a_i^\varepsilon(1 + |p_{x_i}^\varepsilon|^{1+\nu}), \qquad \nu > 0,$$

$$a_i^\varepsilon \cdot |p_{x_i}^\varepsilon|^{1+\nu} = (a_i^\varepsilon)^{1-\mu}((a_i^\varepsilon)^\mu \cdot |p_{x_i}^\varepsilon|^{1+\nu})$$

$$\leqq \frac{1}{\delta}(a_i^\varepsilon)^{(1-\mu)\delta} + \frac{1}{\delta'}[(a_i^\varepsilon)^\mu |p_{x_i}^\varepsilon|^{1+\nu}]^{\delta'},$$

where we choose ν, μ, δ, δ' such that

$$1 > \mu > \frac{1}{2}, \quad 1 < (1-\mu)\delta < \alpha, \quad \delta > 2, \quad 1 + \nu = 2\mu, \quad \frac{1}{\delta} + \frac{1}{\delta'} = 1.$$

The above inequalities show that $\{a_i^\varepsilon \cdot p_{x_i}^\varepsilon\}$ is bounded in $L^\tau(Q)$ with some $\tau > 1$ since $\{a_i^\varepsilon\}$ is bounded in $L^\alpha(Q)$. Then $(a_i^\varepsilon \cdot p_{x_i}^\varepsilon)_{x_i}$ is bounded in $L^\tau(0, T; W^{-1,\tau}(\Omega))$ and by (6.13) we see that $\{p_t^\varepsilon\}$ is bounded in $L^\tau(0, T; W^{-1,\tau}(\Omega))$. By the Aubin compactness result we have:

$$p^\varepsilon \to p \quad \text{strongly in } L^2(Q),$$

$$p_{x_i}^\varepsilon \to p_{x_i} \quad \text{weakly in } L^2(Q),$$

$$p_t^\varepsilon \to p_t \quad \text{weakly in } L^\tau(0, T; W^{-1,\tau}(\Omega)),$$

$$a_i^\varepsilon \cdot p_{x_i}^\varepsilon \to h \quad \text{weakly in } L^\tau(Q).$$

From (6.15), by Egorov's theorem, for every $\eta>0$ there is $Q_\eta \subset Q$ such that mes $(Q-Q_\eta)<\eta$ and $a_i^\varepsilon \to a_i$ uniformly on Q_η. Then

$$a_i^\varepsilon \cdot p_{x_i}^\varepsilon \to a_i \cdot p_{x_i} \quad \text{weakly in } L^2(Q_\eta).$$

When $\eta \to 0$ we deduce that $h(t, x)=a_i(t, x)\cdot p_{x_i}(t, x)$ a.e. Q that is p satisfies (6.8)–(6.10).

LEMMA 6.3. *The application θ_ε is Gâteaux differentiable in $L^2(Q)$ and $r=\nabla\theta_\varepsilon(f)g$ satisfies*

$$r_t-\sum_{i=1}^N (\nabla a_i^\varepsilon(y_{x_i})\cdot r_{x_i})_{x_i}=g \quad \text{a.e. } Q, \tag{6.18}$$

$$r(0, x)=0 \quad \text{a.e. } \Omega, \tag{6.19}$$

$$r(t, x)=0 \quad \text{a.e. } \Gamma\times]0, T[, \tag{6.20}$$

where $f, g\in L^2(Q)$ and $y=\theta_\varepsilon(f)$. Moreover $r\in L^\infty(0, T; L^2(\Omega))\cap L^2(0, T; H_0^1(\Omega))$ and $r_t\in L^\tau(0, T; W^{-1,\tau}(\Omega))$ for some $\tau>1$.

For later use we describe the adjoint of $\nabla\theta_\varepsilon(f)$. We denote $\nabla\theta_\varepsilon(f)^*q=p$ where $f, p, q\in L^2(Q)$. Then p is a solution of

$$p_t+\sum_{i=1}^N (\nabla a_i^\varepsilon(y_{x_i})\cdot p_{x_i})_{x_i}=-q \quad \text{a.e. } Q, \tag{6.21}$$

$$p(T, x)=0 \quad \text{a.e. } \Omega, \tag{6.22}$$

$$p(t, x)=0 \quad \text{a.e. } \Gamma\times]0, T[. \tag{6.23}$$

The existence and uniqueness for problem (6.21)–(6.23) is contained in Theorem 6.2.

LEMMA 6.4. *Denote $y^\varepsilon=\theta_\varepsilon(f)$, $y=\theta(f)$. Then*

$$|y^\varepsilon(t)-y(t)|_{L^2(\Omega)}\leqq C\cdot\varepsilon^{1/2} \quad \forall t\in[0, T].$$

The proofs of these statements follow the same lines as in the previous sections and we omit them.

We recall the approximate control problem

$$\text{Minimize}\left\{\int_0^T L^\varepsilon(y, u)\,dt+\frac{1}{2}\int_0^T |u-u^*|_U^2\,dt\right\} \tag{6.24}$$

subject to

$$y_t-\sum_{i=1}^N (a_i^\varepsilon(y_{x_i}))_{x_i}=Bu \quad \text{a.e. } Q, \tag{6.25}$$

$$y(0, x)=y_0(x) \quad \text{a.e. } \Omega, \tag{6.26}$$

$$y(t, x)=0 \quad \text{a.e. } \Gamma\times]0, T[. \tag{6.27}$$

The results of § 3 enable us to state:

PROPOSITION 6.5. *Problem (6.24)–(6.27) has a solution $\langle y^\varepsilon, u_\varepsilon\rangle\in L^2(0, T; L^2(\Omega))\times L^2(0, T; U)$ and for every $\varepsilon>0$ there exists the adjoint arc $p^\varepsilon\in L^2(0, T; L^2(\Omega))$ such that:*

$$y_t^\varepsilon-\sum_{i=1}^N (a_i^\varepsilon(y_{x_i}^\varepsilon))_{x_i}=Bu_\varepsilon \quad \text{a.e. } Q,$$

$$y^\varepsilon(0) = y_0, \qquad y^\varepsilon|_\Gamma = 0,$$

$$p_t^\varepsilon + \sum_{i=1}^{N} [\nabla a_i^\varepsilon(y_{x_i}^\varepsilon) \cdot p_{x_i}^\varepsilon]_{x_i} = q^\varepsilon \quad \text{a.e. } Q, \tag{6.28}$$

$$p^\varepsilon(T, x) = 0 \quad \text{a.e. } \Omega, \tag{6.29}$$

$$p^\varepsilon(t, x) = 0 \quad \text{a.e. } \Gamma \times]0, T[, \tag{6.30}$$

$$\langle q^\varepsilon(t), B^* p^\varepsilon(t) - u_\varepsilon(t) + u^*(t) \rangle = \partial L^\varepsilon(y^\varepsilon(t), u_\varepsilon(t)) \quad \text{a.e. } [0, T]. \tag{6.31}$$

Moreover, we have $y^\varepsilon \to y^*$ *strongly in* $C(0, T; L^2(\Omega))$, $u_\varepsilon \to u^*$ *strongly in* $L^2(0, T; U)$, $p^\varepsilon \to p$ *weakly* in* $L^\infty(0, T; L^2(\Omega))$ *and* $q^\varepsilon \to q$ *weakly in* $L^1(0, T; L^2(\Omega))$ *where*

$$\langle q(t), B^* p(t) \rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T]. \tag{6.32}$$

Remark 2. Obviously $\{y^\varepsilon\}$ is bounded in $L^\infty(0, T; L^2(\Omega)) \cap L^p(0, T; W_0^{1,p}(\Omega))$ and

$$y_{x_i}^\varepsilon \to y_{x_i}^* \quad \text{strongly in } L^2(Q).$$

We give some more estimates. Multiply (6.28) by p^ε and use (6.5) to obtain that $\{p^\varepsilon\}$ is bounded in $L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$. We also deduce that $\{\nabla a_i^\varepsilon(y_{x_i}^\varepsilon) \cdot |p_{x_i}^\varepsilon|^2\}$ is bounded in $L^1(Q)$. Since $\{\nabla a_i^\varepsilon(y_{x_i}^\varepsilon)\}$ is bounded in $L^{p/(p-2)}(Q)$ by (6.6), we get that $\{\nabla a_i^\varepsilon(y_{x_i}^\varepsilon) \cdot p_{x_i}^\varepsilon\}$ is bounded in $L^\tau(Q)$ with some $\tau > 1$ as in the proof of Theorem 6.2. We can see at once that

$$\nabla a_i^\varepsilon(y_{x_i}^\varepsilon) \cdot p_{x_i}^\varepsilon \to h^i \quad \text{weakly in } L^\tau(0, T; L^\tau(\Omega)), \tag{6.33}$$

$$p_t^\varepsilon \to p_t \quad \text{weakly in } L^1(0, T; W^{-1,\tau}(\Omega)). \tag{6.34}$$

A variant of Barbu [6, Lemma 3] gives

$$\begin{aligned} &\nabla a_i^\varepsilon(y_{x_i}^\varepsilon) \to \partial a_i(y_{x_i}^*) \text{ weakly* in } L^\infty(Q_\sigma) \text{ for every } \sigma > 0 \\ &\text{and with } Q_\sigma \subset Q, \text{ mes}\,(Q - Q_\sigma) < \sigma. \end{aligned} \tag{6.35}$$

To identify functions h^i, we have to make the not restrictive assumption:

$$a_i = \xi_i - \iota_i, \qquad i = \overline{1, N} \tag{6.36}$$

where ξ_i, ι_i are real, convex functions.

The same device as in § 4 gives $\nabla a_i^\varepsilon(y_{x_i}^\varepsilon) \cdot p_{x_i}^\varepsilon \to \partial a_i(y_{x_i}^*) \cdot p_{x_i}$ weakly in $L^1(Q_\sigma)$.

We can state:

THEOREM 6.6. *Let* $\langle y^*, u^* \rangle$ *be an optimal pair for problem* (6.1)–(6.3). *Then there exist functions* $p \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$ *and* $q \in L^2(Q)$ *such that, in a weak sense*:

$$y_t^* - \sum_{i=1}^{N} (a_i(y_{x_i}^*))_{x_i} = Bu^* \quad \text{a.e. } Q,$$

$$y^*(0, x) = y_0(x), \; p(T, x) = 0 \quad \text{a.e. } \Omega,$$

$$y^*(t, x) = 0, \; p(t, x) = 0 \quad \text{a.e. } \Gamma \times]0, T[,$$

$$p_t + \sum_{i=1}^{N} (\partial a_i(y_{x_i}^*) p_{x_i})_{x_i} = q \quad \text{a.e. } Q,$$

$$\langle q(t), B^* p(t) \rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T].$$

Remark 3. If L is quadratic, from the maximum principle, we get $u^* \in L^2(0, T; H_0^1(\Omega)) \cap L^\infty(0, T; L^2(\Omega))$, a regularity result for the optimal control.

7. Delay control systems. Consider the problem:

$$\text{(7.1)} \qquad \text{Minimize} \int_0^T L(y, u)\, dt$$

subject to

$$\text{(7.2)} \qquad y'(t) = A(y(t)) + Dy(t-h) + Eu(t) \quad \text{a.e. } [0, T],$$

$$\text{(7.3)} \qquad y(0) = y_0, \qquad y(s) = \varphi(s) \quad \text{a.e. } [-h, 0].$$

We assume that $L: R^N \times R^m \to]-\infty, +\infty]$ is convex, lower semicontinuous, proper, with finite Hamiltonian, D, E are matrices and $y_0 \in R^N$, $\varphi \in L^2(-h, 0; R^N)$; $A: R^N \to R^N$ is Lipschitz.

We take the spaces $Z = R^N \times L^2(-h, 0; R^N)$ $X = R^N$, $W = R^m$, $Y = R^N \times \{0\}$. Operator $M: Z \to Z$ is given by

$$\text{dom } (M) = \{\langle x, \varphi\rangle \in R^N \times H^1(-h, 0; R^N);\ \varphi(0) = x\},$$

$$M(x, \varphi) = \langle A(x) + D\varphi(-h), \dot{\varphi}\rangle$$

where $\dot{\varphi}$ denotes the derivative with respect to $s \in [-h, 0]$. We keep notation φ' for the derivative with respect to $t \in [0, T]$.

Operator F is given by $\langle E, 0\rangle$ and S is the projection $S(y(t), y') = y(t)$. Here the pair $\langle y(t), y'\rangle$ plays the role of y in § 3.

It is known that M is ω-maximal dissipative in Z. Then the state equation (7.2), with condition (7.3) has a unique strong solution such that $y' \in L^2(0, T; R^N)$.

To obtain M^ε we put instead of A, the regularization A^ε:

$$\text{(7.4)} \qquad A^\varepsilon(y) = \int_{R^N} A(y - \varepsilon x)\rho(x)\, dx, \qquad \varepsilon > 0$$

where ρ is a Friedrichs mollifier as in § 6. The family of mappings A^ε are uniformly Lipschitz and the approximate equation has a strong solution. Then θ and θ_ε are well defined.

The approximate control problem is:

$$\text{(7.5)} \qquad \text{Minimize} \left\{ \int_0^T L^\varepsilon(y(t), u(t))\, dt + \frac{1}{2} \int_0^T |u(t) - u^*(t)|^2_{R^m}\, dt \right\}$$

subject to

$$\text{(7.6)} \qquad y'(t) = A^\varepsilon(y(t)) + Dy(t-h) + Eu(t) \quad \text{a.e. } [0, T],$$

$$\text{(7.7)} \qquad y(0) = y_0, \qquad y(s) = \varphi(s) \quad \text{a.e. } [-h, 0].$$

Now we begin to check conditions (a), (b), (c) from § 3. Let $u_\varepsilon \to u$ weakly in $L^2(0, T; R^m)$.

Multiply (7.6) by $y_\varepsilon = \theta_\varepsilon(Eu_\varepsilon)$ to obtain

$$\frac{1}{2}|y_\varepsilon(t)|^2 - \frac{1}{2}|y_0|^2 = \int_0^t A^\varepsilon(y_\varepsilon(t)) y_\varepsilon(t)\, dt + \int_0^t Dy_\varepsilon(t-h) y_\varepsilon(t)\, dt + \int_0^t Eu_\varepsilon(t) \cdot y_\varepsilon(t)\, dt.$$

As A^ε are uniformly Lipschitz with respect of ε, by Gronwall's lemma we get $\{y_\varepsilon\}$ bounded in $C(0, T, R^N)$ and the same for $\{A^\varepsilon(y_\varepsilon)\}$. From (7.6) it follows that $\{y'_\varepsilon\}$ is bounded in $L^2(0, T; R^N)$ so we can easily pass to the limit and prove $y_\varepsilon \to y = \theta(Eu)$ in $C(0, T; R^N)$.

For (b) we denote $\nabla\theta_\varepsilon(w)v = r$ with w, v in $L^2(0, T; R^N)$. It is immediate to check that r exists (θ_ε is differentiable Gâteaux) and satisfies:

$$r'(t) = \nabla A^\varepsilon(\theta_\varepsilon(w))r(t) + Dr(t-h) + v(t), \tag{7.8}$$

$$r(0) = 0, \qquad r(s) = 0 \quad \text{a.e. } [-h, 0]. \tag{7.9}$$

The adjoint operator $\nabla\theta_\varepsilon(w)^* : L^2(0, T; R^N) \to L^2(0, T; R^N)$ is given by $\nabla\theta_\varepsilon(w)^* q = p$ where;

$$-p'(t) = \nabla A^\varepsilon(\theta_\varepsilon(w))^* p(t) + D^* p(t+h) + q(t), \tag{7.10}$$

$$p(T) = 0, \qquad p(s) = 0 \qquad \text{a.e. } [T, T+h]. \tag{7.11}$$

Here D^*, $\nabla A^\varepsilon(\cdot)^*$ denote the transposed matrices.

Condition (c) can be easily obtained with $\delta(\varepsilon) = \varepsilon$:

$$|\theta_\varepsilon(f)(t) - \theta(f)(t)| \leqq C \cdot \varepsilon \quad \forall t \in [-h, T].$$

Then $L^\varepsilon = L_\varepsilon$, the Yosida regularization of function L.

The approximate optimality system provided by § 3 is

$$y'_\varepsilon(t) = A^\varepsilon(y_\varepsilon(t)) + Dy_\varepsilon(t-h) + Eu_\varepsilon(t) \quad \text{a.e. } [0, T], \tag{7.12}$$

$$-p'_\varepsilon(t) = \nabla A^\varepsilon(y_\varepsilon(t))^* p_\varepsilon(t) + D^* p_\varepsilon(t+h) - q_\varepsilon(t) \quad \text{a.e. } [0, T], \tag{7.13}$$

$$y_\varepsilon(0) = y_0, \quad y_\varepsilon(s) = \varphi(s) \quad \text{a.e. } [-h, 0], \tag{7.14}$$

$$p_\varepsilon(T) = 0, \quad p_\varepsilon(s) = 0 \quad \text{a.e. } [T, T+h], \tag{7.15}$$

$$\langle q_\varepsilon(t), B^* p_\varepsilon(t) - u_\varepsilon(t) + u^*(t)\rangle = \partial L_\varepsilon(y_\varepsilon(t), u_\varepsilon(t)) \quad \text{a.e. } [0, T]. \tag{7.16}$$

We know that:

$$y_\varepsilon \to y^* \quad \text{strongly in } C(0, T; R^N),$$

$$y'_\varepsilon \to y^{*\prime} \quad \text{strongly in } L^2(0, T; R^N),$$

$$u_\varepsilon \to u^* \quad \text{strongly in } L^2(0, T; R^m),$$

$$p_\varepsilon \to p \quad \text{weakly* in } L^\infty(0, T; R^N),$$

$$q_\varepsilon \to q \quad \text{weakly in } L^1(0, T; R^N)$$

and

$$\langle q(t), B^* p(t)\rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T]. \tag{7.17}$$

From (7.13) one can see that $p'_\varepsilon \to p'$ weakly in $L^1(0, T; R^N)$ as all the right-hand side terms are weakly convergent. The Helly compactness principle shows in connection with the Lebesgue convergence theorem that $p_\varepsilon \to p$ strongly in $L^2(0, T; R^N)$ on a convenient subsequence.

Since $\nabla A^\varepsilon(y_\varepsilon)^*$ is bounded in $L^\infty(0, T; R^{N\times N})$, we get $\nabla A^\varepsilon(y_\varepsilon)^* \to \partial A(y^*)$ weakly* in $L^\infty(0, T; R^{N\times N})$ by a variant of Barbu [6, Lemma 3].

We can state the theorem:

THEOREM 7.1. *Let $\langle y^*, u^*\rangle$ be an optimal pair for problem* (7.1)–(7.3). *Then there*

exist functions $p \in W^{1,1}(0, T; R^N)$ *and* $q \in L^1(0, T; R^N)$ *such that*:

$$(y^*)'(t) = A(y^*(t)) + Dy^*(t-h) + Eu^*(t) \quad \text{a.e. } [0, T],$$

$$-p'(t) = \partial A(y^*(t))^* \cdot p(t) + D^*p(t+h) - q(t) \quad \text{a.e. } [0, T],$$

$$y^*(0) = y_0, \quad y^*(s) = \varphi(s) \quad \text{a.e. } [-h, 0],$$

$$p(T) = 0, \quad p(s) = 0 \quad \text{a.e. } [T, T+h],$$

$$\langle q(t), B^*p(t)\rangle \in \partial L(y^*(t), u^*(t)) \quad \text{a.e. } [0, T].$$

Acknowledgment. The author wishes to thank the referees for valuable suggestions concerning the final version of the paper.

REFERENCES

[1] J. P. AUBIN, *Un théorème de compacité*, C.R. Acad. Sci. Paris 256 (1963), pp. 5042–5044.

[2] S. AGMON, A. DOUGLIS AND L. NIRENBERG, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[3] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Ed. Acad.-Noordhoff, 1976.

[4] ———, *Necessary conditions for distributed control problems governed by parabolic variational inequalities*, this Journal, 29 (1981), pp. 64–86.

[5] ———, *Necessary conditions for nonconvex distributed control problems governed by elliptic variational inequalities*, J. Math. Anal. Appl., 80 (1981), pp. 566–597.

[6] ———, *Boundary control problems with nonlinear state equation*, this Journal, 20 (1982), pp. 125–143.

[7] ———, *Boundary optimal control of some free boundary problems* in Evolution Equations and Their Applications, F. Kappel and W. Schappacher eds., Pitman, London, 1982.

[8] V. BARBU AND TH. PRECUPANU, *Convexity and optimization in Banach spaces*, Ed. Acad.-Noordhoff, 1978.

[9] H. BREZIS, *Problèmes unilateraux*, J. Math. Pures Appl., 51 (1972), pp. 1–164.

[10] ———, *Operateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, Math. Studies 5, North-Holland, Amsterdam, 1973.

[11] F. H. CLARKE, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[12] B. DACOROGNA, *Weak Continuity and Weak Lower-Semicontinuity of Nonlinear Functionals*, Lecture Notes in Mathematics 922, Springer-Verlag, Berlin, 1982.

[13] T. KATO AND H. TANABE, *On the abstract evolution equation*, Osaka Math. J., 14 (1962), pp. 107–133.

[14] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[15] F. MIGNOT, *Contrôle dans les inéquations variationelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.

[16] F. MIGNOT AND J. P. PUEL, *Optimal control in some variational inequalities* (to appear).

[17] Z. MEIKE AND D. TIBA, *Optimal control for a Stefan problem*, in Analysis and Optimization of Systems, Proceedings, A. Bensoussan and J. L. Lions eds., Lecture Notes in Control and Information Sciences 44, Springer-Verlag, Berlin, 1982, pp. 776–788.

[18] R. T. ROCKAFELLAR, *Monotone operators associated with saddle functions and minimax problems*, Proc. Symp. Pure Math., vol. XVII, F. Browder ed., Amer. Mathematical Society, 1970.

[19] ———, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.

[20] ———, *La théorie des sousgradients et ses applications à l'optimisation*, Les Presses de l'Univ. de Montréal, 1978.

[21] C. SAGUEZ, *Contrôle optimal de systèmes à frontière libre*, Thèse, Univ. de Technologie de Compiegne. 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

009

BOUNDARY CONTROL AND STABILITY OF LINEAR WATER WAVES*

RUSSELL M. REID† AND DAVID L. RUSSELL‡

Abstract. This paper is the first of two parts which study controllability and stabilizability properties of small amplitude waves on a fluid surface. We first derive an evolution equation describing these waves, and then discuss a suitable form of boundary control.

We show that, for a simple domain geometry, the system is null controllable (can be steered to the zero state) only on an infinite time interval. (The second part of this paper extends the result to finite, irregular domains.) We actually construct the Laplace transform of the open loop null control for infinite time, and show that no null control exists for finite time. We give sufficient conditions for convergence of the series describing the control.

Key words. surface waves, small amplitude waves, linear waves, boundary control

1. Introduction. In this paper we examine controllability properties of a fluid in a two-dimensional region Ω. We assume Ω has a boundary in two parts: Γ a fixed boundary (walls and bottom) and S a free surface. We begin with the case where Ω is semi-infinite, (bottomless) with vertical straight sides.

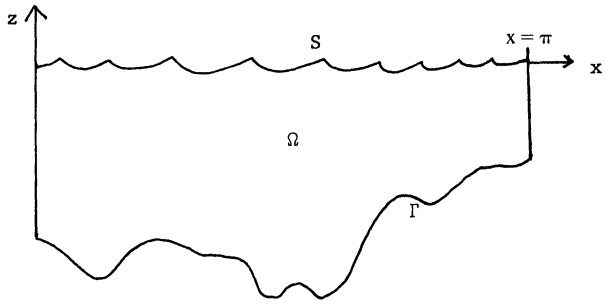


FIG. 1.

We later discuss the case where Ω is a bounded domain, requiring that it have finite arclength and no "beaches"; it must have vertical walls of positive length (see fig. 1).

In § 2 we develop the problem. We show that the system is governed by equations of the form $\dot{x} = A_0 x$ and give existence and uniqueness results for solutions. We then show that a reasonable controlled equation is of the form $\dot{x} = A_0 x + bu$.

In § 3 we look at the controlled system using eigenfunction expansions. We show that the problem of steering the system to zero is equivalent to a moment problem involving the expansion coefficients. The system cannot be steered to zero in finite time.

In § 4 we construct a solution valid in infinite time, and give conditions under which the series describing the formal solution converge.

2. Formulation of the problem. If we assume an irrotational, incompressible, inviscid fluid with constant density, and take a small amplitude linearization of the hydrodynamic equations (dropping second-order terms in amplitude and velocity—see

* Received by the editors November 9, 1982, and in revised form January 25, 1984. This research was supported in part by the Office of Naval Research under grant NR041-404.

† Department of Mathematical and Computer Sciences, Michigan Technological University, Houghton, Michigan 49931.

‡ Department of Mathematics, University of Wisconsin, Madison, Wisconsin 53706.

[1], [7], [20]), then the fluid motion can be expressed in terms of a velocity potential $\phi(x, t)$ defined on $\Omega \times \mathbb{R}$: fluid velocity is $-\text{grad } \phi$. At each moment t, this potential satisfies the Laplace equation in Ω, subject to the boundary conditions

$$\frac{\partial \phi}{\partial n} = 0 \quad \text{on } \Gamma, \qquad -\frac{\partial \phi}{\partial n} = \frac{-\partial \phi}{\partial z} = \frac{\partial \zeta}{\partial t} \quad \text{on the top } z = 0. \tag{2.1}$$

(Here $z = \zeta(\cdot, t)$ is the equation of the actual surface contour, and n is a unit outward normal; it is a result of the linearization process that conditions are imposed at $z = 0$.)

In addition, ϕ must satisfy

$$\frac{\partial \phi}{\partial t} = \zeta \quad \text{on the top } z = 0. \tag{2.2}$$

It is simplest to work with $\Psi = \partial \phi / \partial t$, which is also harmonic in Ω for fixed t. By (2.1), (2.2), Ψ satisfies the boundary conditions

$$\frac{\partial \Psi}{\partial n} = 0 \quad \text{on } \Gamma, \qquad \Psi = \zeta \quad \text{on the top } z = 0, \tag{2.3}$$

so Ψ is well-defined in terms of ζ, and is linear in ζ. In view of (2.1) we want $-\partial \Psi / \partial z|_{\text{top}}$, which will be $\ddot{\zeta}$. We introduce the operator A,

$$A: \zeta \to \left. \frac{\partial \Psi}{\partial z} \right|_{\text{top}} \tag{2.4}$$

where Ψ is harmonic in Ω and satisfies (2.3). Then formally, the evolution equation for ζ is

$$\ddot{\zeta} = -A\zeta. \tag{2.5}$$

A necessary condition for existence of Ψ is $\int_{\text{top}} \zeta(x)\, dx = 0$, so the range of A must lie in the space $L_0^2[0, \pi] = \{f \in L^2[0, \pi]: \int_0^\pi f(x)\, dx = 0\}$. This follows formally from the definition of A, since

$$\int_{\text{top}} A\zeta\, dx = \int_{\text{top}} \frac{\partial \Psi}{\partial n}\, dx = \int_{\partial \Omega} \frac{\partial \Psi}{\partial n}\, dA = \int_\Omega \nabla^2 \Psi \cdot 1\, dA - \int_\Omega \nabla \Psi \cdot \nabla 1\, dA = 0.$$

We complete the definition of A by making it an unbounded operator on $L_0^2[0, \pi]$ with domain $D(A) = H_0^1[0, \pi] \cap L_0^2[0, \pi]$. One can see that A is well-defined from standard results on elliptic operators (see [9]).

LEMMA 2.2. *The operator A is an unbounded, positive, self-adjoint operator with compact resolvent.*

Proof. First we show A is positive. If Ψ_ζ denotes the harmonic function whose trace is ζ, (used to define A above), then

$$\langle A\zeta, \zeta \rangle = \int_S \Psi_\zeta \frac{\partial \Psi_\zeta}{\partial n}\, dl = \int_\Omega \nabla \Psi_\zeta \cdot \nabla \Psi_\zeta\, dA \geqq 0.$$

If $\langle A\zeta, \zeta \rangle = 0$ then $\nabla \Psi = 0$, i.e. Ψ is constant. The requirement $\int_0^\pi \zeta(x)\, dx = 0$ forces that constant to be zero.

Now let $\zeta, \eta \in D(A)$. Integrating by parts, we have

$$\langle A\zeta, \eta \rangle = \int_S \Psi_\eta \frac{\partial \Psi_\zeta}{\partial n}\, dl = \int_{\Gamma \cup S} \Psi_\eta \frac{\partial \Psi_\zeta}{\partial n}\, dl \quad \left(\text{because } \frac{\partial \Psi}{\partial n} = 0 \text{ on } \Gamma \right)$$

$$= \int_{\Gamma \cup S} \frac{\partial \Psi_\eta}{\partial n} \Psi_\zeta\, dl = \langle \zeta, A\eta \rangle.$$

Thus $A^* \supset A$, but the theorem of traces shows that A is a maximal operator on L^2, so A is self-adjoint.

In order to show that A has compact resolvent, we exhibit A^{-1} as an integral operator with weakly singular kernel. Recall that A is defined by the map

$$\zeta \to \Psi_\zeta \to \left.\frac{\partial \Psi_\zeta}{\partial z}\right|_S = A\zeta.$$

Thus $A^{-1}\eta$ will be the boundary values of a harmonic function, specifically that harmonic function whose normal derivative is zero on Γ and η on the surface S. This harmonic function is therefore the single-layer potential for two space dimensions: if $\bar{u} \in \Gamma$, $\bar{x} \in \Omega$, then

$$\Psi(\bar{x}) = \int_\Gamma \eta(\bar{u}) \ln \frac{1}{|\bar{u} - \bar{x}|}\, d\Gamma.$$

This potential is continuous in the plane, and its boundary value is $A^{-1}\eta$. Thus A^{-1} is compact, so the resolvent operator of A is compact. □

We now define energy spaces H_E, H_V either of which may be used in what follows. Denoting by H_0^1 the intersection of H^1 with L_0^2 (see Def. 2.1), we let

$$H_V = H_0^1 \oplus L_0^2, \qquad H_E = L_0^2 \oplus L_0^2,$$

equipped with the inner products

$$\left\langle \begin{pmatrix} \zeta \\ \eta \end{pmatrix}, \begin{pmatrix} z \\ h \end{pmatrix} \right\rangle_V = \langle \eta, h \rangle_{L^2} + \langle \zeta, Az \rangle_{L^2},$$

$$\left\langle \begin{pmatrix} \zeta \\ \eta \end{pmatrix}, \begin{pmatrix} z \\ h \end{pmatrix} \right\rangle_E = \langle \eta, A^{-1}h \rangle_{L^2} + \langle \zeta, z \rangle_{L^2}.$$

Using the self-adjointness of A, it may be easily shown that H_E, H_V are Hilbert spaces, and that the operator

$$A_0 = \begin{pmatrix} 0 & I \\ -A & 0 \end{pmatrix}$$

is skew adjoint on H_V, H_E.

Note that H_E gives the physical energy of the system: $\frac{1}{2}\int_S \zeta^2(x)\, dx$ is the potential energy, and the kinetic energy is

$$\frac{1}{2}\int_\Omega \nabla\phi^2\, d\Omega = \frac{1}{2}\int_\Gamma \phi \frac{\partial \phi}{\partial n}\, d\Gamma = \frac{1}{2}\int_S \left(\frac{\partial \zeta}{\partial t}\right) A^{-1}\left(\frac{\partial \zeta}{\partial t}\right) dx$$

where we recall $\eta = \partial\zeta/\partial t = \partial\phi/\partial n$ on the surface, and since ϕ is harmonic, $A^{-1}(\partial\zeta/\partial t) = A^{-1}(\partial\phi/\partial n) = \phi$ on the surface S.

Both H_E and H_V norms are energies in the sense that they are constants of the motion; for either norm

$$\frac{d}{dt}\left\| \begin{pmatrix} \zeta \\ \eta \end{pmatrix} \right\| = 0,$$

as is easily verified from the dynamic equation (2.5). Henceforth we use these energy spaces as the state spaces. We use H_V, but results and proofs are alike for H_E.

PROPOSITION 2.3. *For initial conditions in H_V, a solution to*

$$\frac{\partial}{\partial t}\begin{pmatrix}\zeta\\ \eta\end{pmatrix}=\begin{pmatrix}0 & I\\ -A & 0\end{pmatrix}\begin{pmatrix}\zeta\\ \eta\end{pmatrix}=A_0\begin{pmatrix}\zeta\\ \eta\end{pmatrix}$$

exists and is unique in H_V on any time interval. If the mapping from initial state z_0 to final state z is denoted by $S(t)$, then $S(t)$ is a strongly continuous semigroup of bounded operators in H_V.

Proof. If A_0 represents the matrix operator, V-skew adjointness of A_0 gives

$$2\|(\lambda - A_0)x\|_V\|x\|_V \geqq \langle(\lambda I - A_0)x, x\rangle_V + \langle x, (\lambda I - A_0)x\rangle_V = 2\operatorname{Re}\lambda\,\|x\|_V^2$$

for $x \in D(A_0)$. This shows that for $\operatorname{Re}(\lambda) > 0$

$$\|(\lambda I - A_0)^{-1}\|_V \leqq \frac{1}{\operatorname{Re}(\lambda)}.$$

So we may use the Hille–Yosida theorem. Existence, uniqueness, and regularity of the solution $S(t)z_0$ follow from standard semigroups results. □

Now consider a controlled wave system of the form

$$\dot{w} = A_0 w + Bu \tag{2.6}$$

with w now representing an element of H_V (or H_E), where B is taken to be a fixed element of H_V. For brevity in what follows we shall just state that a parallel development holds for H_E.

Equation (2.6) reflects a distributed control whose physical form is difficult to visualize. Boundary control, which is easier to implement, may be reduced to the same form and an equivalent moment problem.

Suppose we can control a small-amplitude movement of one end of our tank, the one whose nominal equation is $x = \pi$. Assume the spatial distribution of control is fixed as a function of z; denote it by $F(z)$. We require that the integral of $F(z)$ over the control surface be zero—to ensure conservation of volume—and that $F(z)$ be sufficiently regular that solutions remain in H_V.

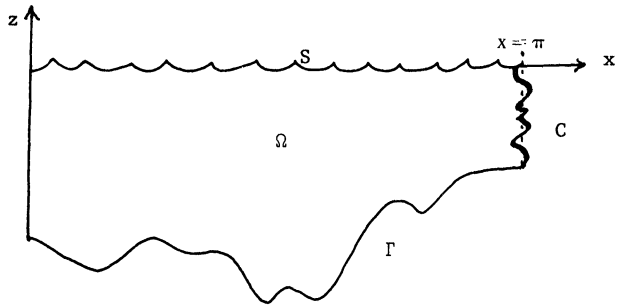


FIG. 2.

Suppose $F(z) \in H^1$ locally, and denote the controlled boundary by C. Since (See (2.2), (2.4))

$$A\zeta = \frac{\partial \eta}{\partial t} = -\frac{\partial}{\partial z}\left(\frac{\partial \Phi}{\partial t}\right),$$

we proceed by finding the harmonic function $\Psi \equiv \partial\Phi/\partial t$. It clearly must satisfy

$$\Psi(x, 0) = \zeta(x), \qquad 0 \leqq x \leqq \pi,$$

$$\frac{\partial \Psi}{\partial n} = 0 \quad \text{on } \Gamma,$$

$$\frac{\partial \Psi}{\partial x} = F(z)U(t) \quad \text{on } C.$$

It is helpful to split Ψ into a sum of two parts; the function Ψ_ζ used to define the operator A, and a potential Θ related only to the control function. In particular there is a harmonic function Θ which satisfies

$$\Theta = 0 \qquad \text{on } S,$$

$$\frac{\partial \Theta}{\partial n} = 0 \qquad \text{on } \Gamma,$$

$$\frac{\partial \Theta}{\partial x} = F(z) \quad \text{on } C.$$

We remark on two properties of Θ: it does not depend on ζ, and it is linear on C in the following sense: If we multiply the boundary condition on C by $U(t)$ for fixed t, i.e.

$$\frac{\partial \Theta}{\partial x} = F(z)U(t) \quad \text{on } C,$$

then the potential satisfying the new boundary conditions is the multiple: $U(t)\Theta$.

The sum $\Psi_\zeta + U(t)\Theta$ is thus the harmonic function Ψ we were seeking because:

$$\Psi_\zeta + U(t)\Theta = \zeta \quad \text{on } S,$$

$$\frac{\partial}{\partial n}(\Psi_\zeta + U(t)\Theta) = 0 \quad \text{on } \Gamma \quad \text{all } t,$$

$$\frac{\partial}{\partial x}(\Psi_\zeta + U(t)\Theta) = F(z)U(t) \quad \text{on } C.$$

We see then that

$$\frac{\partial}{\partial z}(\Psi_\zeta + U(t)\Theta) = A\zeta + U(t)\frac{\partial \Theta}{\partial z},$$

so that our controlled system becomes

$$\frac{\partial}{\partial t}\begin{pmatrix} \zeta \\ \eta \end{pmatrix} = \begin{pmatrix} 0 & I \\ -A & 0 \end{pmatrix}\begin{pmatrix} \zeta \\ \eta \end{pmatrix} + \begin{pmatrix} 0 \\ d(x) \end{pmatrix} U(t), \tag{2.7}$$

where

$$d(x) = \frac{\partial \Theta}{\partial z}(x, 0).$$

We have now included boundary control in the framework of the inhomogenous equation (2.6), except that $b(x) \equiv 0$ for the boundary control. It is interesting to note that the control function is in fact the local acceleration of the boundary wall or membrane.

3. An equivalent moment problem. Since A is self-adjoint with compact resolvent, we know it has a complete orthonormal system of eigenfunctions $\phi_k(x)$ in $L_0^2[0, \pi]$, with associated eigenvalues λ_k. For a particular, very simple geometry we know the λ_k and ϕ_k explicitly: If the tank is infinite depth, with straight sides $x=0$ and $x=\pi$, then one can see that A has eigenvalues $\lambda_k = k$, $k = 1, 2, 3, \cdots$ and eigenvectors $\phi_k = \sqrt{2/\pi} \cos kx$, because $\cos kx\, e^{kz}$ is the potential $\psi(x, z)$ referred to in equation 2.4, i.e. $\psi(x, z)$ has zero normal derivative at $x=0$, $x=\pi$, $z=-\infty$. In this particular case, A^2 is the Sturm–Liouville operator $-d^2/d_{x^2}$, restricted so first derivatives vanish at $x=0$, $x=\pi$. We can separate variables in the controlled system

$$\frac{\partial}{\partial t}\begin{pmatrix}\zeta\\ \eta\end{pmatrix} = \begin{bmatrix}0 & I\\ -A & 0\end{bmatrix}\begin{pmatrix}\zeta\\ \eta\end{pmatrix} + \begin{pmatrix}b\\ d\end{pmatrix}u \tag{3.1}$$

by expanding in terms of the $\phi_k(x)$, giving

$$\begin{aligned}
\zeta(x, t) &= \sum_1^\infty \zeta_k(t)\phi_k(x),\\
\eta(x, t) &= \sum_1^\infty \eta_k(t)\phi_k(x),\\
b(x) &= \sum_1^\infty b_k\phi_k(x),\\
d(x) &= \sum_1^\infty d_k\phi_k(x).
\end{aligned} \tag{3.2}$$

Substituting (3.2) into (3.1), we find that for each k,

$$\frac{\partial}{\partial t}\begin{pmatrix}\zeta_k\\ \eta_k\end{pmatrix} = \begin{bmatrix}0 & 1\\ -\lambda_k & 0\end{bmatrix}\begin{pmatrix}\zeta_k\\ \eta_k\end{pmatrix} + \begin{pmatrix}b_k\\ d_k\end{pmatrix}u(t). \tag{3.3}$$

After the substitution

$$\begin{pmatrix}\zeta_k\\ \eta_k\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ iw_k & -iw_k\end{bmatrix}\begin{pmatrix}x_k\\ y_k\end{pmatrix},$$

where $w_k = \sqrt{\lambda_k}$, then (3.3) may be diagonalized to give the system

$$\frac{\partial}{\partial t}\begin{pmatrix}x_k\\ y_k\end{pmatrix} = \begin{bmatrix}iw_k & 0\\ 0 & -iw_k\end{bmatrix}\begin{pmatrix}x_k\\ y_k\end{pmatrix} + \begin{pmatrix}\gamma_k\\ \delta_k\end{pmatrix}u(t), \tag{3.4}$$

where the control distribution element is now

$$\begin{pmatrix}\gamma_k\\ \delta_k\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix}1 & \dfrac{1}{iw_k}\\ 1 & \dfrac{-1}{iw_k}\end{bmatrix}\begin{pmatrix}b_k\\ d_k\end{pmatrix}. \tag{3.5}$$

Note here that in the boundary control case $b_k = 0$ for all k, but that γ_k and δ_k are nonzero provide d_k is nonzero.

Let us consider finite time null controllability for this system. From (2.6), we see that an initial state ξ_0 is steered to zero in the interval $[0, T]$ provided, with $S(t)$ the semigroup generated by A_0,

$$\|S(T)\xi_0 + \int_0^T S(T-s)Bu(s)\, ds\|_V = 0. \tag{3.6}$$

In order that this norm be zero, each term in the eigenfunction expansion must be zero. Expressing

$$S(T)\xi_0 = \begin{pmatrix} x_0(T) \\ y_0(T) \end{pmatrix}$$

in a series like 3.2, we have

$$\begin{aligned} x_{0,k} + \int_0^T e^{-iw_k s}\gamma_k u(s)\, ds = 0, \\ y_{0,k} + \int_0^T e^{+iw_k s}\delta_k u(s)\, ds = 0 \quad \text{for } k = 1, 2, 3, \cdots. \end{aligned} \tag{3.7}$$

If γ_k and δ_k are not zero, we may consider this a moment problem whose solution gives the control u:

$$\begin{aligned} \int_0^T e^{-iw_k s}u(s)\, ds = \frac{-x_{0,k}}{\gamma_k} \equiv c_k, \\ \int_0^T e^{+iw_k s}u(s)\, ds = \frac{-y_{0,k}}{\delta_k} \equiv d_k. \end{aligned} \tag{3.8}$$

Now let us restrict our attention to the simple case $w_k = \sqrt{k}$ of the infinite-depth tank.

For $T < \infty$ this set of conditions on u does not, in general, have a solution [8], since the w_k have infinite density on the real axis. Physically, this reflects the fact that wave propagation speed nears zero as the wave length nears zero for these (nonviscous) waves. On an infinite interval ($T = \infty$), however, we shall show there is a $u \in L^1 \cap L^2[0, \infty]$ for which the conditions (3.8) hold.

In that case, u will be defined to be a null control provided:

$$\lim_{T\to\infty} \| S(T)\xi_0 + \int_0^T S(T-s)Bu(s)\, ds \|_V = 0. \tag{3.9}$$

(Note that this is a notion similar to asymptotic stability but, rather than being effected by feedback, it is accomplished via the open loop control u.)

The limit is well-defined provided $u \in L^1[0, \infty)$. It will be zero provided the chosen $u \in L^2[0, \infty)$ is a solution of (3.8) with $T = \infty$, i.e.

$$\begin{aligned} \int_0^\infty e^{-iw_k s}u(s)\, ds = \frac{-x_{0,k}}{\gamma_k}, \\ \int_0^\infty e^{+iw_k s}u(s)\, ds = \frac{-y_{0,k}}{\delta_k}. \end{aligned} \tag{3.10}$$

4. Construction of a solution to the moment problem. We shall construct the required function u by finding a set of functions $p_k(s)$ and $q_k(s)$ which are biorthogonal to the exponentials, i.e. which satisfy

$$\begin{aligned} \int_0^\infty e^{iw_k s}\, p_n(s)\, ds = \delta_n^k, \qquad \int_0^\infty e^{iw_k s}\, q_n(s)\, ds = 0, \\ \int_0^\infty e^{-iw_k s}\, p_n(s)\, ds = 0, \qquad \int_0^\infty e^{-iw_k s}\, q_n(s)\, ds = \delta_n^k. \end{aligned} \tag{4.1}$$

Using the functions $p_k(s)$, $q_k(s)$,

$$u(s)=\sum_1^\infty (c_n p_n(s)+d_n q_n(s)) \tag{4.2}$$

will be a formal solution to (3.10); its convergence will be considered separately.

To find the required functions $p_k(s)$, $q_k(s)$, we will construct their Laplace transforms $\hat{p}_n(z)$, $\hat{q}_n(z)$. The properties (4.1) are equivalent to

$$\begin{aligned} &\hat{p}_n(iw_k)=\delta_k^n, \qquad \hat{q}_n(-iw_k)=\delta_k^n, \\ &\hat{p}_n(-iw_k)=\hat{q}_n(+iw_k)=0, \end{aligned} \tag{4.3}$$

for the transforms $\hat{p}_k$, $\hat{q}_k$. We begin by constructing a function $G(z)$ which has simple zeros at $\pm iw_k$ and is bounded in the right half-plane. The $\hat{p}_k(z)$ and $\hat{q}_k(z)$ are then formed by factoring out $(z-iw_k)$ or $(z+iw_k)$, thereby removing one zero at a time.

LEMMA 4.4. *The function of a complex variable z defined by*

$$G(z)=\frac{\Gamma((z+1)^2)}{\Gamma(z^2)[e^{z+1}\Gamma(z+1)]^4} \tag{4.4}$$

is analytic for $\mathrm{Re}(z)\geq -1$ *and uniformly bounded in the right half-plane.* $G(z)$ *has simple zeros at* $z=\pm i\sqrt{k}$ *for* $k=1, 2, 3, \cdots$. *On the imaginary axis, the values of* $G(iw)$ *and its derivatives are asymptotic, as* $w\to\infty$, *to values and derivatives of*

$$G(iw)\sim\frac{1}{4\pi^2 e}\left[1-e^{-2\pi iw^2}\right]. \tag{4.5}$$

Proof. Analyticity of $G(z)$ is clear from the fact that $\Gamma(z)$ is analytic in the right half-plane, and $1/\Gamma(z)$ is entire. G has simple zeros where $1/\Gamma(z^2)$ has simple zeros, i.e. $\pm i\sqrt{k}$. ($G(z)$, of course, has other zeros—but not in the right half-plane.) For an asymptotic estimate of $G(z)$, use Stirling's formula

$$\Gamma(z)=e^{-z}z^{(z-1/2)}(2\pi)^{1/2}(1+O(1/|z|))$$

which is valid as $|z|\to\infty$ in a sector $|\arg z|<\pi$.

We desire an estimate of $G(z)$ in the right half-plane $-\pi/2\leqq\arg z\leqq\pi/2$. First consider the open half-plane $-\pi/2<\arg z<\pi/2$. Then z^2, $(z+1)$, and $(z+1)^2$ are all in the sector where Stirling's estimate is valid, i.e. $|\arg(z^2)|<\pi$, $|\arg(z+1)|<\pi$, $|\arg(z+1)^2|<\pi$. Expanding $G(z)$, we find

$$G(z)=\frac{1}{4\pi^2}e^{-2z-1}\frac{(z+1)^{2z^2-1}}{z^{2z^2-1}}\left[+O\left(\frac{1}{|z|}\right)\right]=\frac{1}{4\pi^2 e}\left(1+O\left(\frac{1}{|z|}\right)\right). \tag{4.6}$$

On the imaginary axis $|\arg z|=\pi/2$, the terms $\Gamma(z+1)$ and $\Gamma\{(z+1)^2\}$ may be estimated as before. Stirling's estimate of $\Gamma(z^2)$ is no longer valid, so we use the reflection principle

$$\frac{1}{\Gamma(z)}\cdot\frac{1}{\Gamma(-z)}=\frac{-z\sin\pi z}{\pi}.$$

Applied to $1/\Gamma(z^2)$, this gives

$$\frac{1}{\Gamma(z^2)}=\frac{-z^2\sin(\pi z^2)}{\pi}\Gamma(-z^2).$$

On the imaginary axis $|\arg z| = \pi/2$, $\arg(-z^2) = 0$ and hence Stirling's formula is valid for $\Gamma(-z^2)$. Expanding as in (4.6), we find for $\pi/2 - \varepsilon \leqq |\arg z| \leqq \pi/2$:

$$G(z) = \frac{1}{4\pi^2} \frac{[(z+1)^2]^{(z^2-1/2)}}{(-z^2)^{(z^2-1/2)}} e^{-2z-1}(2 \sin \pi z^2)\left(1 + O\left(\frac{1}{|z|}\right)\right)$$

$$= \frac{1}{4\pi^2 e}(1 - e^{2\pi i z^2})\left(1 + O\left(\frac{1}{|z|}\right)\right).$$

Written with $z = iw$, this is

$$G(w) = \frac{1}{4\pi^2 e}(1 - e^{-2\pi i w^2})\left(1 + O\left(\frac{1}{|w|}\right)\right)$$

as $|w| \to \infty$ in the sectors noted. Similarly,

$$G'(iw) \sim \frac{-1}{4\pi^2 e} \frac{d}{dw}(e^{-2\pi i w^2}) \tag{4.7}$$

in the same sector.

Combine these results to show that $G(z)$ is bounded in the right half-plane: Let R_1 be large enough that $|G(z)| < 2/4\pi^2 e = M$ if $|z| > R_1$ and $|\arg z| < \pi/2$, and R_2 be large enough that $|G(z)| < 2/4\pi^2 e$ whenever $|\arg z| = \pi/2$ and $|z| > R_2$. Then $G(z)$ is bounded by the larger of M and $\sup |G(z)|$ for $z \leqq R$ in the half-plane.

We now use the function G to construct the Laplace transforms of the desired function p_k, q_k. This is the content of the following lemma.

LEMMA 4.5. *The functions* $\hat{p}_k(z)$, $\hat{q}_k(z)$ *defined by*

$$\hat{p}_k(z) = \frac{i\sqrt{k}\, G(z)}{z(z - i\sqrt{k})G'(i\sqrt{k})}, \qquad \hat{q}_k(z) = \frac{-i\sqrt{k}\, G(z)}{z(z + i\sqrt{k})G'(-i\sqrt{k})}, \tag{4.8}$$

are Laplace transforms of functions in $L^1[0, \infty) \cap L^2[0, \infty)$, *and satisfy the biorthogonality properties* (4.3).

Proof. Since $G(z)$ is uniformly bounded in the right half-plane, and, using 4.7, the quotient

$$\frac{G(x+iy)}{(x+iy)(x+iy-i\sqrt{k})} \cdot \frac{\sqrt{k}}{G'(i\sqrt{k})}$$

is in $L^2(-\infty, \infty)$ as a function of y, uniformly in $x > 0$, with estimates independent of k. Results in Paley and Wiener [10] show that therefore $\hat{p}_k(z)$ and $\hat{q}_k(z)$ are Laplace transforms of functions $p_k(t)$, $q_k(t)$ in $L^2[0, \infty)$. The biorthogonality property (4.3) follows immediately from knowledge of the zeros of $G(z)$ and properties of the Laplace transform.

To show that $p_k(t)$ and $q_k(t)$ are in $L^1 \cap L^2$ and not just in L^2, we note from the asymptotic exapnsion of $G(iw)$ that the derivatives with respect to y of $\hat{p}_k(iy)$ and $\hat{q}_k(iy)$ are in $L^2(-\infty, \infty)$. This forces $tp_k(t)$ and $tq_k(t)$ to be in $L^2[0, \infty)$, and hence $p_k(t)$ and $q_k(t)$ are in $L^1 \cap L^2[0, \infty)$ as desired. □

We may now use the functions $p_k(t)$, $q_k(t)$ to construct a null control for an arbitrary initial state.

Suppose that the initial state ξ_0 is written

$$\begin{pmatrix} \zeta_0 \\ \eta_0 \end{pmatrix} = \sum_1^\infty \alpha_k \Psi_k + \beta_k \Psi_{-k} \tag{4.9}$$

where

$$\Psi_k = \begin{pmatrix} \phi_k \\ iw_k\phi_k \end{pmatrix} \quad \text{and} \quad \Psi_{-k} = \begin{pmatrix} \phi_k \\ -iw_k\phi_k \end{pmatrix}$$

are the eigenvectors of the operator A_0 in the x, y coordinate system. Then

$$u(t) = -\sum_1^\infty \frac{\alpha_k}{\gamma_k} q_k(t) + \frac{\beta_k}{\delta_k} p_k(t) \tag{4.10}$$

is a formal solution to the null control problem, as we can see by calculating the kth condition in (3.10). We find, using (4.1), that

$$\int_0^\infty e^{-iw_k s} u(s)\, ds = \frac{-\alpha_k}{\gamma_k} = \frac{-x_{0,k}}{\gamma_k}$$

and

$$\int_0^\infty e^{+iw_k s} u(s)\, ds = \frac{-\beta_k}{\gamma_k} = \frac{-y_{0,k}}{\delta_k}$$

as required in (3.10).

It remains to show that under suitable conditions, the series (4.10) converges to an admissible control, i.e. one which is in $L^1[0,\infty) \cap L^2[0,\infty)$. We content ourselves here with some sufficient conditions for the convergence of (4.10). We do this by showing that

$$\sum_{k=1}^\infty \frac{|\alpha_k|}{|\gamma_k|} \|q_k(t)\| < \infty \tag{4.11}$$

for both the L^1 and L^2 norms on q_k, and that a similar result holds for the second term in the series (4.10).

Recall that the coefficients γ_k and δ_k depend on our choice of control distribution element.

Let us consider boundary control. It is reasonable to suppose that the control distribution function was chosen to get as strong a control action as possible, which means choosing $d(x)$ in equations (2.7), (3.2), so that the γ_k, δ_k have as slow a decay rate as possible (within the constraint that the d_k be square summable.) For the boundary control, this means $\sqrt{k}\,\gamma_k$ and $\sqrt{k}\,\delta_k$ must be square summable (see e.g. (3.7)). If we suppose that

$$|\gamma_k| \geqq \frac{1}{\sqrt{k}\ln k}, \qquad |\delta_k| \geqq \frac{1}{\sqrt{k}\ln k},$$

then an initial state (4.9) may be steered to zero provided its Fourier coefficients α_k, β_k satisfy

$$|\alpha_k| < \frac{1}{k\sqrt{k}\,(\ln k)^3}$$

with the same estimate for the β_k so that (4.11) holds. This is certainly true if, for example, the coefficient series $\{\alpha_k\}$, $\{\beta_k\}$ are dominated by any p-series with $p > \frac{3}{2}$ (a smoothness condition on the initial state.)

REFERENCES

[1] C. A. COULSON, *Waves: a Mathematical Account of the Common Types of Wave Motion*, Oliver and Boyd, Edinburgh and London, 1966.

[2] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Vol.* I, *General Theory*, Interscience, New York, 1958.

[3] H. O. FATTORINI AND D. L. RUSSELL, *Exact controllability theorems for linear parabolic equations in one space dimension*, Arch. Rat. Mech. Anal., 43 (1971), pp. 272–292.

[4] H. O. FATTORINI, *On complete control of linear systems*, J. Differential Equations, 3 (1967), pp. 391–402.

[5] A. FRIEDMAN AND M. SHINBROT, *The initial value problem for linearized equations of water waves*, J. Math. Mech., 17 (1967), pp. 197–180.

[6] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.

[7] H. LAMB, *Hydrodynamics*, Dover, New York, 1945.

[8] N. LEVINSON, *Gap and density theorems*, Colloquium Publications 26, American Mathematical Society, Providence RI, 1940.

[9] J. L. LIONS AND L. MAGENES, *Inhomogeneous Linear Boundary Value Problems and Applications*, Dunod, Paris, 1968.

[10] R. E. A. C. PALEY AND N. WIENER, *The Fourier transform in the complex domain*, Colloquium Publications, 19, American Mathematical Society, Providence, RI, 1934.

[11] 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. 7–127.

[12] F. RIESZ AND B. SZ-NAGY, *Functional Analysis*, Frederick Ungar C., New York, 1955 (translated from the 2nd French edition by Leo F. Boron).

[13] J. C. ROGERS, *Water waves: uniqueness and continuous dependence on the data*, Technical Report, Johns Hopkins Univ. Applied Physics Laboratory, 1975.

[14] D. L. RUSSELL, *Linear stabilization of the linear oscillator in Hilbert space*, J. Math. Anal. Appl., 25 (1969), pp. 663–675.

[15] ———, *Boundary value control theory of the higher-dimensional wave equation, part* II, this Journal, 9 (1971), pp. 409–419.

[16] ———, *Exact Boundary Value Controllability Theorems for Wave and Heat Processes in Star-Complemented Regions*, Marcel Dekker, New York, 1974.

[17] ———, *Nonharmonic Fourier series in the control of distributed parameter systems*, J. Math. Anal. Appl., 18 (1977), pp. 542–559.

[18] D. L. RUSSELL AND R. M. REID, *Water waves and problems of infinite time control*, in Proc. International Symposium on Analysis and Optimization of Systems, Seminaries IRIA, Rocquencourt, December, 1978.

[19] D. L. RUSSELL, *Canonical forms and spectral determination for a class of hyperbolic distributed parameter control systems*, J. Math. Anal. Appl., 62, (1978), pp. 186–225.

[20] J. J. STOKER, *Water Waves*, Interscience, New York, 1957.

[21] R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383–403.

[22] E. T. WHITTAKER AND G. N. WATSON, *A Course in Modern Analysis*, Cambridge Univ. Press, London, 1920.

[23] D. V. WIDDER, *The Laplace Transform*, Princeton Univ. Press, Princeton, NJ, 1941.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

010

OPTIMAL REPLACEMENT OF ONE-UNIT SYSTEMS UNDER PERIODIC INSPECTION*

SÜLEYMAN ÖZEKICI†

Abstract. In this paper the optimal replacement problem of a one-unit system under periodic inspection is analyzed utilizing Markov decision theory. The problem is formulated in a general setting and it is shown that an age replacement policy is in fact optimal if the component lifetime has monotone failure rate. A necessary and sufficient condition of optimality together with several characterizations of the optimal policy are presented.

Key words. optimal replacement, Markov decision theory, age replacement, periodic inspection

1. Introduction. This paper deals with a classical optimization problem concerning the optimal management of nonrepairable reliability systems. A one-unit system under periodic inspection is considered and the optimal replacement policy is characterized using Markov decision theory. It is shown that an age replacement policy is in fact optimal if the component lifetime has monotone failure rate. In this section we will state and formulate the problem as a Markov decision chain, and in § 2 a necessary and sufficient condition of optimality will be presented. Furthermore, implications of this optimality condition will be analyzed in § 3.

The optimal replacement problem for nonrepairable reliability systems is analyzed by many researchers who have generally constructed their models based on the pioneering papers by Barlow and Proschan [1], [2]. The models constructed vary basically either in the cost structure or in the replacement or preventive maintenance structure. For a review of the literature on the subject we refer our reader to Nakagawa [6] who analyzed various cases by making changes in the model characteristics for both replacement and preventive maintenance problems. Most of the literature on replacement models assume that the item under inspection is replaced upon failure or at a certain age, whichever comes first. Identification of the replacement times as renewal points in time leads to an expression representing the average expected cost through a renewal theoretic argument. The optimal solution thus obtained is optimal within the class of age replacement policies. As we shall see shortly, we will make a general formulation of the replacement problem that will enable us to utilize Markov decision theory.

Consider an item which is inspected periodically at prespecified times $\{0, t_0, 2t_0, 3t_0, \cdots\}$ where t_0 is some positive number. At every inspection a decision is made to replace the item or not. If the item inspected is found to be in a failed state it is replaced immediately by an identical one and a failure cost C_f is incurred. On the other hand, if a functioning item is replaced, a cost of C_p is incurred. We assume that all replacements are immediate and the lifetimes of the successive items installed are independent and identically distributed random variables with common distribution F, and $C_f \geqq C_p \geqq 0$. To avoid some technical difficulties in our presentation we assume that $F(0)=0$ and $F(x)<1$ for all $x \geqq 0$. The periodic discount rate is α $(0<\alpha<1)$ and the problem is to find the optimal replacement policy that minimizes the expected total discounted cost.

* Received by the editors April 14, 1983, and in revised form January 31, 1984.
† Department of Industrial Engineering, Boğaziçi University, Bebek, Istanbul, Turkey.

At each inspection epoch the inspector is to make a decision to replace the item or not. Thus, letting X_n denote the state of the item inspected at time nt_0 for any integer $n \geqq 0$, the decision made should be a function of X_n. We assume that changes in physical and technical performance specifications of the item due to deterioration or aging are undetectable, so what the inspector observes is an item which is either functioning or not functioning. This enables us to let $X = \{X_n; n = 0, 1, \cdots\}$ be the age process associated with the reliability model under study. In other words, if the item inspected at time nt_0 is functioning, then X_n is the total number of inspection periods that item has survived. Otherwise, the item has failed prior to inspection and X_n is set to be equal to some state Δ which denotes failure. Without further mention, we will use discrete time n and nt_0 interchangeably and call X_n the age of the item in use at time n. Letting E be the set of all nonnegative integers and $E_\Delta = E \cup \{\Delta\}$, it is clear that X is a stochastic process with state space E_Δ.

Since the replacement decision is made by observing the present state of the item, we define the set of all admissible policies M by

$$M = \{r: E_\Delta \to \{0, 1\}; r(\Delta) = 1\}$$

where $r(x) = 1$ (resp. $r(x) = 0$) implies that an item observed to be in state $x \in E_\Delta$ is replaced (resp. not replaced) if policy r is used. The following result shows that the problem formulated above is a Markov decision problem.

THEOREM 1.1. *For any policy $r \in M$, $X = \{X_n; n = 0, 1, \cdots\}$ is a Markov chain with state space E_Δ and transition matrix*

$$P_r(i, j) = \begin{cases} p_0 r(i), & j = 1, \\ (1 - p_i)(1 - r(i)) + (1 - p_0) r(i), & j = \Delta, \\ p_i(1 - r(i)), & j = i + 1 \end{cases} \tag{1.2}$$

where

$$p_i = (1 - F((i+1)t_0))/(1 - F(it_0)) \tag{1.3}$$

for all $i \in E$ and $p_\Delta = 0$.

Proof. The fact that X is a Markov chain follows trivially since the successive items installed have independent and identically distributed lifetimes, and the age of the item at each inspection provides all the information about the past of that particular item. For any nonnegative integer n and $i \in E_\Delta$, letting L_n denote the lifetime of the item inspected at time n, it is clear that

$$\begin{aligned} P_r(i, \Delta) &= P\{X_{n+1} = \Delta | X_n = i\} \\ &= \begin{cases} P\{L_n \leqq (i+1)t_0 | L_n > it_0)\} & \text{if } r(i) = 0, \\ P\{L_{n+1} \leqq t_0 | L_n > it_0\} & \text{if } r(i) = 1 \end{cases} \\ &= \begin{cases} 1 - p_i & \text{if } r(i) = 0, \\ 1 - p_0 & \text{if } r(i) = 1 \end{cases} \\ &= (1 - p_i)(1 - r(i)) + (1 - p_0) r(i) \end{aligned}$$

where

$$p_i = P\{L_n > (i+1)t_0 | L_n > it_0\} = (1 - F((i+1)t_0))/(1 - F(it_0))$$

and $p_\Delta = 0$.

A similar argument shows that $P_r(i, i+1)$ and $P_r(i, 1)$ are as given by (1.2).

The cost structure is such that we pay C_f units whenever state Δ is visited and C_p units whenever a preventive replacement is made. Therefore, the expected total discounted cost for any initial state $i \in E_\Delta$ is expressed by

$$V_r(i) = E\left[\sum_{n=0}^{\infty} \alpha^n g_r(X_n) \,\middle|\, X_0 = i\right]$$

where

$$g_r(i) = \begin{cases} C_p r(i) & \text{if } i \neq \Delta, \\ C_f & \text{if } i = \Delta. \end{cases}$$

The optimal control problem is to find V^* and $r^* \in M$ such that $V^* = V_{r^*} \leqq V_r$ for all $r \in M$. We end this section by stating a result which characterizes the return function V_r for all $r \in M$. We would also like to remark that since E_Δ are discrete functions defined on E_Δ and $E_\Delta \times E_\Delta$ are considered as vectors and matrices respectively as in the statement of the following theorem which follows trivially from Çinlar [3, p. 197].

THEOREM 1.4. *For any $r \in M$, there exists a unique function u defined on E_Δ which satisfies*

$$(I - \alpha P_r)u = g_r.$$

Furthermore,

$$u(i) = V_r(i) = E\left[\sum_{n=0}^{\infty} \alpha^n g_r(X_n) \,\middle|\, X_0 = i\right], \qquad i \in E_\Delta.$$

This result simply states that V_r is the unique solution of a system of linear equations given by Theorem 1.4. Note that, V_r satisfies

$$\begin{aligned} V_r(i) = {} & g_r(i) + \alpha p_0 r(i) V_r(1) + \alpha(1-p_i)(1-r(i)) V_r(\Delta) \\ & + \alpha(1-p_0) r(i) V_r(\Delta) + \alpha p_i (1-r(i)) V_r(i+1) \end{aligned} \tag{1.5}$$

for all $i \in E_\Delta$ when expression (1.4) is rewritten in open form.

2. A necessary and sufficient condition of optimality. In this section a necessary and sufficient condition of optimality will be presented so that some interesting results that follow from it can be obtained. The role of the transition matrix P_r is extremely important in the discrete time control of Markov chains. Markov decision chains, as they are frequenly called in the literature, have been studied by many authors who have obtained necessary and sufficient optimality conditions for general models. We refer the reader to Derman [4] and Howard [5] for a detailed treatment of these models.

Our interest lies in characterizing the optimal solution of the specific replacement problem stated and formulated in the previous section. Before we state the main result of this section we introduce the following notation for simplicity. Let B be the set of all bounded functions on E_Δ and define two mappings Γ_0 and Γ_1 on B by,

$$\Gamma_0 f(i) = \alpha p_i f(i+1) + \alpha(1-p_i) f(\Delta), \qquad i \in E, \tag{2.1}$$

$$\Gamma_1 f(i) = \alpha p_0 f(1) + \alpha(1-p_0) f(\Delta), \qquad i \in E, \tag{2.2}$$

$$\Gamma_0 f(\Delta) = \Gamma_1 f(\Delta) = \alpha p_0 f(1) + \alpha(1-p_0) f(\Delta) \tag{2.3}$$

for every $f \in B$. Furthermore, let $\|\cdot\|$ denote the usual supremum norm defined on B, i.e. $\|f\| = \sup_{i \in E_\Delta} |f(i)|$ for any $f \in B$.

Finally, letting $h_0(i)=0$, $h_1(i)=C_p$ for $i\in E$, and $h_0(\Delta)=h_1(\Delta)=C_f$; this notation permits us to write

$$h_{r(i)}(i)=g_r(i),$$

and thus

$$V_r(i)=h_{r(i)}(i)+\Gamma_{r(i)}V_r(i) \tag{2.4}$$

where V_r is as given by (1.5).

THEOREM 2.5. *There is a unique function $V\in B$ which satisfies*

$$V(i)=\min_{x=0,1}\{h_x(i)+\Gamma_x V(i)\},\qquad i\in E_\Delta. \tag{2.6}$$

Furthermore, there exists $r^*\in M$ such that $V=V^*=V_{r^*}\leqq V_r$ for all $r\in M$.

Proof. Define a mapping Γ on B by

$$\Gamma f(i)=\min_{x=1,0}\{h_x(i)+\Gamma_x f(i)\},\qquad i\in E_\Delta \tag{2.7}$$

for $f\in B$. To prove (2.6) it suffices to show that Γ is a contraction mapping since every contraction mapping on the Banach space B has a unique fixed point in B by Banach's contraction mapping theorem. Note that both Γ_0 and Γ_1 are contraction mappings on B since (2.1), (2.2) and (2.3) imply

$$\|\Gamma_0 f-\Gamma_0 g\|\leqq\alpha\|f-g\|,\qquad \|\Gamma_1 f-\Gamma_1 g\|\leqq\alpha\|f-g\|$$

for all f, $g\in B$. Similarly, by (2.7)

$$\Gamma f(i)-\Gamma g(i)=\min_{x=0,1}\{h_x(i)+\Gamma_x f(i)\}-\min_{x=0,1}\{h_x(i)+\Gamma_x g(i)\}.$$

Assuming that $\bar{x}$ minimizes $h_x(i)+\Gamma_x g(i)$, it is clear that

$$\begin{aligned}\Gamma f(i)-\Gamma g(i)&\leqq h_{\bar{x}}(i)+\Gamma_{\bar{x}}f(i)-h_{\bar{x}}(i)-\Gamma_{\bar{x}}g(i)\\&\leqq\Gamma_{\bar{x}}f(i)-\Gamma_{\bar{x}}g(i)\\&\leqq\alpha\|f-g\|,\end{aligned}$$

since $\bar{x}=0$ or 1. A similar argument shows that

$$\Gamma g(i)-\Gamma f(i)\leqq\alpha\|g-f\|.$$

Therefore $\|\Gamma f-\Gamma g\|\leqq\alpha\|f-g\|$, and Γ is a contraction mapping.

To show that $V\leqq V_r$ for all $r\in M$, note that for arbitrary $r\in M$

$$\begin{aligned}V(i)-V_r(i)&=\min_{x=0,1}\{\{h_x(i)+\Gamma_x V(i)\}-h_{r(i)}(i)-\Gamma_{r(i)}V_r(i)\}\\&=\min_{x=0,1}\{\{h_x(i)+\Gamma_x V(i)\}-h_{r(i)}(i)-\Gamma_{r(i)}V(i)+\Gamma_{r(i)}V(i)-\Gamma_{r(i)}V_r(i)\}\\&=h(i)+\Gamma_{r(i)}(V-V_r)(i),\end{aligned}$$

where $h(i)=\min_{x=0,1}\{\{h_x(i)+\Gamma_x V(i)\}-h_{r(i)}(i)-\Gamma_{r(i)}V(i)\}$. Thus, letting $u=V-V_r$, it satisfies

$$u(i)=h(i)+\Gamma_{r(i)}u(i).$$

It now follows from (2.4), (1.5) and Theroem 1.4 that

$$u(i) = E\left[\sum_{n=0}^{\infty} \alpha^n h(X_n) \,\middle|\, X_0 = i\right],$$

and $u = V - V_r \leqq 0$ since $h \leqq 0$.

To complete the proof let $r^*(i)$ be the minimizing value of x in expression (2.6) for $i \in E_\Delta$. To make sure $r^*(i)$ is well-defined, if both $x = 0$ and $x = 1$ minimizes this expression, we set $r^*(i) = 1$. Clearly $r^*(i) = 0$ or 1 and we can let $r^*(\Delta) = 1$ since by (2.3), $\Gamma_0 V(\Delta) = \Gamma_1 V(\Delta)$; thus

$$V(\Delta) = \min_{x=0,1} \{h_x(\Delta) + \Gamma_x V(\Delta)\} = h_1(\Delta) + \Gamma_1 V(\Delta) = C_f + \Gamma_1 V(\Delta). \tag{2.8}$$

Therefore, $r^* \in M$ and $V = V_{r^*}$ since by the way r^* is defined we have

$$V(i) = h_{r^*(i)}(i) + \Gamma_{r^*(i)} V(i), \qquad i \in E_\Delta.$$

This result shows that there is a unique optimal return function and an optimal replacement policy can be found if (2.6) is solved. This could be done by a policy improvement or successive approximations algorithm if the state space of X is finite by making slight changes in the definition of P_r and in the statements of the results stated thereafter. But, as mentioned earlier, our aim is to characterize V^* and r^* in some interesting cases.

3. Characterization of the optimal policy. In all reliability problems, the cases involving increasing failure rate (IFR) and decreasing failure rate (DFR) distributions are of special interest. In this section we will analyze these two cases in more detail and try to find some properties that we intuitively expect V^* and r^* to satisfy. The presentation that follows is designed to show that if F has IFR or DFR an age replacement policy commonly encountered in many reliability problems is in fact optimal.

Recall that our formulation of the optimal control problem involved a periodic inspection scheme. Thus, the statistical data required were $\{p_i\}_{i \in E}$ defined by (1.3) with no further information on the distribution function F. The following result clarifies the relationship between F and $\{p_i\}_{i \in E}$.

LEMMA 3.1. *Let* $\{p_i\}_{i \in E}$ *be as defined by* (1.3). *Then*

a) F *has* IFR *implies* $p_i \geqq p_{i+1}$ *for all* $i \in E$.

b) F *has* DFR *implies* $p_i \leqq p_{i+1}$ *for all* $i \in E$.

Proof. To show a) note that F has IFR implies $(1 - F(x+t))/(1 - F(t))$ is decreasing in t for each $x \geqq 0$. Taking $x = t_0$, and $t_1 = it_0 \leqq t_2 = (i+1)t_0$, we obtain

$$\frac{1 - F((i+1)t_0)}{1 - F(it_0)} \geqq \frac{1 - F((i+2)t_0)}{1 - F((i+1)t_0)},$$

or equivalently $p_i \geqq p_{i+1}$ by (1.3). A similar argument can be made to prove b).

Before we state the main result of this section we need to prove the following lemma.

LEMMA 3.2. *Let* V *be the optimal return function of Theorem* 2.5. *Then*

$$0 \leqq V(i) \leqq V(\Delta) \leqq C_f/(1-\alpha), \qquad i \in E. \tag{3.3}$$

Proof. To prove this result we more generally show that if $f \in B$ satisfies (3.3), then Γf defined by (2.7) also satisfies (3.3). The fact that $\Gamma f(i) \geqq 0$ follows easily since

$f \geqq 0$ implies $\Gamma_x f \geqq 0$, and $h_x \geqq 0$ for $x = 0$ or 1. Now, using the definition of Γf,

$$\Gamma f(i) \leqq h_1(i) + \Gamma_1 f(i) = C_p + \alpha p_0 f(1) + \alpha(1 - p_0) f(\Delta)$$
$$\leqq C_f + \alpha p_0 f(1) + \alpha(1 - p_0) f(\Delta) = \Gamma f(\Delta).$$

Finally, if $f(i) \leqq f(\Delta) \leqq C_f/(1-\alpha)$, then

$$\Gamma f(\Delta) = C_f + \alpha p_0 f(1) + \alpha(1 - p_0) f(\Delta)$$
$$\leqq C_f + \alpha f(\Delta) \leqq C_f + \alpha(C_f/(1-\alpha)) = C_f/(1-\alpha).$$

The results stated above actually describes some properties the optimal return function and periodic survival probabilities are intuitively expected to satisfy. The following clarifies the structure of the optimal return function and optimal replacement policy.

THEOREM 3.4. *Let V and r^* be as given in Theorem* 2.5. *If F has* IFR (*resp.* DFR). *Then*

a) *V is increasing (resp. decreasing) on E,*

b) *r^* is increasing (resp. decreasing) on E.*

Proof. We shall present the proof for the IFR case only to avoid repetition. In light of Lemma 3.2, to prove a) we need to show that if $f \in B$ is increasing on E with $f(i) \leqq f(\Delta)$ for all $i \in E$, then Γf defined by (2.7) is also increasing on E. For arbitrary $i \in E$, note that

$$\Gamma f(i+1) - \Gamma f(i) = \min_{x=0,1} \{h_x(i+1) + \Gamma_x f(i+1)\} - \min_{x=0,1} \{h_x(i) + \Gamma_x f(i)\}.$$

Now, letting $\bar{x}$ be the minimizing value of the first minimization on the right-hand side of this expression we obtain

$$\Gamma f(i+1) - \Gamma f(i) \geqq h_{\bar{x}}(i+1) + \Gamma_{\bar{x}} f(i+1) - h_{\bar{x}}(i) - \Gamma_{\bar{x}} f(i)$$
$$\geqq \Gamma_{\bar{x}} f(i+1) - \Gamma_{\bar{x}} f(i),$$

since $h_{\bar{x}}$ is constant on E for $\bar{x} = 1$ or 0. Therefore, our assertion is trivially satisfied if $\bar{x} = 1$ since $\Gamma_1 f$ is constant on E. If $\bar{x} = 0$ however,

$$\Gamma f(i+1) - \Gamma f(i) \geqq \alpha p_{i+1} f(i+2) + \alpha(1 - p_{i+1}) f(\Delta) - \alpha p_i f(i+1) - \alpha(1 - p_i) f(\Delta)$$
$$\geqq \alpha(p_i - p_{i+1})(f(\Delta) - f(i+1))$$
$$\geqq 0,$$

where the second inequality is obtained by noting that f is increasing on E, and the third inequality follows trivially since p_i is decreasing on E by Lemma 3.1 and $f(j) \leqq f(\Delta)$ for all $j \in E$.

To prove b) we will show that if $r^*(i) = 1$ for some $i \in E$, then $r^*(i+1) = 1$. Noting the way r^* is defined in the proof of Theorem 2.5, $r^*(i) = 1$ implies

$$h_0(i) + \Gamma_0 V(i) \geqq h_1(i) + \Gamma_1 V(i).$$

Now, $h_0(i+1) = h_0(i)$ and it follows similarly as in the proof of part a) that $\Gamma_0 V(i+1) \geqq \Gamma_0 V(i)$. Therefore,

$$h_0(i+1) + \Gamma_0 V(i+1) \geqq h_0(i) + \Gamma_0 V(i)$$
$$\geqq h_1(i) + \Gamma_1 V(i) = h_1(i+1) + \Gamma_1 V(i+1),$$

since h_1 and $\Gamma_1 V$ are constant on E. This in turn implies that $r^*(i+1) = 1$.

This result characterizes the optimal return function and the optimal replacement policy. As we shall see next, we can conclude that a periodic age replacement policy is in fact optimal. Note that for arbitrary F, $r^*(0)=0$ since by Theorem 2.5,

$$\begin{aligned} V(0) &= \min\left(h_0(0)+\Gamma_0 V(0),\ h_1(0)+\Gamma_1 V(0)\right) \\ &= \min\left(\alpha p_0 V(1)+\alpha(1-p_0)V(\Delta),\ C_p+\alpha p_0 V(1)+\alpha(1-p_0)V(\Delta)\right) \\ &= \alpha p_0 V(1)+\alpha(1-p_0)V(\Delta), \end{aligned}$$

which in turn implies by the way r^* is defined that an item of age 0 should not be replaced. Now, the characterization given in Theorem 3.4 directly leads to the following result which we state without proof.

COROLLARY 3.5. *If F has* DFR, *then the optimal replacement policy is to replace the item upon failure only. However, if F has* IFR, *then the optimal replacement policy is a periodic age replacement policy.*

In the DFR case the fact that $r^*(0)=0$ and r^* is decreasing on E implies $r^*(i)=0$ for all $i\in E$. In the IFR case since r is increasing it must have the following structure,

$$r^*(i)=\begin{cases}0, & i\leqq n^*,\\ 1, & i>n^*,\end{cases}$$

for some integer $n^*\geqq 0$ possibly infinite. Thus, the optimal return function V satisfies

$$V(i)=h_0(i)+\Gamma_0 V(i)$$

in the DFR case, and

$$V(i)=\begin{cases}h_0(i)+\Gamma_0 V(i), & i\leqq n^*,\\ h_1(i)+\Gamma_1 V(i), & i>n^*,\end{cases}$$

in the IFR case. Putting these characterizations together the optimal replacement age n^*, and thus V and r^* can be obtained either through an algorithmic approach such as policy improvement or successive approximations, or a renewal theoretic approach as it is commonly done in the literature.

We would like to mention that the assumption $F(x)<1$ for all $x\geqq 0$ is not crucial in our presentation. We could have obtained similar results by dropping this assumption and defining the state space E_Δ of X properly, i.e. $E_\Delta=\{0,1,\cdots,m\}\cup\{\Delta\}$ where m is the largest integer satisfying $F(mt_0)<1$. In this paper our aim has been to illustrate the application of a Markov decision theoretic approach in analyzing the discrete time control of a simple periodic replacement model. The flavor of the results obtained suggest that Markov decision theory could be extremely fruitful in analyzing far more general discrete or continuous time control problems concerning reliability models.

REFERENCES

[1] R. E. BARLOW AND F. PROSCHAN, *Optimum preventive maintenance policies*, Oper. Res., 8 (1960), pp. 90–100.

[2] ———, *Mathematical Theory of Reliability*, John Wiley, New York, 1965.

[3] E. ÇINLAR, *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[4] C. DERMAN, *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.

[5] R. A. HOWARD, *Dynamic Programming and Markov Processes*, John Wiley, New York, 1960.

[6] T. NAKAGAWA, *Studies on optimum maintenance policies for reliable systems*, Ph.D. dissertation, Kyoto University, Tokyo, Japan, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

011

SOME INFORMATION THEORETIC SADDLEPOINTS*

J. MARTIN BORDEN†, DAVID M. MASON‡ AND ROBERT J. McELIECE§

Abstract. In order to study communication in the presence of jamming (or vice-versa), one can formulate a two person zero sum game with mutual information as the payoff function. Since mutual information is convex-concave in just the right way, one often finds *saddlepoint strategies* (which are simultaneously optimal for both players). In this paper we shall discuss these saddlepoints for four distinct cases: Additive noise, average power constraints, and any combination of "hard"/"soft" input/output quantization. One of these cases (soft/soft) has been previously studied (Shannon/Dobrushin/Blachman), but the other three appear to be new.

Key words. information theory, game theory, communications, jamming

1. Introduction. Consider a discrete-time additive channel which accepts, during every unit of time, a real number X. Suppose that this input is garbled in transmission and instead of X,

$$Y = X + Z$$

is transmitted to the receiver, where Z is the "noise" associated with the transmission. Assume that X and Z are independent random variables. A good measure, perhaps the best measure, of the average amount of information conveyed by the channel per unit of time is the *mutual information-I*$(X; Y)$ between X and Y. We define $I(X; Y)$ as in Pinsker [7] and Gray and Kieffer [4] to be

$$I(X; Y) = \sup_{\mathscr{P}_1, \mathscr{P}_2} \sum_{A \in \mathscr{P}_1} \sum_{B \in \mathscr{P}_2} P_{X,Y}(A \times B) \log \left[\frac{P_{X,Y}(A \times B)}{P_X(A) P_Y(B)} \right], \tag{1}$$

where the supremum is taken over all finite partitions $\mathscr{P}_1$ and $\mathscr{P}_2$ of $\mathbb{R}$, $P_{X,Y}$ is the probability measure on $\mathbb{R}^2$ induced by (X, Y), and P_X, P_Y are the probability measures induced on $\mathbb{R}$ by X, Y respectively.

Let us adopt the viewpoint that this is a game with two players. Player I, also called *the Coder*, controls the input X. Player II, which we call *the Jammer*, controls the noise. The Coder's goal is to make $I(X; Y)$ as large as possible, and the Jammer's goal is to make it as small as possible. To emphasize the dependence of $I(X; Y)$ on the independent random variables X and Z, we introduce the notation

$$\phi(X, Z) = I(X; Y); \tag{2}$$

and in line with our game theoretic viewpoint we call ϕ the game's *payoff function.*

This game will be meaningless and trivial unless we place restrictions on the players. We suppose that the Coder's choice of X must lie in a certain set S, the set of *allowable inputs*, and that Z must lie in T, the set of *allowable noises.* Thus associated with this game are two programs.

Program I (*Coder's program*):

$$C' = \sup_{X \in S} \inf_{Z \in T} \phi(X, Z).$$

* Received by the editors October 4, 1983, and in final form March 19, 1983. This research was supported in part by the Air Force Office of Scientific Research under grant AFOSR-83-0296.

† Worcester Polytechnic Institute, Worcester, Massachusetts 01609.

‡ University of Delaware, Newark, Delaware 19711.

§ University of Illinois, Urbana, Illinois 61801.

Program II (*Jammer's program*):

$$C'' = \inf_{Z \in T} \sup_{X \in S} \phi(X, Z).$$

A strategy $X_0 \in S$ such that

$$\inf_{Z \in T} \phi(X_0, Z) = C' \tag{3}$$

is called an *optimal strategy* for the Coder. The significance is that (3) implies

$$\phi(X_0, Z) \geqq C' \tag{4}$$

for all allowable noises Z. Hence, if the Coder chooses the input X_0, he is guaranteed a payoff of at least C', regardless of the Jammer's strategy. Sometimes no such optimal strategy exists, and the Coder must be satisfied with an ε-optimal strategy, i.e. an $X_0 \in S$ such that for a given choice of $\varepsilon > 0$,

$$\phi(X_0, Z) \geqq C' - \varepsilon \tag{5}$$

for all allowable noises Z.

Similarly, we define optimal strategies and ε-optimal strategies for the Jammer, and in place of (3), (4), and (5), we have for a given $Z_0 \in T$

$$\sup_{X \in S} \phi(X, Z_0) = C'', \tag{6}$$

$$\phi(X, Z_0) \leqq C'' \quad \text{for all } X \in S, \tag{7}$$

$$\phi(X, Z_0) \leqq C'' + \varepsilon \quad \text{for all } X \in S \tag{8}$$

(depending on whether the strategy Z_0 is optimal or ε-optimal).

If it happens that $C' = C''$, then combining (4) and (7), we have

$$\phi(X_0, Z_0) = C' = C'', \tag{9}$$

$$\phi(X, Z_0) \leqq \phi(X_0, Z_0) \leqq \phi(X_0, Z) \tag{10}$$

for every choice of allowable X and Z. If (9) holds, which is equivalent to (10), the common value, denoted by C, is called the value of the game. The pair (X_0, Z_0) of optimal strategies is called a *saddlepoint.* In absence of other information, the Coder will want to play strategy X_0, and the Jammer will want to play Z_0.

In this paper, we will study four closely-related special cases of this problem. In each case we assume that the "signal-to-noise" ratio

$$E(X^2)/E(Z^2) = A,$$

where A is a fixed positive number. Without loss of generality we state this assumption as a pair of restrictions on the players:

$$E(X^2) = A, \tag{11}$$

$$E(Z^2) = 1. \tag{12}$$

The four cases that we consider are distinguished by whether the input and/or output are subject to *binary quantization.*

Binary input quantization.

$$X = \begin{cases} \sqrt{A} & \text{with probability (w.p.) } \frac{1}{2}, \\ -\sqrt{A} & \text{w.p. } \frac{1}{2}. \end{cases} \tag{13}$$

Binary output quantization.

$$Y = \mathrm{HD}(X+Z), \tag{14}$$

where HD ("hard decision") is a random variable which equals +1 if its argument is positive, −1 if its argument is negative, and if its argument is zero, is equally likely to be +1 or −1. Our four cases are then as follows:

Case	Restrictions
I(∞/∞)	(11), (12)
II(2/2)	(11), (12), (13), (14)
III(∞/2)	(11), (12), , (14)
IV(2/∞)	(11), (12), (13).

In each case, we will show that the game has a value, which we denote by $C^*(A)$. In Cases I, III, and IV, we will identify the optimal strategies. In case II, we will find that for $A \geqq 1$, the Jammer has no optimal strategy, but will exhibit an ε-optimal strategy for each $\varepsilon > 0$, whereas for $0 < A < 1$ the Jammer does have an optimal strategy.

2. Statements of the main results. In this section, we state our results, postponing the proofs until § 3.

In the following,

$$h(p) = -p \log p - (1-p)\log(1-p) \quad \text{for } p \in [0,1]$$

denotes the binary entropy function.

THEOREM 1. *Case* I(∞/∞). *For every* $A > 0$

$$C^*(A) = \tfrac{1}{2}\log(1+A); \tag{15}$$

and both the Coder and the Jammer have unique optimal strategies given as

$$X_0 = N(0, A), \tag{16}$$

$$Z_0 = N(0, 1). \tag{17}$$

($N(\mu, \sigma^2)$ *denotes a normal random variable with mean* μ *and variance* σ^2.)

THEOREM 2. *Case* II(2/2).

$$C^*(A) = \begin{cases} 0 & \text{for } A < 1, \\ \log 2 - h\left(\dfrac{1}{2A}\right) & \text{for } A \geqq 1; \end{cases} \tag{18}$$

and for $A < 1$, *an optimal strategy for the Jammer is given by*

$$Z_0 = \begin{cases} 1 & \text{w.p. } \tfrac{1}{2}, \\ -1 & \text{w.p. } \tfrac{1}{2}, \end{cases} \tag{19}$$

whereas for $A \geqq 1$, *the Jammer has no optimal strategy. However the Jammer does have, for each* $\varepsilon > 0$, *an* ε-*optimal strategy given by*

$$Z_0 = \begin{cases} \sqrt{A+\delta} & \text{w.p. } 1/(2(A+\delta)), \\ 0 & \text{w.p. } 1 - 1/(A+\delta), \\ -\sqrt{A+\delta} & \text{w.p. } 1/(2(A+\delta)), \end{cases} \tag{20}$$

for $\delta > 0$ *sufficiently small.* (Note that in this case the input random variable is fixed.)

To describe the optimal strategies in Case III($\infty/2$), we introduce the following class of distribution functions $\{F_\mu: \mu \in (0, \log 2]\}$: Choose any $\mu \in (0, \log 2]$. Observe that since h strictly increasing on $(0, \frac{1}{2}]$, there exists a unique $x \in (0, \frac{1}{2}]$ such that $h(x) = \mu$. For any $\mu \in (0, \log 2]$ define

$$\begin{aligned} \lambda = \lambda(\mu) &= 2\int_0^\mu h^{-1}(y)\,dy \\ &= 2\mu x - 2\int_0^x h(t)\,dt = 2\mu x - x - (1-x)^2 \log(1-x) + x^2 \log x, \end{aligned} \tag{21}$$

where $x = h^{-1}(\mu)$. For any such choice of μ, let F_μ be the unique distribution function symmetric about zero satisfying

$$h(F_\mu(z)) = \begin{cases} -\lambda z^2 + \mu & \text{if } |z| \leqq \sqrt{\mu/\lambda}, \\ 0 & \text{if } |z| > \sqrt{\mu/\lambda}. \end{cases} \tag{22}$$

When $\mu = \log 2$, F_μ is differentiable everywhere, but for $\mu < \log 2$, F_μ is differentiable everywhere except at zero where it has a jump discontinuity of height $1 - 2h^{-1}(\mu)$. A simple calculation shows that for each $\mu \in (0, \log 2]$

$$\int z^2\,dF_\mu(z) = 1.$$

THEOREM 3. *Case* III($\infty/2$).

$$C^*(A) = \begin{cases} A\left(\log 2 - \frac{1}{2}\right) & \text{for } A \leqq 1, \\ \left(\log 2 - \frac{1}{2}\right) + \frac{x}{2}\log x - \frac{(1-x)^2}{2x}\log(1-x) & \text{if } A \geqq 1, \end{cases} \tag{23}$$

where $x = 1/(2A)$; *and the optimal strategies for the Jammer and the Coder are given by the following.*

For $A < 1$, Z_0 *has distribution function* $G_0 = F_\mu$ *with* $\mu = \log 2$, *and* X_0 *has distribution function* F_0 *satisfying*

$$F_0'(z) = AG_0'(z) \quad \text{for } z \neq 0, \qquad F_0(0) = 1 - \frac{A}{2};$$

and for $A \geqq 1$, Z_0 *has distribution function* $G_0 = F_\mu$ *with* $\mu = h1/(2A)$, *and* X_0 *has distribution function* F_0 *satisfying*

$$F_0'(z) = AG_0'(z) \quad \text{for } z \neq 0, \qquad F_0 \text{ is continuous at } z = 0.$$

It is easy to show that $EX^2 = A$.

In the final case, we do not present closed forms for the value of the game $C^*(A)$ and the distribution of the optimal strategy for the Jammer. It is doubtful that such closed forms exist. Rather, we demonstrate that for each power constraint $A > 0$ there exists an optimal Jammer strategy Z_0, and describe the type of random variable that Z_0 must be.

THEOREM 4. *Case* IV($2/\infty$). *For each* $A > 0$ *there exists a constant* a *and an optimal Jammer strategy* Z_0 *such that* Z_0 *is a discrete random variable taking on values* $a + 2k\sqrt{A}$ *for every integer* $-\infty < k < \infty$ *and only these values, each with positive probability.*

(For each $A > 0$, $C^*(A)$, the choice of a, and the distribution of Z_0 must be determined numerically.)

3. Proofs of the theorems. If a random variable Y has a density f, let $\tilde{h}(Y) = -\int_{-\infty}^{\infty} f(y) \log f(y)\, dy$ denote the *differentiable entropy* of Y.

Proof of Theorem 1. We require the following lemma.

LEMMA 1. *Let X, Z, and $W_1, W_2, \cdots$, be random variables such that for each $n \geqq 1$ X, Z, and W_n are independent. Assume that $X+Z+W_n$ converges in distribution to $X+Z$. Then*

$$\lim_{n\to\infty} I(X+Z+W_n; X) = I(X+Z; X), \tag{24}$$

and for each $n \geqq 1$

$$I(X+Z+W_n; X) \leqq I(X+Z; X). \tag{25}$$

Proof. Gel'fand and Yaglom [3, item II, p. 206] show

$$\liminf_{n\to\infty} I(X+Z+W_n; X) \geqq I(X+Z; X). \tag{26}$$

Applying an identity due to Dobrushin [2] (see Pinsker [7, p. 45]) we have for each $n \geqq 1$

$$I((X+Z, W_n); X) + I(X+Z; W_n) = I(W_n; (X+Z, X)) + I(X+Z; X).$$

Since by assumption for each $n \geqq 1$ W_n is independent of X and $X+Z$, we have, in addition, that

$$I(X+Z; W_n) = I(W_n; (X+Z, X)) = 0.$$

Hence for each $n \geqq 1$

$$I(X+Z; X) = I((X+Z, W_n); X) \geqq I(X+Z+W_n; X). \tag{27}$$

(For the appropriate facts refer to Pinsker [7, items (1) and (4), p. 11].) (26) and (27) complete the proof of the lemma. □

Let X_0 and Z_0 be as in (16) and (17). Choose any $A > 0$ and X such that $EX^2 \leqq A$. Also choose a $N(0, 1)$ random variable V independent of X, X_0 and Z_0; and set $W_n = n^{-1/2} V$ for each $n \geqq 1$.

Observe that by [3, item II]

$$\phi(X, Z_0) = I(X; X+Z_0) \leqq \liminf_{n\to\infty} I(X+W_n; X+Z_0+W_n).$$

Now for each $n \geqq 1$

$$I(X+W_n; X+Z_0+W_n) = \tilde{h}(X+Z_0+W_n) - \tilde{h}(Z_0+W_n). \tag{28}$$

(28) follows from [6, problem 1.27, p. 44] and the well-known fact that both $X+Z_0+W_n$ and $X+W_n$ have bounded continuous densities. (See, for instance, Breiman [1, Thm. 8.29, p. 178].) According to [6, Thm. 1.11] if W is a random variable with density such that $E(W^2) \leqq B$ then

$$\tilde{h}(W) \leqq \tfrac{1}{2} \log (2\pi e B),$$

with equality if and only if W is $N(0, B)$. Since

$$E(X+Z_0+W_n)^2 = EX^2 + E(Z_0+W_n)^2 \leqq A + 1 + n^{-1},$$

and Z_0+W_n is $N(0, 1+n^{-1})$ it follows from (28) that

$$\phi(X, Z_0) \leqq \tfrac{1}{2} \log (1+A) = \phi(X_0, Z_0).$$

To prove the opposite inequality, let Z satisfy $EZ^2 = 1$, but otherwise be arbitrary; and let W_n for $n \geqq 1$ be defined as above independent of X_0 and Z. By Lemma 1,

$$\phi(X_0, Z) = \lim_{n\to\infty} \phi(X_0, Z + W_n), \tag{29}$$

and as above for each $n \geqq 1$

$$\phi(X_0, Z + W_n) = \tilde{h}(X_0 + Z + W_n) - \tilde{h}(Z + W_n). \tag{30}$$

Now a theorem of Shannon (see [8]) says that if X and Y are independent random variables having differential entropies, then

$$\exp(2\tilde{h}(X + Y)) \geqq \exp(2\tilde{h}(X)) + \exp(2\tilde{h}(Y)), \tag{31}$$

with equality if and only if X and Y are both normal. Applying this to the term $\tilde{h}(X_0 + Z + W_n)$ in (30), we get that

$$\tilde{h}(X_0 + (Z + W_n)) \geqq \tfrac{1}{2}\log(2e\pi A + e^{2\tilde{h}(Z+W_n)}). \tag{32}$$

(Note that since X_0 is normal, $h(X_0) = \frac{1}{2}\log(2\pi eA)$.) Hence combining (30) and (32), we have for each $n \geqq 1$

$$\phi(X_0, Z + W_n) \geqq \tfrac{1}{2}\log(2\pi eA + e^{2\tilde{h}(Z+W_n)}) - \tilde{h}(Z + W_n).$$

Since $\tilde{h}(Z + W_n) \leqq \frac{1}{2}\log(2\pi e(1 + n^{-1}))$, it is not difficult to show that this last expression is

$$\geqq \tfrac{1}{2}[\log(A + 1 + n^{-1}) - \log(1 + n^{-1})].$$

(29) completes the proof. □

Before we prove Theorem 2, we need some additional definitions.

If X is a discrete random variable taking on values in a countable set S, we define the *entropy of* X to be

$$H(X) = -\sum_{x\in S} \log(P(X = x))P(X = x); \tag{33}$$

and if (X, Y) is a bivariate random variable taking on values in a countable set S^*, we define the *expected conditional entropy of* X *given* Y to be

$$H(X|Y) = \sum_{(x,y)\in S^*} \log(P(X = x, Y = y)/P(Y = y))P(X = x, Y = y). \tag{34}$$

It can be shown that $H(X) = I(X; X)$. In fact for general random variables X, $H(X)$ is defined in this way. (Refer to Pinsker [7].) In the proof of Theorem 4 we use the more general definition of $H(X|Y)$ given in [7].

Proof of Theorem 2. Let X_0 and Y be given as in (13) and (14). Then

$$I(X_0; Y) = H(X_0) - H(X_0|Y)$$

(see [6, p. 25]), which

$$= \log 2 - H(X_0|Y).$$

By Fano's inequality [6, p. 23] this last expression is

$$\leqq \log 2 - h(p), \tag{35}$$

where $p = P(X_0 \neq Y)$.

Now

$$P(Y \neq X_0) = \tfrac{1}{2}P(\sqrt{A}\, Y \neq X_0 | X_0 = \sqrt{A}) + \tfrac{1}{2}P(\sqrt{A}\, Y \neq X_0 | X_0 = -\sqrt{A}). \tag{36}$$

But since $Y = HD(X_0 + Z)$,

$$P(\sqrt{A}\, Y \neq X_0 | X = \sqrt{A}) = \tfrac{1}{2}P(Z = -\sqrt{A}) + P(Z < -\sqrt{A}), \tag{37}$$

$$P(\sqrt{A}\, Y \neq X_0 | X = -\sqrt{A}) = \tfrac{1}{2}P(Z = \sqrt{A}) + P(Z > A). \tag{38}$$

Combining (36), (37), and (38) we have

$$P(\sqrt{A}\, Y \neq X_0) = \tfrac{1}{4}P(Z^2 = A) + \tfrac{1}{2}P(Z^2 > A) \leqq \tfrac{1}{2}P(Z^2 \geqq A), \tag{39}$$

with equality if and only if $P(Z^2 = A) = 0$.

By Markov's equality

$$P(Z^2 \geqq A) \leqq \frac{1}{A}, \tag{40}$$

(recall that we assume that $EZ^2 = 1$). It is easy to see that we have equality in (40) only when both $A \geqq 1$ and $P(Z^2 = A) = P(Z^2 \neq 0) = 1/A$. Thus (39) and (40) imply that

$$p = P(Y \neq X_0) < \frac{1}{2A}. \tag{41}$$

The inequality is strict because the conditions for equality in (39) and (40) cannot be satisfied simultaneously.

Now for $0 \leqq p \leqq \frac{1}{2}$, $h(p)$ is monotone increasing. Thus for $A \geqq 1$, (35) and (41) together imply that

$$I(X_0; Y) > \log 2 - h\left(\frac{1}{2A}\right). \tag{42}$$

Hence for $A \geqq 1$, we see that

$$C^*(A) \geqq \log 2 - h\left(\frac{1}{2A}\right).$$

On the other hand for Z_0 as described in (20)

$$I(X_0; Y_0) = \log 2 - h\left(\frac{1}{2(A+\delta)}\right), \tag{43}$$

where $Y_0 = \mathrm{HD}(X_0 + Z_0)$. Since (43) can be made arbitrarily close to the right side of (42), this completes the proof of the $A \geqq 1$ part of Theorem 2.

Suppose $A < 1$; then for Z_0 as described in (19) $I(X_0; Y_0) = 0$. □

Proof of Theorem 3. Choose any $A > 0$, and let X_0 and Z_0 be as described in the statement of Theorem 3. We begin by showing that

$$\phi(X, Z_0) \leqq \phi(X_0, Z_0)$$

for any X satisfying $EX^2 \leqq A$. The payoff function can be written as

$$\phi(X, Z_0) = I(X; Y) = H(Y) - H(Y|X),$$

where the expected conditional entropy $H(Y|X)$ is the expectation with respect to X of the quantity

$$H(Y|X = x) = h(P(Y = -1|X = x)). \tag{44}$$

But

$$P(Y=-1|X=x)=P(X+Z_0<0|X=x)+\tfrac{1}{2}P(X+Z_0=0|X=x),$$
$$=P(Z_0<-x)+\tfrac{1}{2}P(Z_0=x).$$

So

$$P(Y=-1|X=x)=\begin{cases}G_0(-x) & \text{for } x\neq 0,\\ \frac{1}{2} & \text{for } x=0.\end{cases} \tag{45}$$

It follows from (44), (45) and (22) that

$$H(Y|X=x)=\begin{cases}-\lambda x^2+\mu & \text{for } x\neq 0,\\ \log 2 & \text{for } x=0.\end{cases} \tag{46}$$

Since $\mu \leqq \log 2$, we have (denoting the distribution of X by F) that

$$H(Y|X)\geqq \int (-\lambda x^2+\mu)\, dF(x)\geqq -\lambda A+\mu. \tag{47}$$

Also observe that since $Y=\mathrm{HD}(X+Z_0)$ assumes only two values, we have

$$H(Y)\leqq \log 2. \tag{48}$$

Hence we have in combination with (47) that

$$\phi(X, Z_0)\leqq \log 2+\lambda A-\mu, \tag{49}$$

with equality if and only if

(50a) the distribution of X is symmetric about zero, which is the condition for equality in (48);

and

(50b) $P(X=0)=0$ if $\mu<\log 2$, $P(|X|\geqq\sqrt{\mu/\lambda})=0$, $EX^2=A$, which are the conditions for equality in (47).

The random variable X_0 has all these properties. Thus we have shown that

$$\phi(X, Z_0)\leqq \phi(X_0, Z_0) \tag{51}$$

for all random variables satisfying $EX^2\leqq A$. This completes the first half of the proof of optimality of the pair (X_0, Z_0). Incidentally, elementary calculations show that $\phi(X_0, Z_0)=C^*(A)$.

To finish the proof of optimality, we need to establish that

$$\phi(X_0, Z_0)\leqq \phi(X_0, Z) \tag{52}$$

for all Z satisfying $EZ^2=1$.

We indicate the dependence of $\phi(X_0, Z)$ on the distribution G (say) of Z by writing

$$\omega(G)=\phi(X_0, Z)=H(Y)-H(Y|X_0)=h\left(\int G^*(-x)\, dF_0(x)\right)-\int h(G^*(-x))\, dF_0(x),$$

where $G^*(x)=\frac{1}{2}[G(x)+G(x-)]$. We will show that

$$\phi(X_0, Z_0)=\min\{\omega(G_0+\eta)\colon \eta\in\mathscr{A}\}, \tag{53}$$

where $\mathscr{A}$ denotes the class of all real valued functions of bounded variation on $(-\infty, \infty)$ such that

$$\int d\eta(x) = 0, \tag{53a}$$

$$\int x^2 \, d\eta(x) \leqq 0, \tag{53b}$$

$$0 \leqq G_0(x) + \eta(x) \leqq 1 \quad \text{for } -\infty < x < \infty. \tag{53c}$$

Observe that $\mathscr{A}$ is convex. Any η satisfying these conditions will be called admissible. Once (53) is proven we will have (52) since $\eta = G - G_0$ is in $\mathscr{A}$.

Note that the admissible η's may be assumed to have support in $(-\sqrt{\mu/\lambda}, \sqrt{\mu/\lambda})$. Consider a distribution function $G = G_0 + \eta$, whose support does not lie entirely in $(-\sqrt{\mu/\lambda}, \sqrt{\mu/\lambda})$. Define

$$G_1(x) = \begin{cases} G(x) & \text{for } x \in (-\sqrt{\mu/\lambda}, \sqrt{\mu/\lambda}), \\ 0 & \text{for } x \leqq -\sqrt{\mu/\lambda}, \\ 1 & \text{for } x \geqq \sqrt{\mu/\lambda}. \end{cases}$$

Since $F_0'(x) = 0$ when $|x| > \sqrt{\mu/\lambda}$, it is easy to see that $\omega(G) = \omega(G_1)$. Moreover, we can write $G_1 = G_0 + \eta_1$, where η_1 has support in $(-\sqrt{\mu/\lambda}, \sqrt{\mu/\lambda})$.

The Gâteaux derivative of ω at G_0 in any admissible direction η, $D_\eta \omega(G_0)$, splits naturally into the difference of two terms. The first term is

$$h'\left(\int G_0^*(-x) \, dF_0(x)\right) \cdot \int \eta(-x) \, dF_0(x).$$

However, the argument of h' is $P(Y = -1) = \frac{1}{2}$ (by symmetry of G_0 about zero) and thus this term vanishes. Therefore,

$$D_\eta \omega(G_0) = -\int h'(G_0^*(-x)) \eta(-x) \, dF_0(x).$$

By the remarks about F_0' just given, this integral can be restricted to the interval $[-\sqrt{\mu/\lambda}, \sqrt{\mu/\lambda}]$. Also note that by symmetry of G_0 for $x \neq 0$ $G_0(-x) = 1 - G_0(x)$ so that $G_0'(-x) = G_0'(x)$. Hence by the properties of G_0 and G_0' given in the statement of Theorem 3 for $x \neq 0$

$$-2\lambda x = \frac{d}{dx} h(G_0^*(-x)) = -h'(G_0(-x)) G_0'(-x)$$

$$= -h'(G_0(-x)) G_0'(x) = -\frac{1}{A} h'(G_0^*(-x)) F_0'(x).$$

Thus for $x \neq 0$

$$h'(G_0^*(-x)) \, dF_0(x) = 2A\lambda x \, dx.$$

On the other hand when $x = 0$, either $G_0^*(-x) = \frac{1}{2}$ or $F_0(x)$ is continuously differentiable, so that the integral for $D_\eta \omega(G_0)$ receives no contribution at $x = 0$. We have thus deduced that

$$D_\eta \omega(G_0) = -\int_{-\sqrt{\mu/\lambda}}^{\sqrt{\mu/\lambda}} \eta(-x) 2A\lambda x \, dx.$$

Applying integration by parts and (53a) we see that this last expression equals

$$-A\lambda \int_{-\sqrt{\mu/\lambda}}^{\sqrt{\mu/\lambda}} y^2 \, d\eta(y),$$

which by the fact that $-A\lambda \leqq 0$ along with (53b) is nonnegative. This proves that G_0 is a local minimizer of ω. However, since ω is a convex function of $G = G_0 + \eta$, G_0 is in fact a global minimizer of ω on $\mathscr{A}$. This completes the proof of Theorem 3. □

Proof of Theorem 4. As in Theorem 2, the input X_0 is fixed and only Z is allowed to vary; accordingly, we find it advantageous to view the solution of the Jammer's problem as a function of the noise-to-signal ratio $E = 1/A$. Hence, instead of (11) and (12), we take

$$E(X^2) = 1, \tag{54}$$

$$E(Z^2) \leqq E, \tag{55}$$

and prove that for each $E > 0$ there exists a discrete optimal Jammer strategy Z_0 which has support $\{a + 2k | k \text{ is integral}\}$ for some real constant a.

Let us outline the approach. The Jammer wishes to choose Z so as to minimize $\phi(X_0, Z) = H(X_0) - H(X_0|Y)$, or what is the same, maximize the concave function $J(Z)$ defined by

$$J(Z) = H(X_0|Y) \tag{56}$$

(a precise description of J follows). Thus we wish to solve the concave program

$$\zeta(E) = \sup J(Z) \quad \text{such that } E(Z^2) \leqq E. \tag{57}$$

Although Z is allowed to be any measurable function, we use linear programming ideas to prove

LEMMA 2. *If Z satisfies $E(Z^2) < \infty$ and $\varepsilon > 0$ then there exists a discrete random variable W having $J(W) \geqq J(Z)$, $E(W^2) \leqq E(Z^2) + \varepsilon$, and such that the support of W is contained in $\{x | x \equiv a_1 \text{ or } a_2 \bmod 2\}$ for some real constants a_1, a_2.*

Using a limiting argument, we obtain

LEMMA 3. *For any $E > 0$ there exists a discrete optimal Jammer density p having second moment not exceeding E which is a convex combination of two lattice densities of span* 2.

We next wish to show that the support of an optimal p is precisely one lattice set. To facilitate a Lagrangian analysis, we introduce, for any discrete set C, an ancillary function

$$\zeta_C(E) = \max J(p) \tag{58}$$

where the maximum is taken over all nonnegative discrete functions p with support in C, total mass 1, and second moment not exceeding E. We record a number of observations:

(i) ζ is a strictly increasing function of E.

(ii) If $\zeta_C(E) > 0$ is attained by p then there exists Lagrange multipliers $\mu = \mu(E)$ and $\lambda = \lambda(E) \geqq 0$ such that p maximizes

$$L(r, \mu, \lambda) = J(r) - \mu(M_0(r) - 1) - \lambda(M_2(r) - E) \tag{59}$$

among all admissible functions r. (Here $M_0(r)$ and $M_2(r)$ are respectively the total mass and second moment of r.) Also, if $\lambda(E) > 0$ then $M_2(r) = E$.

(iii) In the same situation as in (ii), p satisfies the necessary condition

$$J'(p)_c = \mu(E) + \lambda(E)c^2. \tag{60}$$

where

$$J'(p)_c = \tfrac{1}{2}\log(1 + p(c+2)/p(c))(1 + p(c-2)/p(c)), \tag{61}$$

for each $c \in C$ such that $p(c) > 0$.

(iv) The support of any optimal solution to (57) is closed under translation by ± 2.

(v) Suppose p_i attains $\zeta(E_i)$ and $E_1 < E_2$. Take C to be the union of the supports of p_1 and p_2 in (58). Then any choices of corresponding Lagrange multipliers from (ii) satisfy $\lambda(E_1) > \lambda(E_2)$.

With these observations established we can complete the proof of the theorem. Suppose that the theorem is false, that is, for some $E_0 > 0$ the value of $\zeta(E_0)$ cannot be attained by a lattice density. From Lemma 2, there exists a p_0 of the form $p_0 = \alpha p_1 + (1-\alpha)p_2$, $0 < \alpha < 1$, where p_1 places its mass on a lattice A and p_2 places its mass on a lattice B (necessarily $A \cap B = \varnothing$) and such that $J(p_0) = \zeta(E_0)$. Observe that each of the quantities $E_1 = M_2(p_1)$, $E_2 = M_2(p_2)$ and $E_0 = \alpha E_1 + (1-\alpha)E_2$ must be distinct for otherwise it would be possible to use p_1 or p_2 to attain $\zeta(E_0)$. Of course, $J(p_1) = \zeta_A(E_1)$ and $J(p_2) = \zeta_B(E_2)$. Using the concavity of ζ we have that

$$\begin{aligned}
\zeta(E_0) &\geqq \alpha\zeta(E_1) + (1-\alpha)\zeta(E_2) \\
&\geqq \alpha\zeta_A(E_1) + (1-\alpha)\zeta_B(E_2) \\
&= \alpha J(p_1) + (1-\alpha)J(p_2) \\
&= J(p) = \zeta(E_0).
\end{aligned}$$

Equality must hold throughout so that in fact $J(p_1) = \zeta(E_1)$ and $J(p_2) = \zeta(E_2)$.

Let $C = A \cup B$ be the support of p_0. By observations (iii) and (iv) we have

$$J'(p_1)_a = \mu(E_1) + \lambda(E_1)a^2 \quad \text{for } a \in A,$$

and

$$J'(p_0)_c = \mu(E_0) + \lambda(E_0)c^2 \quad \text{for } c \in C.$$

Observe however that (61) shows that $J'(p_0)_c$ is homogeneous of degree zero, meaning that for all $\alpha > 0$, $J'(\alpha p_0)_c = J(p_0)_c$. Since $p_0(a \pm 1) = \alpha p_1(a \pm 1)$ whenever $a \in A$, we thus find that $J'(p_0)_a = J'(\alpha p_1)_a = J'(p_1)_a$ for all $a \in A$. Therefore $\mu(E_1) + \lambda(E_1)a^2 = \mu(E_0) + \lambda(E_0)a^2$ for all $a \in A$. Since the set A is infinite we must have $\lambda(E_1) = \lambda(E_0)$. This contradiction to (v) will complete the proof of the theorem.

We turn now to establishing all of the claims above. Recall that X_0 has support $\{-1, +1\}$, so that by a straightforward calculation, (1) and (2) become

$$\phi(X_0, Z) = \log 2 - \inf_{\mathscr{P}} \sum_{B \in \mathscr{P}} \log R(\tfrac{1}{2}P_Z(B-1), \tfrac{1}{2}P_Z(B+1)), \tag{62}$$

where

$$R(x, y) = (x+y)^{x+y}/x^x y^y = (1 + y/x)^x (1 + x/y)^y$$

for $x \geqq 0$, $y \geqq 0$ (and $0^0 = 1$). We define, for any finite partition $\mathscr{P}$ of $\mathbb{R}$,

$$J_{\mathscr{P}}(Z) = \sum_{B \in \mathscr{P}} \log R(\tfrac{1}{2}P_Z(B-1), \tfrac{1}{2}P_Z(B+1)), \tag{63}$$

and

$$J(Z) = \inf_{\mathscr{P}} J_{\mathscr{P}}(Z), \tag{64}$$

where the infimum is taken over all finite partitions $\mathcal{P}$ of $\mathbb{R}$. $J(Z)$ is the expected conditional entropy of X_0 given Y. When Z has a discrete density p, we can write

$$J(p) = J(Z) = \sum_y \log R(\tfrac{1}{2}p(y-1), \tfrac{1}{2}p(y+1)), \tag{65}$$

where the sum is taken over all possible values of $Z + X_0$. Notice that the right-hand side of (65) remains meaningful, although possibly infinite, when p is an arbitrary nonnegative discrete function. Making use of the inequality

$$R(x_1, y_1)R(x_2, y_2) \leqq R(x_1 + x_2, y_1 + y_2),$$

which is readily established using the inequality of the arithmetic and geometric means, and the fact that J is homogeneous of degree one, we obtain the important inequality

$$\alpha_1 J(p_1) + \alpha_2 J(p_2) \leqq J(\alpha_1 p_1 + \alpha_2 p_2), \tag{66}$$

valid whenever p_1 and p_2 are nonnegative discrete functions and α_1 and α_2 are nonnegative constants. In particular, J is concave.

Proof of Lemma 2. For arbitrary integers k and integer $m \geqq 1$ let $I_{m,k} = [2(k-1)/m, 2k/m)$. Let $\mathcal{P}_m$ be the partition $\mathcal{P}_m = \{I_{m,k}: -\infty < k < \infty\}$. Clearly $\mathcal{P}_m$ can be obtained as a limit of refinements of finite partitions of $\mathbb{R}$ so that, from (62),

$$J(Z) \leqq J_{\mathcal{P}_m}(Z). \tag{67}$$

It is routine to show that for all integers m sufficiently large

$$\sum_{k=-\infty}^{\infty} \left(\frac{2k}{m}\right)^2 P_Z(I_{m,k}) \leqq M_2(Z) + \varepsilon.$$

Choose any such m and let U be the discrete random variable that places mass $P_U(2k/m) := P_Z(I_{m,k})$ at the point $2k/m$ for each integer k. Thus $M_2(U) \leqq M_2(Z) + \varepsilon$. Now, since $P_U(2k/m \pm 1) = P_Z(I_{m,k} \pm 1)$, we have

$$\begin{aligned} J(U) &= \sum_{k=-\infty}^{\infty} \log R\left(\tfrac{1}{2}P_U\left(\frac{2k}{m} - 1\right), \tfrac{1}{2}P_U\left(\frac{2k}{m} + 1\right)\right) \\ &= \sum_{k=-\infty}^{\infty} \log R(\tfrac{1}{2}P_Z(I_{m,k} - 1), \tfrac{1}{2}P_Z(I_{m,k} + 1)) = J_{\mathcal{P}_m}(Z) \geqq J(Z) \end{aligned}$$

by virtue of (67).

U is a lattice random variable with span $2/m$ which is not quite what is desired. For each integer j, $0 \leqq j < m$, consider the mass placed by U on the set of points $\{j/m + 2k: k \in \mathbb{Z}\}$. Clearly we can choose a discrete density p_j whose support is a subset of this set, and a nonnegative constant α_j such that $P_U(j/m + 2k) = \alpha_j p_j(j/m + 2k)$ for all k. In this way we view P_U as a convex combination of lattice densities, each having span 2:

$$P_U = \sum_{j=0}^{m-1} \alpha_j p_j, \quad \text{where } 0 \leqq \alpha_j \leqq 1 \quad \text{and} \quad \sum_{j=0}^{m-1} \alpha_j = 1.$$

Observe that equality holds in (66) whenever p_1 and p_2 have disjoint supports. Hence

$$J(U) = \sum_{j=0}^{m-1} \alpha_j J(p_j) \quad \text{and} \quad M_2(U) = \sum_{j=0}^{m-1} \alpha_j M_2(p_j).$$

Consider the following linear program.

Maximize

$$\sum_{j=0}^{m-1} y_j J(p_j) \quad \text{such that}$$

$$\sum_{j=0}^{m-1} y_j = 1, \quad \sum_{j=0}^{m-1} y_j M_2(p_j) \leqq M_2(U), \quad y_j \geqq 0.$$

Since the feasible region of this program is nonempty (it contains $(\alpha_0, \cdots, \alpha_{m-1})$) and is obviously bounded, there exists a finite optimal solution to the program. Moreover, there exists a basic optimal solution (see for example [5]), which means here that there is an optimal solution $(y_0^*, \cdots, y_{m-1}^*)$ having at most two nonzero components (corresponding to the two constraints). Define $p = \sum_{j=0}^{m-1} y_j^* p_j$ and let W have density p. Then W is a discrete random variable of the desired form, $M_2(W) \leqq M_2(U) \leqq M_2(Z) + \varepsilon$, and $J(W) \geqq J(U) \geqq J(Z)$. □

We wish to let $\varepsilon \to 0$ in Lemma 1 and the appropriate device is that of convergence in distribution. We require some basic facts, all of which can be found in Breiman [1]. Let $\{F_m\}$ be a sequence of distributions on $\mathbb{R}$. If the F_m have uniformly bounded second moments then $\{F_m\}$ is mass-preserving, that is, for each $\gamma > 0$ there exists a constant c such that $F_m(c) - F_m(-c) > 1 - \gamma$, independently of m. This implies that there exists a distribution F and a subsequence m_i such that F_{m_i} converges to F in distribution. We can also assert that for an open interval I and continuous function g,

$$\int_I g(x)\, dF(x) \leqq \liminf \int_I g(x)\, dF_{m_i}(x). \tag{68}$$

Finally, suppose that F_m is a discrete distribution placing mass $p_m(a_{mk})$ on a_{mk} $(-\infty < k < \infty)$ and that F_m converges in distribution to F, where F concentrates mass $p(a_k)$ at a_k. Then if $a_{mk} \to a_k$ as $m \to \infty$ we also have that $p_m(a_{mk}) \to p(a_k)$ (see, for example, [1, problem 8.3]).

Proof of Lemma 3. Choose any sequence $\varepsilon_m \downarrow 0$ and let Z_m be chosen so that $M_2(Z_m) \leqq E$ and $J(Z_m) + \varepsilon_m \geqq \zeta(E)$. According to Lemma 1, we can choose a discrete density p_m of the form $p_m = \alpha_m p_{1m} + (1 - \alpha_m) p_{2m}$, where $0 \leqq \alpha_m \leqq 1$, p_{1m} places all its mass on $A_m = \{a_m + 2k;\ k \in \mathbb{Z}\}$, p_{2m} places all its mass on $B_m = \{b_m + 2k:\ k \in \mathbb{Z}\}$, $M_2(p_m) \leqq E + \varepsilon_m$, and $J(p_m) \geqq J(Z_m)$. Without loss of generality $a_m, b_m \in [0, 2]$ so that by passing to a subsequence if necessary, we assume that $a_m \to a$, $b_m \to b$, and $\alpha_m \to \alpha$ as $m \to \infty$.

Case $0 < \alpha < 1$. For m sufficiently large $\alpha_m \geqq \alpha/2$ and $M_2(\alpha_m p_{1m}) \leqq M_2(p_m) \leqq 2E$, so that $M_2(p_{1m}) \leqq 4E/\alpha$. Therefore distributions F_{1m} (say) of the p_{1m} have uniformly bounded second moments and by passing to a subsequence we may assume that the F_{1m} converge in distribution to F_1 (say). Let $A = \{a + 2k:\ k \in \mathbb{Z}\}$ and let I be any open interval whose closure contains no points of A. For all m sufficiently large $I \cap A_m = \varnothing$, so that from (68),

$$\int_I dF_1(X) \leqq \liminf \int_I dF_{1m}(X) = 0.$$

This is enough to prove that the density p_1 of F_1 places all its mass on A. By further thinning the sequence and following the same procedure, we may assume that p_{2m} converges (in this same sense) to a density p_2 placing mass only on $B = \{b + 2k:\ k \in \mathbb{Z}\}$. Hence p_m converges (in this sense) to $p = \alpha p + (1 - \alpha) p_2$. Using $g(x) = x^2$ in (68) we see that

$$M_2(p) \leqq \liminf M_2(p_m) \leqq \liminf E + \varepsilon_m = E.$$

From our comments above $p_m(a_m+kd)\to p(a+kd)$ for each k as $m\to\infty$. Thus, for any K,

$$\sum_{k=-K}^{K}\log R(\tfrac{1}{2}p_{1m}(a_m+2k-1),\tfrac{1}{2}p_{2m}(a_m+2k+1))$$

$$\to\sum_{k=-K}^{K}\log R(\tfrac{1}{2}p_1(a+2k-1),\tfrac{1}{2}p_1(a+2k+1)).$$

We use the easily established inequality

$$\log R(x,y)\leqq(x+y)\log 2 \tag{69}$$

to bound the contribution to $J(p_1)$ made by the terms where k falls outside of $[-K,K]$ by

$$(\log 2)\sum_{|k|>K}\tfrac{1}{2}p_{1m}(am+2k-1)+\tfrac{1}{2}p_{1m}(a_m+2k+1).$$

Since $\{F_{1m}\}$ is mass-preserving, by taking K sufficiently large this contribution can be made arbitrarily small and we deduce that $J(p_{1m})\to J(p_1)$ as $m\to\infty$. Similarly $J(p_{2m})\to J(p_2)$ and we have $\zeta(E)\geqq J(p)\geqq\lim J(p_m)\geqq\lim J(Z_m)=\zeta(E)$. Thus, p satisfies the conditions of the lemma.

Case $\alpha=0$ *or* $\alpha=1$. By a straightforward modification of the above argument one shows that there exists a lattice density p with $M_2(p)\leqq E$ and $J(p)=\zeta(E)$.

Proof of assertions (i)–(v).

(i) It is plain from (57) that ζ is increasing. Suppose however that ζ is constant on an interval. From the fact that a local maximum of a concave function is a global maximum, we infer that there exists a finite E_1 such that $\zeta(E_1)=\sup_E\zeta(E)=\log 2$. If the discrete random variable Z attains $\zeta(E_1)$, then $\phi(X_0,Z)=I(X_0;X_0+Z)=0$, that is, X_0 and X_0+Z are independent. But this is impossible: choose any constant c such that $P(Z=c)\neq P(Z=c+1)$ to see that $P(X_0=1,X_0+Z=c+1)=\frac{1}{2}P(Z=c)\neq\frac{1}{2}\{\frac{1}{2}P(Z=c)+\frac{1}{2}P(Z=c+1)\}=P(X_0=1)P(X_0+Z=c+1)$. Thus ζ must be strictly increasing.

(ii) Given a discrete set C let $\mathscr{F}_C$ denote the class of nonnegative functions r defined on C having second moment $M_2(r)<\infty$. (56) is equivalent to $\zeta_C(E)=\max J(r)$ such that

$$r\in\mathscr{F}_C,$$

$$M_0(r)=1,$$

$$M_2(r)\leqq E.$$

Here J is concave, $\mathscr{F}_C$ is a convex set, M_0 and M_2 are linear and (ii) follows by applying for example, Hestenes [5, Thm. 3.1].

(iii) Fix $c\in C$ such that $p(c)>0$ and consider any function $r\in\mathscr{F}_C$ that disagrees with p only at c. By (ii), p maximizes (59) over $\mathscr{F}_C$ so that the right-hand Gâteaux derivative

$$\lim_{\theta\downarrow 0}[L(p+\theta(r-p),\mu(E),\lambda(E))-L(p,\mu(E),\lambda(E))]/\theta\leqq 0. \tag{70}$$

By a routine calculation this becomes

$$[J'(p)_c-\mu(E)-\lambda(E)c^2](r(c)-p(c))\leqq 0.$$

Since $r(c)\geqq 0$ is arbitrary, this implies (60).

(iv) It suffices to argue that if $p(c)=0$, then $p(c\pm 2)=0$. Suppose $p(c)=0$. Again choose r to disagree with p only at c. By applying the mean value theorem to the left-hand side of (70), we see that $J'(p)_c=\infty$, which violates (61), unless $p(c+2)=0$ and $p(c-2)=0$.

(v) From (ii), we have

$$J(p_1)=\max_{r\in\mathcal{F}_C} L(r,\mu(E_1),\lambda(E_1)),$$

so that

$$J(p_1)\geqq J(p_2)-\mu(E_1)(M_0(p_2)-1)-\lambda(E_1)(M_2(p_2)-E_1).$$

Necessarily $M_0(p_2)=1$, so that we can rewrite this to give

$$\zeta(E_1)-\zeta(E_2)=J(p_1)-J(p_2)\geqq\lambda(E_1)(E_1-E_2).$$

Reversing the roles of p_1 and p_2 in the argument yields the inequality

$$\lambda(E_2)(E_1-E_2)\geqq\zeta(E_1)-\zeta(E_2).$$

By (i), $\zeta(E_1)-\zeta(E_2)<0$, which entails that $\lambda(E_1)>\lambda(E_2)$. □

REFERENCES

[1] L. BREIMAN (1968), *Probability*, Addison-Wesley, Reading, MA.

[2] R. L. DOBRUSHIN (1959), *General formulation of Shannon's main theorem in information theory*, Uspekhi Mat. Nauk 14, 6(90), pp. 3–104, English translation in American Mathematical Society Translations, Ser. 2, vol. 33, 1963.

[3] I. M. GEL'FAND AND A. M. YAGLOM (1957), *Calculation of the amount of information about random functions contained in another such function*, Uspekhi Mat. Nauk (N.S.) 12, no. 1(73), pp. 3–52. (In Russian.)

[4] R. M. GRAY AND J. C. KIEFFER (1980), *Mutual information rate, distortion, and quantization in metric spaces*, IEEE Trans. Inform. Theory, vol. IT-26, pp. 412–422.

[5] M. R. HESTENES (1975), *Optimization Theory: The Finite Dimensional Case*, John Wiley, New York.

[6] R. J. MCELIECE (1977), *Theory of Information and Coding*, Addison-Wesley, Reading, MA.

[7] M. S. PINSKER (1964), *Information and Information Stability of Random Variables and Processes*, Holden-Day, San Francisco.

[8] C. E. SHANNON (1948), *A mathematical theory of communication*, Bell. Syst. Tech. J., October 1948, Appendix 6.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

012

ON STABILIZABILITY OF LINEAR SPECTRAL SYSTEMS VIA STATE BOUNDARY FEEDBACK*

RUTH F. CURTAIN†

Abstract. This paper gives conditions under which a class of linear-distributed systems may be stabilized by state boundary feedback control. The system operator is assumed to be spectral with a discrete spectrum, but it may have infinitely many unstable eigenvalues.

Key words. spectral operators, stabilizability state boundary feedback, distributed systems

1. Introduction. It is well known that in finite dimensional state problems controllability of a linear system implies arbitrary spectrum assignability by state feedback. In infinite dimensions this is not in general true, although several authors [1], [7], [8], [10] have established sufficient conditions for spectrum assignability for certain classes of infinite dimensional systems by state feedback.

In [10] Sun considered the following control system on a Hilbert space H:

$$\dot{z} = Az + bu(t), \qquad z(0) = z_0 \tag{1.1}$$

under the state feedback

$$u(t) = \langle z(t), g\rangle, \tag{1.2}$$

where $g, b \in H$ and $\langle \cdot, \cdot \rangle$ denotes the inner product in H. He assumed further that A satisfies the condition F:

(F1) A is an unbounded spectral operator with discrete spectrum, $\sigma = \{\lambda_k; k = 1, 2, \cdots\}$. and normalized eigenvectors $\{\phi_k; k = 1, 2, \cdots\}$. The eigenvalues are distinct and the eigenspaces are one-dimensional.

(F2) $\inf_{i \neq j} |\lambda_i - \lambda_j| = \delta > 0$.

$$\text{(F3)} \qquad \sup_{i \geqq k < \infty} \sum_{\substack{j=1 \\ j \neq k}}^{\infty} \frac{1}{|\lambda_j - \lambda_k|^2} < \infty.$$

His main result was:

THEOREM 1. *Suppose that A satisfies condition* F *and $b \in H$. Then for all $g \in H$, the closed operator $A + \langle \cdot, g\rangle b$ is spectral. Furthermore for a given sequence of complex numbers $\Lambda = \{\gamma_1, \gamma_2, \cdots\}$, in order that there should exist a $g \in H$ so that the spectrum of the closed loop operator $A + \langle \cdot, g\rangle b$ satisfies*

$$\sigma(A + \langle \cdot, g\rangle b) = \sigma_p(A + \langle \cdot, g\rangle) = \Lambda,$$

a necessary and sufficient condition is that

$$b_k = \langle \phi_k, b\rangle \neq 0, \qquad k \geqq 1, \tag{1.3}$$

$$\sum_{k=1}^{\infty} \left| \frac{\lambda_k - \gamma_k}{b_k} \right|^2 < \infty. \tag{1.4}$$

In his proof Sun assumes "without loss of generality" that the $\{\phi_k\}$ can be taken to be orthonormal. While this is correct, the expression (1.3) for b_k is only valid for

* Received by the editors June 7, 1983, and in revised form March 2, 1984.

† Rijksuniversiteit Groningen Mathematisch Instituut, Postbus 800, 9700 AV Groningen, the Netherlands.

the orthogonal case (A normal): we give a modified result later. Then (1.3) is seen to be a controllability condition on (A, b) [2] and (1.4) a restriction on the separation of λ_k and γ_k. Since $\sum_{k=1}^{\infty} |b_k|^2 < \infty$ and since $\{\lambda_k; k \geqq 1\}$ have no finite limit point (A has compact resolvent), this means that if you try to shift infinitely many eigenvalues, as $k \to \infty$, the most by which you can separate $|\lambda_k - \gamma_k|$ is of the order of $|b_k|^2$ and this tends to zero as $k \to \infty$.

The class of distributed systems generated by spectral operators is an important one and includes the many differential operators arising in boundary-value problems involving nonsymmetric linear differential operators, whose eigenfunction expansions converge is much the same way as Fourier series [5], [9]. Thus they enjoy many of the properties of systems generated by self-adjoint operators. Because of their importance in applications, we devote § 2 to a discussion of the properties of spectral systems.

In § 3 we derive our main result: conditions for stabilizability of linear spectral systems via boundary state feedback. The type of boundary control action is of *integral* type such as was introduced in [3] and further discussed in [4]. Although this does not lead to a precise generalization of the nice finite-dimensional form, it could be argued that it is a more practical control implementation for boundary controlled systems. It is an alternative to the method proposed in [11], for example. An illustrative example of a class of hyperbolic systems is included.

2. Spectral systems. We consider the following system in a separate Hilbert space H

$$\dot{z} = Az + Bu, \qquad z(0) = z_0, \tag{2.1}$$

where A is a regular spectral operator on H, U is a Hilbert space, $B \in \mathscr{L}(U, H)$ and $z_0 \in H$. Following Schwarz in [9] we define:

DEFINITION 1. A *regular spectral operator*, A, is an operator with compact resolvent whose spectrum is not the entire plane and whose spectral measures $E(\lambda_i)$ for $\lambda_i \in \sigma(A)$ generate a uniformly bounded Boolean algebra. We remark that if C is a closed curve surrounding only the eigenvalue λ_i, then we have

$$E(\lambda_i) = \frac{1}{2\pi i} \int_C (\lambda - A)^{-1}\, d\lambda. \tag{2.2}$$

The above definition agrees with the one used by Sun in [10], but in order to discuss the semigroup, we need to assume further the completeness property:

$$\sum_{i=1}^{\infty} E(\lambda_i) = I, \tag{2.3}$$

where the convergence is in the strong topology.

We quote the following result from [9].

THEOREM 2. *If A is a regular spectral operator with only finitely many multiple eigenvalues and* (2.3) *holds, then for every function f which is uniformly bounded on $\sigma(A)$ and which is C^{k_i} near λ_i (k_i is the multiplicity of λ_i), the following operator $f(A)$ is well-defined*:

$$f(A) = \sum_{i=1}^{\infty} f(\lambda_i) E(\lambda_i) \tag{2.4}$$

where the convergence is in the strong topology and moreover,

$$\|f(A)\| \leqq K \max_{\lambda \in \sigma(A)} |f(\lambda) \tag{2.5}$$

where K is a constant independent of f.

An easy consequence of this theorem is that A is the infinitesimal generator of a strongly continuous semigroup.

THEOREM 3. *If A is a regular spectral operator with all but finitely many simple eigenvalues and* (2.3) *holds, then A generates a strongly continuous semigroup T_t provided that $\omega = \sup\{\operatorname{Re}\lambda ; \lambda \in \sigma(A)\} < \infty$. Moreover we have*

$$T_t = \sum_{i=1}^{\infty} e^{\lambda_i t} E(\lambda_i) \tag{2.6}$$

and

$$\|T_t\| \leqq K e^{\omega t}. \tag{2.7}$$

Proof. From Theorem 2 it follows that the resolvent $R(\lambda, A)$ is given by

$$R(\lambda, A) = \sum \frac{1}{\lambda - \lambda_i} E(\lambda_i) \tag{2.8}$$

and for all λ not within a radius ε of any multiple pole of the resolvent,

$$\|R(\lambda, A)^r\| \leqq K \max_{\mu \in \sigma(A)} |(\lambda - \mu)^{-r}|, \qquad r = 1, 2, \cdots. \tag{2.9}$$

Thus we obtain the estimate for real $\lambda > \omega$

$$\|R(\lambda, A)^r\| \leqq \frac{K}{(\lambda - \omega)^r}$$

which by the Hille–Yosida theorem [2] proves that A generates a strongly continuous semigroup T_t and

$$\|T_t\| \leqq K e^{\omega t}. \tag{2.10}$$

It remains to establish (2.6). From Theorem 2, we have that e^{At} is well-defined and is given by

$$e^{At} = \sum_{i=1}^{\infty} e^{\lambda_i t} E(\lambda_i). \tag{2.11}$$

We can take the Laplace transform for $\operatorname{Re}\lambda > \omega$

$$\begin{aligned} \int_0^{\infty} e^{-\lambda t} e^{At} z \, dt &= \sum_{i=1}^{\infty} \int_0^{\infty} e^{(\lambda_i - \lambda)t} E(\lambda_i) z \, dt \\ &= \sum_{i=1}^{\infty} \frac{1}{\lambda - \lambda_i} E(\lambda_i) z \\ &= R(\lambda, A) z \quad \text{also by Theorem 2.} \end{aligned}$$

But for $\operatorname{Re}\lambda > \omega$, we also know that

$$\int_0^{\infty} e^{-\lambda t} T_t z \, dt = R(\lambda, A) z \qquad ([2, \text{p. } 17]).$$

Thus we have for $\operatorname{Re}\lambda > \omega$ that

$$\int_p^{\infty} e^{-\lambda t} (e^{At} - T_t) z \, dt = 0,$$

and (2.6) now follows from the uniqueness of the Laplace transform. □

We remark that the estimate (2.7) means that the spectrum of A determines the decay constant of the semigroup; this is commonly called the spectrum determined growth assumption, and is *not* automatically satisfied (see [2, p. 74]).

We also remark that if we know that the resolvent of A contains the sector $S=\{\lambda \in C: |\arg \lambda| < \omega;\ \pi/2 < \omega < \pi\}$, then it follows from the proof that A generates an analytic semigroup (see [2, p. 40]).

It is of course useful to know when perturbations of a spectral operator retain the spectral property and in this respect we quote the following from [9].

THEOREM 4. *If A is a regular spectral operator and* (2.3) *holds and all but finitely many $E(\lambda_i)$ are one-dimensional, then $A+B$ is a regular spectral operator for all bounded perturbations B provided that*

$$\sum_{n=1}^{\infty} \frac{1}{d_n^2} < \infty, \quad \text{where } d_n = \inf_{\lambda \in \Sigma} |\lambda - \lambda_n| \text{ and } \Sigma = \sigma(A) - \{\lambda_n\}. \tag{2.12}$$

We see that these bounded perturbations retain the spectrum determined growth assumption, which is useful in applications.

Finally it is of interest to speculate on the nature of the projections $E(\lambda_i)$ and we take that simplest case where $\dim E(\lambda_i) = 1$ for all i. If A is normal, then the projections are orthogonal and we have

$$E(\lambda_i) = \langle \cdot, \phi_i \rangle \phi_i \tag{2.13}$$

where ϕ_i are the normalized eigenfunctions of A.

From Young [12], we know that if (2.3) holds there exists a biorthogonal sequence ψ_n such that $\langle \phi_n, \psi_n \rangle = \delta_{mn}$ and

$$E(\lambda_i) = \langle \cdot, \psi_i \rangle \phi_i. \tag{2.14}$$

Finally it is interesting to mention that any uniform Boolean algebra of projections in a Hilbert space can be reduced to a Boolean algebra of orthogonal projections by an inner automorphism $E \to D^{-1}ED$, where D and $D^{-1} \in \mathscr{L}(H)$ (Schwarz [9, p. 424]). The system theoretic interpretation of this is that we can replace (2.1) by the equivalent system

$$\dot{\tilde{z}} = \tilde{A}\tilde{z} + \tilde{B}u, \qquad \tilde{z}(0) = \tilde{z}_0 \tag{2.15}$$

where $\tilde{z} = D^{-1}z$, $\tilde{B} = D^{-1}B$ and $\tilde{A} = D^{-1}AD$. $\tilde{A}$ is a regular spectral operator with orthogonal eigenprojections. It is easily proved that it generates the semigroup $\tilde{T}_t = D^{-1}T_tD$ and that the growth constants of $\tilde{T}_t$ and T_t are identical. This fact has already been exploited by Sun in [10].

3. Main result. We consider the following linear system on a Hilbert space Z.

$$\dot{z} = \mathscr{A}z, \tag{3.1}$$

$$\tau z = u, \tag{3.2}$$

where $\mathscr{A}$ is a closed operator on Z and τ is a linear operator with $D(\mathscr{A}) \subseteq D(\tau)$ and the restriction of τ to $D(\mathscr{A})$ is continuous with respect to the graph norm of $\mathscr{A}$. Typically $\mathscr{A}$ is a partial differential operator and τ is a boundary operator. We suppose that $u(t)$ is the scalar control. We define the associated operator A on Z by

$$D(A) = \{z \in D(\mathscr{A}) / \tau z = 0\} \quad \text{and} \quad Az = \mathscr{A}z \quad \text{in } D(A) \tag{3.3}$$

and we assume that A is the infinitesimal generator of a strongly continuous semigroup

$T(t)$ on Z. Our final assumption is that there exists a $b \in D(\mathcal{A})$ so that

$$\tau(bu) = u \quad \text{for all } u \in R. \tag{3.4}$$

Under these assumptions $q = \mathcal{A}b \in Z$, and the following inhomogeneous system is well-defined

$$\dot{v} = Av - b\dot{u} + qu, \qquad v(0) = v_0 \tag{3.5}$$

and has the unique solution

$$v(t) = T(t)v_0 + \int_0^t T(t-s)qu(s)\,ds - \int_0^t T(t-s)b\dot{u}(s)\,ds, \tag{3.6}$$

provided $v_0 \in D(A)$ and $\dot{u}(s)$ is continuously differentiable. It is then easily verified that

$$z(t) = v(t) + bu(t) \tag{3.7}$$

is a solution of (3.1), (3.2), and conversely, with of course $v_0 = z_0 - bu(0)$. (We shall choose $u(0) = 0$.)

The above was first used by Fattorini in [6] and was recently utilized in [3] together with the following idea of an extended system on the state space $R \oplus Z$:

$$\dot{\tilde{v}} = \begin{pmatrix} 0 & 0 \\ q & A \end{pmatrix} \tilde{v} + \begin{pmatrix} I \\ -b \end{pmatrix} \tilde{u} = \tilde{A}\tilde{v} + \tilde{b}\tilde{u} \tag{3.8}$$

where

$$\tilde{u} = \dot{u} \quad \text{and} \quad \tilde{v} = \begin{pmatrix} u \\ v \end{pmatrix}.$$

Then we have that

$$z = (bI)\tilde{v} = \tilde{C}\tilde{v}. \tag{3.9}$$

We remark that (3.8) is now of the form (1.1) considered by Sun in [10] and we can apply his Theorem 1 to obtain conditions for feedback spectral assignability for (3.3). In his proof, however, he has assumed that the eigenprojections are orthogonal and we do not wish to do this. Following the remarks at the end of § 2 it is clear that the assumptions on A in Theorem 1 remain unchanged, but that b_k should actually read

$$b_k = \langle b, \psi_k \rangle \tag{3.10}$$

where ψ_k is the biorthogonal vector such that $\langle \phi_m, \psi_n \rangle = \delta_{mn}$.

With this modification we proceed to apply Theorem 1. The assumptions needed are F on $\tilde{A}$ or equivalently F on A and the following

(F4) $$0 \notin \sigma(A), \qquad \inf |\lambda_i| > 0,$$

$$\sum_{j=1}^{\infty} \frac{1}{|\lambda_j|^2} < \infty$$

together with (1.3) and (1.4) for $(\tilde{A}, \binom{I}{-b})$. The eigenvectors of $\tilde{A}$ are

$$\left\{ \tilde{\phi}_0 = \begin{pmatrix} 1 \\ h \end{pmatrix} \quad \text{and} \quad \tilde{\phi}_k = \begin{pmatrix} 0 \\ \phi_k \end{pmatrix}; k \geqq 1 \right\}$$

where ϕ_k are the eigenvectors of A and h is the solution of the boundary value problem

$$q + Ah = 0, \qquad h \in D(A). \tag{3.11}$$

The biorthogonal system for $\tilde{\phi}_k$ is given by

$$\tilde{\psi}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \tilde{\psi}_k = \begin{pmatrix} x_k \\ \psi_k \end{pmatrix}, \quad \text{where } x_k = \frac{1}{\bar{\lambda}_k}\langle q, \psi_k \rangle, \; k \geqq 1$$

and so

$$\tilde{b}_k = \left\langle \begin{pmatrix} I \\ -b \end{pmatrix}, \tilde{\psi}_k \right\rangle = \begin{cases} x_k - b_k, & k \geqq 1, \\ 1, & k = 0 \end{cases} \tag{3.12}$$

where the inner product is in $R \oplus Z$. We remark that if $q = 0$, then $h = 0$ and $\tilde{b}_0 = 1$, $\tilde{b}_k = -b_k$. We now state our main result.

THEOREM 5. *If the operator A satisfies assumptions* (F1)–(F4) *and the completeness assumption* (2.3), *and there exists an index set J such that* $\operatorname{Re} \lambda_k \geqq 0$, $k \in J$ *and* $\operatorname{Re} \lambda_k < 0$ *for $k \notin J$, then the system* (3.1), (3.2) *can be stabilized by integral state feedback if and only if*

$$\begin{aligned} &\text{(i)} && \tilde{b}_k \neq 0, \qquad k \in J, \\ &\text{(ii)} && \sum_{k \in J} \frac{|\operatorname{Re} \lambda_k|^2}{|\tilde{b}_k|^2} < \infty, \end{aligned} \tag{3.13}$$

where $\tilde{b}_k$ is defined by (3.12).

Proof. We apply Theorem 1 to our system (3.8). Under (F1)–(F4) and (3.13) there exists a $\binom{a}{g}$ in $R \oplus Z$ such that with the state feedback control

$$\tilde{u}(t) = \left\langle \begin{pmatrix} a \\ g \end{pmatrix}, \begin{pmatrix} u(t) \\ v(t) \end{pmatrix} \right\rangle = a u(t) + \langle g, v(t) \rangle \tag{3.14}$$

the eigenvalues of the closed loop system operator will be at λ_k, $k \notin J$ and for $k \in J$, $\gamma_k = \lambda_k - \operatorname{Re} \lambda_k - \varepsilon_k |\tilde{b}_k|$, where $\tilde{b}_k$ is given by (3.12). Provided $\varepsilon_k > 0$ and $\sum_{k \in J} \varepsilon_k^2 < \infty$, we have that $\operatorname{Re} \lambda_k = -\varepsilon_k |\tilde{b}_k| < 0 \; \forall k \in J$ and

$$\sum_{k \in J} \frac{|\nu_k - \lambda_k|^2}{|\tilde{b}_k|^2} \leqq 2 \sum_{k \in J} \left| \frac{\operatorname{Re} \lambda_k}{\tilde{b}_k} \right|^2 + 2 \sum_{k \in J} \varepsilon_k^2 < \infty.$$

It is clear that such a form for ν_k is also necessary.

The completeness assumption (2.3) on A implies the same on $\tilde{A}$ and so by Theorem 4 the closed loop operator

$$\tilde{A} + \begin{pmatrix} I \\ -b \end{pmatrix} \left\langle \begin{pmatrix} a \\ g \end{pmatrix}, \cdot \right\rangle \quad \text{on } R \oplus Z$$

satisfies the spectrum determined growth assumption and thus (3.8) is stabilized under the state feedback law (3.14) (the growth rate is zero). Since $z(t) = (b \;\; I)\tilde{v}$, we can say the same for the solution of (3.1), (3.2). It remains to interpret the control law (3.14) in terms of the original system. Writing it in terms of $u(t)$, we obtain

$$\frac{du}{dt} - a u(t) = \langle g, v(t) \rangle = \langle g, z(t) \rangle - \langle g, b \rangle u(t), \tag{3.15}$$

which has the solution

$$u(t) = C e^{\beta t} + \int_0^t e^{\beta(t-s)} \langle g, z(s) \rangle \, ds \tag{3.16}$$

where $\beta = a - \langle g, b \rangle$.

The obvious solution is to choose $C=0$ and so we have an integral feedback of the state z:

$$u(t)=\int_0^t e^{\beta(t-s)}\langle g, z(s)\rangle\, ds. \tag{3.17}$$

□

We remark that this stabilizability result does not guarantee exponential stabilizability with arbitrary decay rate. This will only be possible in the case that J is finite, since $\sum|\tilde{b}_k|^2<\infty$. We do have the following result, however.

COROLLARY 1. *If the operator A satisfies the assumptions* (F1)–(F4) *and the completeness assumption* (2.3) *and* $\operatorname{Re}\lambda_k \geqq -\alpha^2$ *for $k\in J$, a finite set, and* $\operatorname{Re}\lambda_k<-\alpha^2$ *for $h\notin J$, then the system* (3.1), (3.2) *can be stabilized by integral state feedback with decay rate $-\alpha^2$ provided that*

$$\tilde{b}_k\neq 0 \quad \text{for } k\in J. \tag{3.18}$$

If one is prepared to choose b as well as g and α, then the following result is a direct consequence of a result in Sun [10].

THEOREM 6. *If the operator A satisfies conditions* (F1)–(F4) *and the completeness assumption* (2.3) *and there exists an index set J such that* $\operatorname{Re}\lambda_k>0$, $k\in J$ *and* $\operatorname{Re}\lambda_k\leqq 0$, $k\in J$, $k\geqq 1$, *then in order that there should exist g, $b\in Z$ and $\alpha\in R$ such that the system* (3.1), (3.2) *under the integral feedback* (3.17) *be stable, a necessary and sufficient condition is that*

$$\sum_{k\in J}\operatorname{Re}\lambda_k<\infty.$$

As an illustration of the usefulness of the preceding results we examine a class of hyperbolic systems studied in [2, Example 2.16]. They have spectral operators whose eigenvectors are not orthogonal.

Example. Consider the following second order system;

$$\ddot{z}+\alpha\dot{z}+Az=0, \qquad z(0)=z_0, \quad \dot{z}(0)=z_1, \tag{3.19}$$

where α is a real constant, A is a positive, self-adjoint operator on a real Hilbert space, H, with compact resolvent. Then A has positive eigenvalues $\{\mu_n\}_{n=1}^{\infty}$ with $\mu_n\to\infty$ and orthonormal eigenvectors $\{e_n\}_{n=1}^{\infty}$ which form a basis for H. We suppose that the eigenvalues are distinct and the eigenspaces one-dimensional. Following [2], we reformulate (3.19) on the product space $X=D(A^{1/2})\oplus H$ with the inner product

$$\langle x, \bar{x}\rangle_X=\langle Ax_1, \bar{x}_1\rangle_H+\langle x_2, \bar{x}_2\rangle_H, \tag{3.20}$$

where

$$x=\begin{pmatrix}x_1\\ x_2\end{pmatrix}, \qquad \bar{x}=\begin{pmatrix}\bar{x}_1\\ \bar{x}_2\end{pmatrix}.$$

Then (3.19) becomes the following system on X

$$\dot{x}=\tilde{A}x, \qquad x(0)=x_0, \tag{3.21}$$

where

$$x=\begin{pmatrix}z\\ \dot{z}\end{pmatrix} \quad \text{and} \quad \tilde{A}=\begin{pmatrix}0 & I\\ -A & -\alpha I\end{pmatrix}.$$

$\tilde{A}$ has eigenvalues $\lambda_n=-\alpha/2\pm 1/2\sqrt{\alpha^2-4\mu_n}$, $n=1,\cdots$. For convenience we suppose that all λ_n are complex and we write

$$\lambda_n=-\alpha/2+i/2\sqrt{4\mu_n-\alpha^2} \quad \text{and} \quad \lambda_{-n}=\bar{\lambda}_n. \tag{3.22}$$

The corresponding eigenvectors are

$$\tilde{\phi}_n = \begin{pmatrix} e_n \\ \lambda_n e_n \end{pmatrix}, \qquad \tilde{\phi}_{-n} = \begin{pmatrix} e_n \\ \lambda_{-n} e_n \end{pmatrix},$$

and since $\{e_n\}_{n=1}^{\infty}$ is complete in H the set $\{\tilde{\phi}_n\}$ is complete in X. The eigenvectors of the adjoint operator $\tilde{A}^* = \left(\begin{smallmatrix} 0 & -I \\ A & -\alpha I \end{smallmatrix}\right)$ are

$$\tilde{\psi}_n = \begin{pmatrix} e \\ -\lambda_n e_n \end{pmatrix} \quad \text{and} \quad \tilde{\psi}_{-n} = \begin{pmatrix} e_n \\ -\lambda_n e_n \end{pmatrix}$$

respectively, and we have the following relations:

$$\langle \tilde{\phi}_n, \tilde{\psi}_n \rangle_X = 2\mu_n \delta_{mn}, \qquad \|\tilde{\phi}_n\| = \|\tilde{\psi}_n\| = \sqrt{2\mu_n}. \tag{3.23}$$

Choosing $\phi_n = 1/\sqrt{2\mu_n}\ \tilde{\phi}_n$ and $\psi_n = 1/\sqrt{2\mu_n}\ \tilde{\psi}_n$, we obtain the biorthogonal pair (ϕ_n, ψ_n) with $\langle \phi_n, \psi_m \rangle = \delta_{mn}$.

To establish that $\tilde{A}$ is a regular spectral operator it only remains to show that the spectral measures, $E(\lambda_n)$, are uniformly bounded in n. Since μ_n is simple, so is λ_n and

$$E(\lambda_n) = \langle \cdot, \psi_n \rangle_X \phi_n \quad \text{and so} \quad \|E(\lambda_n)\| \leqq \|\psi_n\| \, \|\phi_n\| = 1. \tag{3.24}$$

So $\tilde{A}$ is a regular spectral operator and assumption (F1) is satisfied. For (F2) and (F3) we note that

$$|\lambda_n - \lambda_m| = \tfrac{1}{2}|\sqrt{4\mu_n - \alpha^2} - \sqrt{4\mu_m - \alpha^2}| \tag{3.25}$$

and this yields readily verifiable conditions for (F2) and (F3) in terms of the eigenvalues of A. (F4) will be satisfied if

$$\alpha(\alpha^2 - 4\mu_n) \neq 0 \quad \text{for all } n, \qquad \sum_{n=0}^{\infty} \frac{1}{\mu_n} < \infty. \tag{3.26}$$

Condition (3.3) of Theorem 5 reduces to the controllability condition $\tilde{b}_k \neq 0$ and (ii) reduces to

$$\sum_{k \in J} 1/|\tilde{b}_k|^2 < \infty. \tag{3.27}$$

Since $\sum |\tilde{b}_k|^2 < \infty$, (3.27) will only hold if J is a finite set.

The formulation of the extended bounded system depends on the particular A operator and some examples of this are to be found in [3] and [4]. Here we consider the special case

$$\begin{aligned} &\frac{\partial^2 z}{\partial t^2} + \alpha \frac{\partial z}{\partial t} - \frac{\partial^2 z}{\partial x^2} = 0 \quad \text{on } H = L_2(0, 1), \\ &z(0, t) = 0, \qquad z(1, t) = u(t). \end{aligned} \tag{3.28}$$

Define

$$\mathscr{A} = -\frac{d^2}{dx^2} \quad \text{with } D(\mathscr{A}) = \{h \in H: \mathscr{A}h \in H;\ h(0) = 0\},$$

$$A = -\frac{d^2}{dx^2} \quad \text{with } D(A) = \{h \in H: \mathscr{A}h \in H;\ h(0) = 0 = h(1)\}.$$

Then (3.28) can be written as the system on $H \oplus H$:

$$\dot{x} = \tilde{\mathscr{A}}x, \qquad \tau x = u \tag{3.29}$$

where

$$\tilde{\mathscr{A}} = \begin{pmatrix} 0 & I \\ -\mathscr{A} & -\alpha I \end{pmatrix} \quad \text{and} \quad \tau \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1(1).$$

Then $\tilde{\mathscr{A}}$ and τ satisfy our assumptions and the following b and q in $X = D(A^{1/2}) \oplus H$ suffice

$$b = \begin{pmatrix} \xi \\ 0 \end{pmatrix}, \qquad q = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{3.30}$$

Condition (3.13) reduces here to $\int_0^1 \xi \sin k\pi\xi \, d\xi = 0$, which is always satisfied. So the essential condition for the theory is that there be finitely many unstable eigenvalues for this class of second order systems.

A similar analysis works for parabolic systems and for the flexible beam example of [3].

REFERENCES

[1] B. M. N. CLARKE AND D. WILLIAMSON, *Control canonical forms and eigenvalue assignment by feedback for a class of linear hyperbolic systems*, this Journal, 19 (1981), pp. 711–729.

[2] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences 8, Springer-Verlag, Berlin, 1978.

[3] R. F. CURTAIN, *Finite dimensional compensators for some hyperbolic systems with boundary control*, in Control Theory for Distributed Parameter Systems and Applications, F. Kappel, K. Kunisch and W. Schappacher, eds., Lecture Notes in Control and Information Sciences 54, Springer-Verlag, Berlin, 1983, pp. 77–91.

[4] ———, *On semigroup formulations of unbounded observations and control in distributed parameter systems*, Mathematical Theory of Networks and Systems, Proc. MTNS-83, Int. Symposium, Beersheva, Israel, Lecture Notes in Control and Information Sciences 58, Springer-Verlag, Berlin, 1984.

[5] N. DUNFORD AND J. SCHWARZ, *Linear Operators Part* III: *Spectral Operators*, Interscience, New York, 1971.

[6] H. O. FATTORINI, *Boundary control systems*, this Journal, 6 (1968), pp. 379–385.

[7] D. L. RUSSELL, *Control theory of hyperbolic equations related to certain questions in harmonic analysis and spectral theory*, J. Math. Anal. Appl., 40 (1972), pp. 336–368.

[8] ———, *Canonical forms and spectral determination for a class of hyperbolic distributed parameter control systems*, J. Math. Anal. Appl., 40 (1978), pp. 186–225.

[9] J. I. SCHWARTZ, *Perturbation on spectral operators and applications*, Pacific J. Math, 4 (1954), pp. 415–548.

[10] S. H. SUN, *On spectrum distribution of completely controllable linear systems*, Acta Mathematica Sinica, 21, pp. 193–205 (in Chinese); English translation, this Journal, 19 (1981), pp. 730–743.

[11] ———, *Boundary stabilization of hyperbolic systems with no dissipative conditions*, this Journal, 20 (1982), pp. 862–883.

[12] R. M. YOUNG, *An Introduction to Nonharmonic Fourier Series*, Academic Press, London, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 1, January 1985

013

APPROXIMATE STABILIZABILITY VIA THE ALGEBRAIC RICCATI EQUATION*

N. LEVAN†

Abstract. This note will study stabilizability of Hilbert space C_0 semigroups by means of a state feedback involving a solution of the algebraic Riccati equation. The notion of "approximate" stability is introduced and it is shown that stabilizability in this case is, in general, only approximate in the sense that the feedback semigroup is stable on a dense subspace—instead of on the whole space.

Key words. stabilizability, algebraic Riccati equation, quasi-affine transforms of contraction semigroups

1. Introduction. In this note we will be dealing with complex Hilbert spaces. Inner product and norm are denoted by $[\cdot, \cdot]$ and $\|\cdot\|$, respectively.

Let A be the generator of a C_0 semigroup $T(t)$, $t \geqq 0$, over a Hilbert space H. The algebraic Riccati equation (A.R.E.) [1] associated with the semigroup is the inner product equation:

$$[Ax, Px]+[Px, Ax]-[PBB^*Px, x]+[Rx, x]=0 \tag{1.1}$$

for x in the domain $\mathcal{D}(A)$ of A. Here B is a bounded linear operator from another Hilbert space to H, B^* is the adjoint of B, R is bounded linear, self-adjoint and nonnegative on H, $R \geqq 0$, and so is the operator solution P, $P \geqq 0$.

Now, since P is self-adjoint, (1.1) can be written as

$$2 \operatorname{Re}[PAx, x]=\|B^*Px\|^2-[Rx, x] \quad \text{for } x \text{ in } \mathcal{D}(A), \tag{1.2}$$

from which it follows that, for x in $\mathcal{D}(A)$:

$$2 \operatorname{Re}[P(A-BB^*P)x, x]=-\|B^*Px\|^2-[Rx, x] \leqq 0. \tag{1.3}$$

Thus if P is the identity operator I then $A-BB^*$ is dissipative; therefore, since A is a generator and $-BB^*$ is bounded linear, it generates a C_0 semigroup of contractions on H. This, of course, need not be the case when $P \neq I$.

Denote by $S(t)$, $t \geqq 0$, the semigroup generated by $A-BB^*P$, then, as we shall see, for x in H:

$$[PS(t)x, S(t)x] \leqq [Px, x] \quad \text{for } t \geqq 0. \tag{1.4}$$

Now if P is boundedly invertible then (1.4) implies that $S(t)$, $t \geqq 0$, is similar to a contraction semigroup on H. If P is only quasi-invertible (or quasi-affine), i.e., it has dense range and trivial kernel, then it follows from (1.4) that $S(t)$, $t \geqq 0$, is a quasi-affine transform [2] of a contraction semigroup.

This note will study stability of the semigroup $[S(t), t \geqq 0]$ and hence stabilizability of the original semigroup $T(t)$, $t \geqq 0$, using the state feedback $-B^*P$, where P is a solution of the A.R.E. The main results are given in § 2. First we obtain a condition for a solution P of the A.R.E. to be quasi-affine. Then we find sufficient conditions for a quasi-affine transform of a contraction semigroup to be weakly stable on a dense subspace. This is shown to be the "weakest" type of Lyapunov stability. Stabilizability

* Received by the editors April 12, 1983, and in revised form September 15, 1983. This research was supported by the U.S. Air Force Office of Scientific Research, Directorate of Mathematical and Information Sciences, under grant AFOSR 79-0053C.

† Department of System Science, University of California, Los Angeles, California 90024.

via the A.R.E. is then studied in some detail. The key result is that stabilizability via the A.R.E. is, in general, only weakly on a dense subspace, hence it is called "approximate stabilizability."

2. The main results. To begin with, we obtain from (1.3) the following important relation, for x in H and $t \geqq 0$:

$$[PS(t)x, S(t)x] - [Px, x] = -\int_0^t \|B^*PS(\sigma)x\|^2 \, d\sigma - \int_0^t [RS(\sigma)x, S(\sigma)x] \, d\sigma. \tag{2.1}$$

Therefore

$$[PS(t)x, S(t)x] \leqq [Px, x] \quad \text{for } x \text{ in } H \text{ and } t \geqq 0. \tag{2.2}$$

To proceed, we define the set

$$M(S) = \{x \text{ in } H: [PS(t)x, S(t)x] = [Px, x],\ t \geqq 0\}. \tag{2.3}$$

Then, since $[P - S(t)^*PS(t)] \geqq 0$, $t \geqq 0$, by (2.2), it is plain that x belongs to $M(S)$ if and only if it belongs to $\ker [P - S(t)^*PS(t)]$, $t \geqq 0$. Therefore $M(S)$ is a closed subspace and it is clearly invariant for $S(t)$, $t \geqq 0$. More is true; we have from (2.1)

$$x \text{ in } M(S) \Rightarrow B^*PS(t)x = 0 = RS(t)x \text{ for } t \geqq 0. \tag{2.4}$$

It follows from this

$$\text{for } x \text{ in } M(S) \cap \mathcal{D}(A): \quad Ax = (A - BB^*P)x.$$

Therefore, since $M(S) \cap \mathcal{D}(A)$ is dense in $M(S)$ by the fact that $M(S)$ is invariant for $S(t)$, $t \geqq 0$,

$$\text{for } x \text{ in } M(S): \quad T(t)x = S(t)x, \quad t \geqq 0. \tag{2.5}$$

This implies that $M(S)$ is also invariant for $T(t)$, $t \geqq 0$, and from (2.4) it is evident that

$$M(S) \subseteq \{x \text{ in } H: RT(t)x = 0,\ t \geqq 0\} = M_{uc}(A^*, R) \quad \text{(say)}. \tag{2.6}$$

We note that $\ker P$ is a subspace of $M(S)$ and it is also invariant for $S(t)$, $t \geqq 0$, by (2.2), hence for $T(t)$, $t \geqq 0$, by (2.5).

We summarize the above in

LEMMA 1. *Let $P \geqq 0$ be a solution of the algebraic Riccati equation and let $M(S)$ be as defined by* (2.3). *Then*

$$\ker P \subseteq M(S) \subseteq M_{uc}(A^*, R) \subseteq \ker R.$$

Moreover, $\ker P$ *and $M(S)$ are invariant subspaces of $S(t)$, $t \geqq 0$, and $T(t)$, $t \geqq 0$.*

The subspace $M_{uc}(A^*, R)$ is the uncontrollable subspace of the pair (A^*, R) [1]. Therefore if (A^*, R) is approximately controllable, i.e., $M_{uc}(A^*, R) = \{0\}$, then $\ker P$ and $M(S)$ are trivial. In finite dimension this will result in stability of the semigroup $S(t)$, $t \geqq 0$: $\|S(t)x\| \to 0$, $t \to \infty$, for each x in H. This is easily seen from (1.3) since, if $(A - BB^*P)\phi = \lambda\phi$ then $\operatorname{Re} \lambda \leqq 0$. But $\operatorname{Re} \lambda = 0$ implies that $[PS(t)\phi, S(t)\phi] = [P\phi, \phi]$ which contradicts the fact that $M(S)$ is trivial. Thus we only have eigenvalues with negative real part. Let $Q^2 = P$; then since Q^{-1} exists as soon as $P > 0$—in finite dimension, (2.2) becomes

$$\|QS(t)Q^{-1}x\| \leqq \|x\| \quad \text{for } x \text{ in } H \text{ and } t \geqq 0.$$

This shows that $S(t)$, $t \geqq 0$, is similar to the contraction semigroup $QS(t)Q^{-1}$, $t \geqq 0$. We summarize the above in

PROPOSITION 1. *Let H be a finite dimensional Hilbert space and $S(t)$, $t \geqq 0$, be the semigroup generated by $A - BB^*P$, where P satisfies the A.R.E.* (1.1). *If (A^*, R) is controllable then $P > 0$ and $S(t)$, $t \geqq 0$, is stable. Moreover, $S(t)$, $t \geqq 0$, is similar to the contraction semigroup $QS(t)Q^{-1}$, $t \geqq 0$, where $Q^2 = P$.*

If H is infinite dimensional and $P > 0$ then it does not necessarily follow that $S(t)$, $t \geqq 0$, is similar to a contraction semigroup. However we can still have $M(S) = \{0\}$ as soon as the pair (A^*, R) is controllable. This is the key idea in the following development.

We must note that if the semigroup $S(t)$, $t \geqq 0$, is similar to the contraction semigroup $QS(t)Q^{-1} = W(t)$ (say), $t \geqq 0$, then the subspace $M(S)$ becomes

$$M(W) = \{x \text{ in } H: \|W(t)x\| = \|x\|, t \geqq 0\} \tag{2.7}$$

which is invariant for $W(t)$, $t \geqq 0$, and $W(t)|M(W)$ is isometric. Hence requiring $M(W)$ to be trivial is equivalent to saying that the contraction semigroup $W(t)$, $t \geqq 0$, is completely nonisometric, and hence completely nonunitary also. The notion of a completely nonunitary contraction was introduced by Nagy and Foias in their study of model theory of Hilbert space contractions [2].

We recall that a C_0 semigroup $Z(t)$, $t \geqq 0$, on H is e(exponentially)-stable if $\|Z(t)\| \leqq M e^{-\alpha t}$ for some $M \geqq 1$ and $\alpha > 0$; s(strongly)-stable if for x in H: $\|Z(t)x\| \to 0$, $t \to \infty$; and w(weakly)-stable if for x and y in H: $[Z(t)x, y] \to 0$, $t \to \infty$.

We define

DEFINITION. A C_0 semigroup over H is *approximately* s-stable (resp. w-stable) if it is s-stable (resp. w-stable) on a dense subspace of H.

Let $Z(t)$, $t \geqq 0$, be a C_0 semigroup on H and suppose that there exists a bounded linear, self-adjoint and positive operator P on H, $P > 0$, such that

$$[PZ(t)x, Z(t)x] \leqq [Px, x] \quad \text{for } x \text{ in } H \text{ and } t \geqq 0. \tag{2.8}$$

Then, since the function $[PZ(t)x, Z(t)x]$ is nonincreasing, $\lim_{t\to\infty}[PZ(t)x, Z(t)x]$ always exists. We then have, for x in H and $s \geqq 0$:

$$\lim_{t\to\infty}[PZ(t+s)x, Z(t+s)x] = \lim_{t\to\infty}[PZ(t)x, Z(t)x],$$

or,

$$\lim_{t\to\infty}[(P - Z(s)^*PZ(s))Z(t)x, Z(t)x] = 0.$$

This, by the fact that $P - Z(s)^*PZ(s) \geqq 0$ for $s \geqq 0$, is equivalent to:

$$\lim_{t\to\infty}\|(P - Z(s)^*PZ(s))Z(t)x\| = 0, \quad \text{for } x \text{ in } H \text{ and } s \geqq 0. \tag{2.9}$$

Using (2.9) in the inequality

$$|[Z(t)^*(P - Z(s)^*PZ(s))x, y]| \leqq \|x\| \cdot \|(P - Z(s)^*PZ(s))Z(t)y\|,$$

we have

$$\text{for any } x \text{ and } y \text{ in } H: \quad \lim_{t\to\infty}[Z(t)^*(P - Z(s)^*PZ(s))x, y] = 0. \tag{2.10}$$

To proceed, we define

$$H_* = \{x \text{ in } H: \lim_{t\to\infty}[Z(t)^*x, y] = 0, y \text{ in } H\}, \tag{2.11}$$

which is a subspace of H and need not be closed. It then follows from (2.10) and

(2.11) that

$$\text{for } s \geqq 0: \quad [P - Z(s)^* P Z(s)]H \subset \bar{H}_* \quad \text{(the closure of } H_*).$$

This is equivalent to

$$H_*^{\perp} \subset \bigcap_{s \geqq 0} \ker [P - Z(s)^* P Z(s)] = M(Z),$$

where $H_*^{\perp}$ is the orthogonal complement in H of $\bar{H}_*$. We conclude that $M(Z) = \{0\} \Rightarrow \bar{H}_* = H$, and of course, for any x and y in H_*: $\lim_{t \to \infty} [Z(t)^* x, y] = 0 = \lim_{t \to \infty} [Z(t) y, x]$.

Let $Q^2 = P$ then (2.8) becomes

$$\|QZ(t)x\| \leqq \|Qx\| \quad \text{for } x \text{ in } H \text{ and } t \geqq 0. \tag{2.12}$$

Set

$$QZ(t)x = C(t)Qx \quad \text{for each } x \text{ in } H \text{ and } t \geqq 0. \tag{2.13}$$

Then it is evident from (2.12), and since $\mathcal{R}(Q)$ is dense, that $C(t)$, $t \geqq 0$, is a well-defined semigroup of contractions on H. The semigroup $Z(t)$, $t \geqq 0$, is therefore a quasi-affine transform of the contraction semigroup $C(t)$, $t \geqq 0$.

We conclude from the above that

THEOREM 1. *If a C_0 semigroup $Z(t)$, $t \geqq 0$, on H is a quasi-affine transform of a contraction semigroup, and its subspace $M(Z)$ is trivial then it is approximately weak stable.*

It follows at once from this theorem that

COROLLARY 1. (i) *If the semigroup $Z(t)$, $t \geqq 0$, is uniformly bounded, $\|Z(t)\| \leqq M$ $(\geqq 1)$ for $t \geqq 0$, then the conditions of the theorem are sufficient for it to be weakly stable.*

(ii) *The condition $M(Z) = \{0\}$ of the theorem can be replaced by: there exists $t_0 > 0$ such that $\|QZ(t_0)x\| < \|Qx\|$ for each x in H.*

Proof. The proof is all but trivial. For (i) we only have to observe that if $Z(t)$, $t \geqq 0$, is uniformly bounded then the subspace H_* (see (2.11)) is closed. For (ii) we note that, by assumption, for each x in H and $t \geqq 0$: $\|QZ(t + t_0)x\| \leqq \|QZ(t_0)x\| < \|Qx\| \Rightarrow M(Z) = \{0\}$. This finishes the proof of the corollary.

The stability result of Theorem 1 can be regarded as a Lyapunov type result. Indeed, for Hilbert space C_0 semigroups Datko gives the following extension of Lyapunov's stability result—in finite dimension:

THEOREM 2 (Datko [3]). *A C_0 semigroup $Z(t)$, $t \geqq 0$, with generator A in H is exponentially stable if and only if there exist on H two bounded linear and self-adjoint operators*: (i) $W \geqq kI$, $k > 0$, *i.e., W is strictly positive, and* (ii) $P > 0$, *such that*:

$$2 \operatorname{Re} [PAx, x] = -[Wx, x] \quad \text{for } x \text{ in } \mathcal{D}(A). \tag{2.14}$$

Suppose now that (2.14) holds for some W which is only positive, but is not strictly positive. Then, of course, the semigroup is not exponentially stable! Returning to (2.8) we see that it is equivalent to

$$2 \operatorname{Re} [PAx, x] \leqq 0 \quad \text{for } x \text{ in } \mathcal{D}(A). \tag{2.15}$$

This clearly is "weaker" than (2.14), even when W is only positive—which of course implies that $P > 0$. We conclude that if there exists $P > 0$ satisfying (2.14) in which W is *not* strictly positive, then (2.15) also holds and as a consequence the result of Theorem 1 applies. Theorem 1 is, perhaps, the "weakest" type of Lyapunov stability.

We are now ready to state our stabilizability results.

THEOREM 3. *Let $S(t)$, $t \geqq 0$, be the C_0 semigroup on H with generator $A - BB^*P$, where A generates a C_0 semigroup $T(t)$, $t \geqq 0$, on H, and $P \geqq 0$ is a solution of the A.R.E. If R is positive, or (A^*, R) is approximately controllable then*: (i) $P > 0$, *and* (ii) $S(t)$, $t \geqq 0$, *is approximately weak stable, i.e., $T(t)$, $t \geqq 0$, is approximately weak stabilizable.*

Proof. The proof is all but trivial. The conditions of the theorem imply that the subspace $M(S)$ (see (2.3)) is trivial and P is positive, by Lemma 1. Therefore $S(t)$, $t \geqq 0$, is approximately weak stable by Theorem 1.

Note that we also have from the A.R.E. (1.1), for $t \geqq 0$ and x in H:

$$[PT(t)x, T(t)x] - [Px, x] = -\int_0^t [RT(\sigma)x, T(\sigma)x]\, d\sigma + \int_0^t \|B^*PT(\sigma)x\|^2\, d\sigma.$$

Therefore

$$[PT(t)x, T(t)x] \leqq [Px, x] + \int_0^t \|B^*PT(\sigma)x\|^2\, d\sigma,$$

and

$$[Px, x] \leqq [PT(\sigma)x, T(\sigma)x] + \int_0^t [RT(\sigma)x, T(\sigma)x]\, d\sigma. \tag{2.16}$$

It follows from these inequalities that

$$x \text{ belongs to } M_{uc}(A^*, R) \cap \{x \text{ in } H\colon B^*PT(t)x = 0,\ t \geqq 0\}$$
$$\Rightarrow [PT(t)x, T(t)x] = [Px, x] \text{ for } t \geqq 0.$$

This shows that the set $M(T)$ of the semigroup $T(t)$, $t \geqq 0$, i.e.,

$$M(T) = \{x \text{ in } H\colon [PT(t)x, T(t)x] = [Px, x],\ t \geqq 0\} \tag{2.17}$$

(which need not be a subspace) satisfies:

$$M(T) \supseteq M_{uc}(A^*, R) \cap \{x \text{ in } H\colon B^*PT(t)x = 0,\ t \geqq 0\}. \tag{2.18}$$

This shows that controllability of the pair (A^*, R) has no effect on $M(T)$. It is interesting to observe from (2.5) and (2.17) that $M(S)$ is also a subspace of the set $M(T)$.

We now consider the extreme case in which $P = I$ is a solution of the A.R.E. (1.1). Then, as indicated in the Introduction, the semigroup $S(t)$, $t \geqq 0$, generated by $A - BB^*$ is a contraction semigroup, and $M(S)$ becomes

$$M(S) = \{x \text{ in } H\colon \|S(t)x\| = \|x\|,\ t \geqq 0\}.$$

Also, it follows from (2.4) that

$$x \text{ in } M(S) \Rightarrow B^*S(t)x = 0 = RS(t)x \quad \text{for } t \geqq 0. \tag{2.19}$$

Then as before we have $M(S) \subseteq M_{uc}(A^*, R)$. However, more is true in this case. It is evident from (2.19) and (2.5) that $M(S)$ is also contained in the subspace

$$\{x \text{ in } H\colon B^*T(t)x = 0,\ t \geqq 0\} = M_{uc}(A^*, B) \quad \text{(say)},$$

which is the uncontrollable subspace of the pair (A^*, B). We then conclude that

$$M(S) \subseteq M_{uc}(A^*, R) \cap M_{uc}(A^*, B).$$

Hence, as a corollary of Theorem 3, we have

COROLLARY 2. *Let A generate the C_0 semigroup $T(t)$, $t \geqq 0$, and*

$$[Ax, x] + [x, Ax] - \|B^*x\|^2 + [Rx, x] = 0 \quad \textit{for } x \textit{ in } \mathcal{D}(A).$$

Then $A-BB^*$ generates a contraction semigroup $S(t)$, $t\geqq 0$, which is weakly stable as soon as (A^*, R) or (A^*, B) is approximately controllable.

Finally, let us consider a generalization of Corollary 2. Let $P\neq I$ be a solution of the A.R.E.; what we ask is: "*When does $A-BB^*P$ generate a contraction semigroup*?" We know that if P is strictly positive: $P\geqq cI$ for some $c>0$, then it is boundedly invertible. Therefore the semigroup $S(t)$, $t\geqq 0$ (generated by $A-BB^*P$), is similar to a contraction semigroup, consequently it is uniformly bounded. This suggests that we should expect P to be, at least, strictly positive for $A-BB^*P$ to generate a contraction semigroup. But, if P is strictly positive it can be written as

$$P=cI+Q \quad \text{for some } c>0, \text{ and some } Q\geqq 0. \tag{2.20}$$

Substituting this in the A.R.E. we find, for x in $\mathcal{D}(A)$:

$$\begin{aligned}&[Ax, Qx]+[Qx, Ax]-\|B^*Qx\|^2+[Rx, x]\\&\quad+c\{2\,\mathrm{Re}\,[(A-BB^*Q)x, x]-c\|B^*x\|^2\}=0.\end{aligned} \tag{2.21}$$

Therefore if Q also satisfies the A.R.E. then, for x in $\mathcal{D}(A)$:

$$2\,\mathrm{Re}\,[(A-BB^*Q)x, x]=c\|B^*x\|^2,$$

or

$$2\,\mathrm{Re}\,[(A-BB^*P)x, x]=-c\|B^*x\|^2\leqq 0, \tag{2.22}$$

i.e., $A-BB^*P$ iş dissipative. Conversely, if (2.22) holds then it is clear from (2.21) that Q is a solution of the A.R.E. We obtain yet another corollary of Theorem 3.

COROLLARY 3. *If the pair (A^*, R) is approximately controllable and P is strictly positive then $A-BB^*P$ generates a weakly stable and uniformly bounded semigroup. In particular, if $P=cI+Q$ where $c>0$ and $Q\geqq 0$, then $A-BB^*P$ generates a contraction semigroup if and only if Q also satisfies the algebraic Riccati equation.*

We close by noting that if $P=I$ then Theorem 1 becomes a special case of the following result due to Foguel [4]:

THEOREM 4 (Foguel [4]). *If $T(t)$, $t\geqq 0$, is a contraction semigroup over H then for any x and y in H:*

$$\lim_{t\to\infty}[T(t)x, y]=0 \Leftrightarrow \lim_{t\to\infty}[T(t)^*x, y]=0.$$

Therefore

$$H_w(T)=\{x \text{ in } H: \lim_{t\to\infty}[T(t)x, y]=0, y \text{ in } H\}=H_w(T^*)$$

is a reducing subspace. Moreover, $T(t)|H_w(T)^\perp$ is unitary.

It follows at once from the above and from the Nagy–Foias canonical decomposition of contractions [2] that $H_w(T)^\perp$ is a subspace of the maximal unitary subspace of the semigroup, i.e., the subspace

$$H_u=\{x \text{ in } H: \|T(t)x\|=\|x\|=\|T(t)^*x\|, t\geqq 0\}.$$

This subspace in turn is a subspace of the isometric subspace

$$M(T)=\{x \text{ in } H: \|T(t)x\|=\|x\|, t\geqq 0\}.$$

Therefore if $M(T)$ is trivial then the semigroup is weakly stable. This certainly is the case when $P=I$ in Theorem 1. Indeed the proof of Theorem 1 follows that of Theorem 4.

Foguel's result is the key in stabilizing Hilbert space contraction semigroups [5]. Here we have obtained a method for stabilizing a class of quasi-affine transforms of contraction semigroups, in which the quasi-affinity comes from an algebraic Riccati equation. For exponential stabilizability and strong stabilizability via the algebraic Riccati equation we refer to the work of Zabczyk [6] and Balakrishnan [7].

We now close the note with some examples.

Examples.

1. We begin with a simple example showing that one can find a solution $P \geqq 0$ of the A.R.E. such that the semigroup $S(t)$, $t \geqq 0$, is stable on the range of P.

Let f be such that $Af = (i\omega)f$ and choose $B = f$, and $R = BB^*$. Then the A.R.E. becomes

$$[Ax, Px] + [Px, Ax] - |[f, Px]|^2 + |[f, x]|^2 = 0 \quad \text{for } x \text{ in } \mathscr{D}(A).$$

Thus, as in [7], this equation admits the nonnegative solution P given by:

$$Px = 0 \quad \text{for all } x \text{ orthogonal to } f,$$

and

$$Pf = f.$$

We then have

$$(A - BB^*P)f = (i\omega - \|f\|^2)f.$$

Therefore

$$S(t)f = e^{-\|f\|^2 t} e^{i\omega t} f \to 0, \qquad t \to \infty.$$

Hence the semigroup $S(t)$, $t \geqq 0$, is stable on the range $\mathscr{R}(P)$ of P.

2. To generalize the above example we now consider the case in which A is self-adjoint, and suppose that

$$Af_k = \lambda_k f_k, \qquad k = 0, 1, 2, \cdots,$$

where the λ_k's are real and nonnegative, and the sequence $\{f_k\}$ is an orthonormal basis.

Next, we define, for each x in H:

$$Bx = \sum_k b_k [x, f_k] f_k,$$

and

$$Rx = \sum_k r_k^2 [x, f_k] f_k,$$

where b_k's and r_k's are scalars. The A.R.E. now becomes, for x in $\mathscr{D}(A)$:

$$[Ax, Px] + [Px, Ax] - \sum_k |b_k|^2 |[Px, f_k]|^2 + \sum_k r_k^2 |[x, f_k]|^2 = 0.$$

To solve for P we set

$$Px = \sum_k p_k [x, f_k] f_k,$$

where the p_k's are to be determined. Substituting this in the A.R.E. we find

$$|b_k|^2 P_k^2 - 2\lambda_k p_k - r_k^2 = 0 \quad \text{for } k = 0, 1, 2, 3, \cdots.$$

Hence p_k is the positive root of this equation.

Finally, we have

$$(A - BB^*P)f_k = (\lambda_k - |b_k|^2 p_k)f_k, \qquad k = 0, 1, 2, \cdots.$$

But

$$\lambda_k - |b_k|^2 p_k = -\frac{r_k^2}{p_k} - \lambda_k < 0.$$

This shows that the semigroup $S(t)$, $t \geqq 0$, is stable on the span of $\{f_k\}$ which is dense in H. It is noted that the pair (A^*, R) is controllable in this example.

3. This example illustrates the results of Corollary 2. Consider

$$H = L^2[0, 2\pi], \qquad A = -\frac{\partial}{\partial z},$$

and

$$\mathcal{D}(A) = \{x \text{ in } H\colon x \text{ is absolutely continuous, } \dot{x} \text{ in } H \text{ and } x(0) = x(2\pi)\}.$$

Take B to be an element of H such that $[B(z), e^{inz}/\sqrt{2\pi}] \neq 0$ for all integers n, and take $R = BB^*$. We have, for x in $\mathcal{D}(A)$:

$$[Ax, x] + [x, Ax] - \|B^*x\|^2 + [BB^*x, x] = 0,$$

since $A = -A^*$. Moreover, for x in $\mathcal{D}(A)$:

$$[(A - BB^*)x, x] + [x, (A - BB^*)x] = -2\|B^*x\|^2 \leqq 0.$$

Therefore the semigroup $S(t)$, $t \geqq 0$, generated by $A - BB^*$ is contractive. Finally, it is plain that the pair (A^*, R) is controllable, since

$$RT(t)x = B[B, T(t)x] = B \sum_n e^{-int}[e^{inz}/\sqrt{2\pi}, x][B, e^{inz}/\sqrt{2\pi}]$$

$$= 0 \quad \text{for all } t \geqq 0$$

$$\Rightarrow x = 0.$$

Thus the semigroup $S(t)$, $t \geqq 0$, is stable as expected.

REFERENCES

[1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, second ed., Springer-Verlag, New York, 1981.

[2] B. SZ-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, American Elsevier, New York, 1970.

[3] R. DATKO, *Extending a theorem of A. M. Liapunov to Hilbert space*, J. Math. Anal. Appl., 32 (1970), pp. 610–616.

[4] S. R. FOGUEL, *Powers of a contraction in Hilbert space*, Pacific J. Math., 13 (1961), pp. 551–562.

[5] C. D. BENCHIMOL, *A note on weak stabilizability of contraction semigroups*, this Journal, 16 (1978), pp. 373–379.

[6] J. ZABCZYK, *Remarks on the algebraic Riccati equation in Hilbert space*, Appl. Math. Optim., 23 (1976), pp. 251–256.

[7] A. V. BALAKRISHNAN, *Strong stabilizability and the steady state Riccati equation*, Appl. Math. Optim., 7 (1981), pp. 335–345.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

VISCOSITY SOLUTIONS FOR THE MONOTONE CONTROL PROBLEM*

EMMANUEL N. BARRON†

Abstract. The value function, V, of the optimal control problem in which the controls must be monotone nondecreasing is a $W^{1,\infty}$ generalized solution of the quasi-variational inequality (QVI) $\max\{LV, V_y\} = 0$. Here $LV = V_t + f \cdot \nabla_x V + h$ and f, h the dynamics of the control problem. We extend the Crandall, Evans, Lions definition of viscosity solution for nonlinear first-order p.d.e.s. to systems like (QVI) and prove that V is the *unique* viscosity solution of (QVI). Further, V is the smallest function satisfying the inequalities $LV \leqq 0$, $V_y \leqq 0$.

Key words. monotone control functions, optimal control, Bellman equation, first order quasi-variational inequalities, viscosity solution

Introduction. In recent years M. G. Crandall and P. L. Lions [8] have mounted a formidable attack against unresolved problems in the theory of first order partial differential equations. In particular they have managed to prove the existence of a *unique* generalized solution to highly nonlinear first order equations by introducing the seminal idea of viscosity solutions which, as it turns out, is the same solution everybody else proved existed [6], [10], [12], [14]. The idea of viscosity solution was later reformulated and simplified in Crandall, Evans and Lions [9]. This paper presents an application of the idea of viscosity solution to the QVI arising in the monotone control problem. This problem is currently receiving renewed interest because of variational inequalities.

The central result of this paper is that the value function, $V(t, x, y)$, associated with the optimal control problem in which the controls must be monotone nondecreasing in [0, 1] is the unique viscosity solution of the quasi-variational inequality

$$\text{(QVI)} \qquad \max\{\partial V/\partial t + f \cdot \nabla_x V + h, \partial V/\partial y\} = 0.$$

This problem was first studied in Barron and Jensen [4] and is motivated by applications of optimal control in which one is controlling exhaustible resources (money, trees, oil, etc.). Accordingly, this problem is of substantial importance in economics and resource management.

We first extend the Crandall, Evans, Lions definition of "viscosity" solution to q.v.i.'s of the type (QVI). We prove that the value function then is, in fact, the viscosity solution. The key idea is the principle of optimality and has been used by Lions [14], Barron, Evans and Jensen [6] and others [10], [16]. Next we show that viscosity solutions can also be obtained by classical "vanishing viscosity" techniques for a penalized equation. Furthermore, using viscosity solution methods, we show that the solution V of (QVI) is the limit as $M \to \infty$ of the viscosity solution V^M of the equation $V_t^M + f \cdot \nabla_x V^M + h + M(V_y)^+ = 0$. This result is used in the next theorem where we prove that V is the smallest function satisfying $V_t + f \cdot \nabla_x V + h \leqq 0$, $V_y \leqq 0$.

Finally, we prove that V is the unique viscosity solution of (QVI) using the fundamental idea of Crandall, Lions and Evans. The proof is substantially simplified, however, because of the special nature of the problem at hand.

Similar results can be shown to hold for the control problem with controls of uniformly bounded variation [4] and the differential game with monotone controls [5].

* Received by the editors October 18, 1983, and in final form March 27, 1984. Part of this paper was completed while the author was at Bell Laboratories, Naperville, Illinois 60540.

† Department of Mathematical Sciences, Loyola University of Chicago, Chicago, Illinois 60626.

Also Barles [1] has obtained results for the first order obstacle problem. It is interesting to note that the value function for a monotone differential game satisfies a *system* of quasi-variational inequalities

$$V:[0, T]\times R^m\times[0, 1]\times[0, 1]\to R^1,$$

$$\min\left\{\frac{\partial V}{\partial z}, \max\left\{LV, \frac{\partial V}{\partial y}\right\}\right\}=0,$$

$$\max\left\{\frac{\partial V}{\partial y}, \min\left\{LV, \frac{\partial V}{\partial z}\right\}\right\}=0$$

a.e. in $[0, T)\times R^m\times[0, 1)\times[0, 1)$, with

$$V(T, x, y, z)=g(x), \quad V(t, x, 1, z)=\phi(t, x, z), \quad V(t, x, y, 1)=\psi(t, x, y).$$

1. The monotone value function. In this section we present the optimal control problem used as a model for the problems studied in this paper.

Consider the dynamical system defining a trajectory in R^m, $\xi=\xi(\tau)$, on the interval $0\leqq t\leqq\tau\leqq T$, $t<T$:

$$\text{(ODE)} \qquad \frac{d\xi}{d\tau}=f(\tau, \xi(\tau), \eta(\tau)), \qquad \xi(t)=x\in R^m.$$

We assume that $f:[0, T]\times R^m\times[0, 1]\to R^1$ is a given bounded function, uniformly Lipschitz continuous in the arguments (t, x, y) in $[0, T]\times R^m\times[0, 1]$.

The system (ODE) is controlled by choosing the control functions $\eta=\eta(\tau)$, $\eta:[t, T]\to[0, 1]$, which must be measurable and not decreasing. Given $y\in[0, 1]$ we define

$$Y_y[t, T]=\{\eta:[t, T]\to[0, 1]: \eta(t)=y, \eta(\cdot) \text{ not decreasing on } [t, T]\}.$$

A function η in Y_y is called a monotone control starting at y. Each control in Y_y gives a unique associated trajectory ξ on $[t, T]$ as the solution of (ODE).

The functions η are chosen from Y_y in order to maximize the payoff

$$P_{t,x,y}(\eta)=g(\xi(T))+\int_t^T h(s, \xi(s), \eta(s))\, ds$$

where ξ is the trajectory associated with η on $[t, T]$.

For the function $g:R^m\to R^1$ and $h:[0, T]\times R^m\times[0, 1]\to R^1$ we assume: g is bounded and uniformly Lipschitz; h is bounded, Lipschitz continuous uniformly in (t, x, y).

If we allow the initial conditions (t, x, y) of the problem to vary, we obtain the value function:

DEFINITION. $V:[0, T]\times R^m\times[0, 1]\to R^1$ given by

$$V(t, x, y)=\sup\{P_{t,x,y}(\eta);\ \eta\in Y_y[t, T]\}$$

is called the *monotone value function.* Any control $\eta^*\in Y_y$ satisfying $V(t, x, y)=P_{t,x,y}(\eta^*)$ is *monotone optimal.*

The following proposition summarizes the major properties of V. (See Barron and Jensen [4].)

PROPOSITION 1. (i)

$$V(T, x, y) = g(x), \qquad V(t, x, 1) = g(\xi_1(T)) + \int_t^T h(s, \xi_1(s), 1)\, ds,$$

where ξ_1 *is the trajectory corresponding to* $\eta(\tau) = 1$.

(ii) $V(\cdot, \cdot, y)$ *is not increasing in* y.

(iii) *For each* (t, x, y) *in* $[0, T] \times R^m \times [0, 1]$ *there is an optimal monotone control* $\eta^* \in Y_y[t, T]$: $V(t, x, y) = P_{t,x,y}(\eta^*)$.

(iv) *If* $\eta \in Y_y$ *is monotone optimal and* $\eta^+ = \eta(t+)$, *then* $V(t, x, y) = V(t, x, y^1)$ *if* $y \leqq y^1 \leqq \eta^+$.

(v) *If* $0 < \delta < T - t$ *and* $\eta \in Y_y$, $0 \leqq y \leqq 1$, *is monotone optimal then* $V(t, x, y) = \int_t^{t+\delta} h(s, \xi(s), \eta(s))\, ds + V(t+\delta, \xi(t+\delta), \eta(t+\delta))$, *where* ξ *is the trajectory corresponding to* η.

(vi) V *is bounded and uniformly Lipschitz continuous in* t, x, y. *Therefore* V *is differentiable almost everywhere and in* $W^{1,\infty}$.

Finally, using the proof in Barron and Jensen [4] we have

THEOREM 1. *The monotone value function satisfies the quasivariational inequality*

$$\text{(QVI)} \qquad \max\left\{LV(t, x, y), \frac{\partial V(t, x, y)}{\partial y}\right\} = 0 \quad \text{a.e. in } [0, T) \times R^m \times [0, 1),$$

where

$$LV = \frac{\partial V}{\partial t} + f \cdot \nabla_x V + h.$$

V *also satisfies*

$$V(T, x, y) = g(x), \qquad (x, y) \in R^m \times [0, 1]$$

and

$$V(t, x, 1) = g(\xi_1(T)) + \int_t^T h(s, \xi_1(s), 1)\, ds \quad \text{on } [0, T] \times R^m$$

where ξ_1 *is the trajectory corresponding to* $\eta(\tau) = 1$.

Throughout this paper we put $r(t, x) = g(\xi_1(T)) + \int_t^T h(s, \xi_1(s), 1)\, ds$ on $[0, T] \times R^m$. The function r is uniformly Lipschitz and $r(T, x) = g(x)$.

2. Viscosity solutions of (QVI). We will define the notion of a weak solution to (QVI) stemming from the idea of viscosity solutions to first order p.d.e.s. developed by Crandall, Evans, Lions [9]. We will prove below that the monotone value function is the unique viscosity solution of (QVI) and use viscosity methods to develop some additional results concerning its characterization and asymptotic properties.

Let BUC (Q) be the space of bounded, uniformly continuous functions on $Q = [0, T] \times R^m \times [0, 1]$.

DEFINITION. A function u in BUC (Q) is a viscosity solution of

$$\text{(QVI)} \qquad \max\left\{Lu, \frac{\partial u}{\partial y}\right\} = 0 \quad \text{in } Q$$

with $Lu = \partial u/\partial t + f \cdot \nabla_x u + h$, if and only if for each ψ in $C^1(Q)$

(i) if $u - \psi$ has a local max at $(t_0, x_0, y_0) \in Q$, then

$$\max\left\{L\psi(t_0, x_0, y_0), \frac{\partial \psi(t_0, x_0, y_0)}{\partial y}\right\} \geqq 0$$

and

(ii) if $u-\psi$ has a local min at $(t_0, x_0, y_0) \in Q$, then

$$\max\left\{L\psi(t_0, x_0, y_0), \frac{\partial\psi(t_0, x_0, y_0)}{\partial y}\right\} \leqq 0.$$

Finally,

$$u(T, x, y) = g(x), \qquad u(t, x, 1) = r(t, x).$$

An important property of this definition is that u may not have derivatives of any sort.

First we show that the monotone value function is, in fact, a viscosity solution. This will follow from Proposition 1(v).

THEOREM 2. *The monotone value function is a viscosity solution of* (QVI) *on* $[0, T]\times R^m \times[0, 1]$.

Proof. We already have V in BUC (Q) and $V(T, x, y) = g(x)$, $V(t, x, 1) = r(t, x)$ so only (i), (ii) are left to prove.

To prove (i) suppose $\psi \in C^1(Q)$ with $V-\psi$ attaining a max at (t_0, x_0, y_0) and $\psi_y(t_0, x_0, y_0) < 0$. We must show that

$$L\psi(t_0, x_0, y_0) = (\psi_t + f\cdot\nabla_x\psi + h)(t_0, x_0, y_0) \geqq 0.$$

We claim that if $\psi_y(t_0, x_0, y_0) < 0$, then if $\eta \in Y_{y_0}[t_0, T]$ is monotone optimal starting at y_0 and $\xi(\cdot)$ is the corresponding trajectory on $[t_0, T]$, $\xi(t_0) = x_0$, then $\eta^+ \equiv \eta(t_0^+) = \eta(t_0) = y_0$. Indeed, if $\eta^+ > y_0$, then there is $e > 0$ with $y_0 < y_0 + e < \eta^+$ and

$$\psi(t_0, x_0, y_0 + e) < \psi(t_0, x_0, y_0).$$

Since $V-\psi$ has a local max at (t_0, x_0, y_0), we can choose e small enough so that

$$V(t_0, x_0, y_0 + e) - V(t_0, x_0, y_0) \leqq \psi(t_0, x_0, y_0 + e) - \psi(t_0, x_0, y_0) < 0.$$

So

$$V(t_0, x_0, y_0 + e) < V(t_0, x_0, y_0).$$

But by Proposition 1(iv) $V(t_0, x_0, y_0) = V(t_0, x_0, y_0 + e) = V(t_0, x_0, \eta^+)$. Hence $\eta^+ = y_0$ if $\psi_y(t_0, x_0, y_0) < 0$ and $V-\psi$ has a local max at (t_0, x_0, y_0).

Let $\xi^+(\cdot)$ be the trajectory on $[t_0, T]$ corresponding to the constant monotone control η^+ in $Y_{\eta^+}[t_0, T]$ with $\xi^+(t_0) = x_0$. Suppose that $L\psi(t_0, x_0, y_0) \leqq -C < 0$ for some constant $C > 0$. Then for each $0 < \delta < T - t_0$, since $\eta^+ = \eta(t_0^+) = y_0$ and $\psi \in C^1$ we have for δ sufficiently small

$$\begin{aligned}
&\psi(t_0 + \delta, \xi^+(t_0 + \delta), \eta^+) - \psi(t_0, x_0, y_0) \\
&\quad = \int_{t_0}^{t_0+\delta} \frac{d}{ds}\psi(s, \xi^+(s), \eta^+)\, ds \\
&\quad = \int_{t_0}^{t_0+\delta} (\psi_t + f\cdot\nabla_x\psi)(s, \xi^+(s), \eta^+)\, ds \\
&\quad = \int_{t_0}^{t_0+\delta} L\psi(s, \xi^+(s), \eta^+)\, ds - \int_{t_0}^{t_0+\delta} h(s, \xi^+(s), \eta^+)\, ds \\
&\quad \leqq -C\delta - \int_{t_0}^{t_0+\delta} h(s, \xi^+(s), \eta^+)\, ds.
\end{aligned}$$

Since $V-\psi$ has a local max at (t_0, x_0, y_0), we then have

$$V(t_0+\delta, \xi^+(t_0+\delta), \eta^+) - V(t_0, x_0, y_0) \leqq -C\delta - \int_{t_0}^{t_0+\delta} h(s, \xi^+(s), \eta^+)\, ds.$$

By Gronwall's inequality it readily follows that $\|\xi^+(s)-\xi(s)\|_\infty = o(\delta)$ if $t_0 \leqq s \leqq t_0+\delta$. Using the continuity of V and h we get for some constant c independent of δ

$$V(t_0+\delta, \xi(t_0+\delta), \eta(t_0+\delta)) - V(t_0, x_0, y_0) \leqq -c\delta - \int_{t_0}^{t_0+\delta} h(s, \xi(s), \eta(s))\, ds.$$

So

$$V(t_0+\delta, \xi(t_0+\delta), \eta(t_0+\delta)) + \int_{t_0}^{t_0+\delta} h(s, \xi(s), \eta(s))\, ds < V(t_0, x_0, y_0)$$

for small δ. This contradicts the principle of optimality 1(v) and thus $L\psi(t_0, x_0, y_0) \geqq 0$.

To prove (ii) holds let $\psi \in C^1$ with $V-\psi$ attaining a local min at (t_0, x_0, y_0). We must show that

$$\psi_y(t_0, x_0, y_0) \leqq 0, \qquad L\psi(t_0, x_0, y_0) \leqq 0.$$

Suppose that $\psi_y(t_0, x_0, y_0) > 0$. Then for some small $e > 0$ $\psi(t_0, x_0, y_0+e) > \psi(t_0, x_0, y_0)$. Since $V-\psi$ has a min at (t_0, x_0, y_0),

$$V(t_0, x_0, y_0) - V(t_0, x_0 y_0+e) \leqq \psi(t_0, x_0, y_0) - \psi(t_0, x_0, y_0+e) < 0.$$

which contradicts the fact that V is not increasing in y. Hence $\psi_y(t_0, x_0, y_0) \leqq 0$.

Now suppose $L\psi(t_0, x_0, y_0) \geqq c > 0$. Let $\bar{\xi}(\cdot)$ on $[t_0, T]$ be the trajectory corresponding to the constant control y_0 in $Y_{y_0}[t_0, T]$ with $\bar{\xi}(t_0) = x_0$. Then for small $0 < \delta < T - t_0$ we have

$$\begin{aligned}
\psi(t_0+\delta, \bar{\xi}(t_0+\delta), y_0) - \psi(t_0, x_0, y_0) &= \int_{t_0}^{t_0+\delta} \frac{d}{ds}\psi(s, \bar{\xi}(s), y_0)\, ds \\
&= \int_{t_0}^{t_0+\delta} L\psi(s, \bar{\xi}(s), y_0)\, ds - \int_{t_0}^{t_0+\delta} h(s, \bar{\xi}(s), y_0)\, ds \\
&\geqq c\delta - \int_{t_0}^{t_0+\delta} h(s, \bar{\xi}(s), y_0)\, ds.
\end{aligned}$$

Since $V-\psi$ has local min at (t_0, x_0, y_0), we have

$$V(t_0+\delta, \bar{\xi}(t_0+\delta), y_0) - V(t_0, x_0, y_0) \geqq c\delta - \int_{t_0}^{t_0+\delta} h(s, \bar{\xi}(s), y_0)\, ds$$

for small δ. Thus

$$V(t_0+\delta, \bar{\xi}(t_0+\delta), y_0) + \int_{t_0}^{t_0+\delta} h(s, \bar{\xi}(s), y_0)\, ds > V(t_0, x_0, y_0)$$

which is a contradiction of the fact that

$$\begin{aligned}
V(t_0, x_0, y_0) &= \sup_{\eta \in Y_{y_0}[t_0, t_0+\delta]} \left\{ \int_{t_0}^{t_0+\delta} h(s, \xi(s), \eta(s))\, ds + V(t_0+\delta, \xi(t_0+\delta), \eta(t_0+\delta)) \right\} \\
&\geqq \int_{t_0}^{t_0+\delta} h(s, \bar{\xi}(s), y_0)\, ds + V(t_0+\delta, \bar{\xi}(t_0+\delta), y_0).
\end{aligned}$$

Hence $L\psi(t_0, x_0, y_0) \leqq 0$ and our proof is complete.

3. Viscosity solutions by viscosity methods. In this section we will show that a viscosity solution of (QVI) can be obtained as the limit of the solution to a second order "viscous" parabolic equation.

To this end, consider the backward parabolic problem for $u_\varepsilon = u_\varepsilon(t, x, y)$ on Q:

$$(\text{PDE})_\varepsilon \qquad \begin{aligned} &\frac{\partial u_\varepsilon}{\partial t} + \varepsilon \Delta_{xy} u_\varepsilon + f(t, x, y) \cdot \nabla_x u_\varepsilon + h(t, x, y) + \frac{1}{\varepsilon}\left(\frac{\partial u_\varepsilon}{\partial y}\right)^+ = 0, \\ &u_\varepsilon(T, x, y) = g(x), \qquad u_\varepsilon(t, x, 1) = \gamma_\varepsilon(t, x) \end{aligned}$$

where

$$a^+ = \max(a, 0), \qquad \gamma_\varepsilon(t, x) \equiv E_{t,x}\left\{ g(\xi_1^\varepsilon(T)) + \int_t^T h(s, \xi_1^\varepsilon(s), 1)\, ds \right\}$$

and ξ_1^ε solves the Itô equation $d\xi_1^\varepsilon = f(\tau, \xi_1^\varepsilon(\tau), 1)\, d\tau + \sqrt{2\varepsilon}\, dw(\tau)$; w is an m-dimensional Brownian motion on $[t, T]$ and $\xi_1^\varepsilon(t) = x$.

By standard results there is a unique smooth solution of $(\text{PDE})_\varepsilon$ for each $\varepsilon > 0$.

LEMMA. *There are constants C_1, C_2 independent of ε so that*

$$\|u_\varepsilon\|_{L^\infty} \leqq C_1 \quad \textit{and} \quad \|Du_\varepsilon\|_{L^\infty} \leqq C_2.$$

Sketch of proof. Compare the solution u_ε of $(\text{PDE})_\varepsilon$ to the solutions of the equation for $W_\varepsilon = W_\varepsilon(t, x)$

$$\frac{\partial W_\varepsilon}{\partial t} + \varepsilon \Delta_x W_\varepsilon + \min_{0 \leqq y \leqq 1} \{ f(t, x, y) \cdot \nabla_x W_\varepsilon + h(t, x, y) \} = 0$$

and for $Z_\varepsilon = Z_\varepsilon(t, x)$

$$\frac{\partial Z_\varepsilon}{\partial t} + \varepsilon \Delta_x Z_\varepsilon + \max_{0 \leqq y \leqq 1} \{ f(t, x, y) \cdot \nabla_x Z_\varepsilon + h(t, x, y) \} = 0$$

with $\psi(T, x) = g(x)$, $\psi = W_\varepsilon, Z_\varepsilon$. The conclusion of the lemma holds for both W_ε and Z_ε and we have from the maximum principle that $W_\varepsilon(t, x) \leqq u_\varepsilon(t, x, y) \leqq Z_\varepsilon(t, x)$. (Note that $\gamma_\varepsilon(t, x) \to r(t, x)$ and $W_\varepsilon(t, x) \leqq \gamma_\varepsilon(t, x) \leqq Z_\varepsilon(t, x)$ by standard stochastic control results (Friedman [11]).)

Hence $\|u_\varepsilon\|_{L^\infty} \leqq C_1$.

For the second part of the proof, approximate $(\cdot)^+$ by smooth functions and differentiate $(\text{PDE})_\varepsilon$ with respect to x to bound $D_x u_\varepsilon$. The details are a modification of those in Barron, Evans and Jensen [6] and are omitted.

By the lemma, there is a subsequence, denoted $\{u_\varepsilon\}$ and a function u in BUC (Q) so that $u_\varepsilon \to u$ uniformly on $[0, T] \times R^m \times [0, 1]$.

THEOREM 3. *u is a viscosity solution of* (QVI).

Proof. Clearly $u(T, x, y) = g(x)$ and $u(t, x, 1) = \lim_{\varepsilon \to 0} \gamma_\varepsilon(t, x) = r(t, x)$.

Now, let $\psi \in C^2$ so that $u - \psi$ has a (strict) local max at (t_0, x_0, y_0). Since $u_\varepsilon \to u$ there is a sequence $(t_\varepsilon, x_\varepsilon, y_\varepsilon) \to (t_0, x_0, y_0)$ and, if ε is small enough, $(t_\varepsilon, x_\varepsilon, y_\varepsilon)$ is a local max of $u_\varepsilon - \psi$. Since u_ε is smooth we have $Du_\varepsilon = D\psi$ and $\Delta_{x,y} u_\varepsilon \leqq \Delta_{x,y} \psi$ at $(t_\varepsilon, x_\varepsilon, y_\varepsilon)$. Also, if $\psi_y(t_0, x_0, y_0) < 0$ then $\psi_y(t_\varepsilon, x_\varepsilon, y_\varepsilon) \leqq 0$ for small ε. Then $(1/\varepsilon)(\psi_y)^+ = 0$ and, using $(\text{PDE})_\varepsilon$ we get

$$(\psi_t + \varepsilon \Delta \psi + f \cdot \nabla_x \psi + h)(t_\varepsilon, x_\varepsilon, y_\varepsilon) \geqq 0.$$

Letting $\varepsilon \to 0$ we conclude that

$$(\psi_t + f \cdot \nabla_x \psi + h)(t_0, x_0, y_0) \geqq 0$$

if $\psi_y(t_0, x_0, y_0) < 0$. Hence in any case we have

$$\max\{L\psi, \psi_y\}(t_0, x_0, y_0) \geqq 0.$$

If ψ is only C^1, we approximate by C^2 functions and the same conclusion obtains.

Next, let $\psi \in C^2$ with $u - \psi$ attaining a (strict) local min at (t_0, x_0, y_0). Let $(t_\varepsilon, x_\varepsilon, y_\varepsilon) \to (t_0, x_0, y_0)$ with $u_\varepsilon - \psi$ attaining a local min at $(t_\varepsilon, x_\varepsilon, y_\varepsilon)$ for small ε. Then $Du_\varepsilon = D\psi$ and $\Delta_{x,y} u_\varepsilon \geqq \Delta_{x,y}\psi$ at $(t_\varepsilon, x_\varepsilon, y_\varepsilon)$. Since $(\cdot)^+ \geqq 0$, we have then that

$$(\psi_t + \varepsilon\Delta\psi + f \cdot \nabla_x \psi + h)(t_\varepsilon, x_\varepsilon, y_\varepsilon) \leqq 0.$$

Letting $\varepsilon \to 0$ we get $(\psi_t + f \cdot \nabla_x \psi + h)(t_0, x_0, y_0) \leqq 0$.

To see that $\psi_y(t_0, x_0, y_0) \leqq 0$, we have

$$\psi_t + \varepsilon\Delta\psi + f \cdot \nabla_x \psi + h \leqq \psi_t + \varepsilon\Delta\psi + f \cdot \nabla_x \psi + \frac{1}{\varepsilon}(\psi_y)^+ + h \leqq 0$$

at $(t_\varepsilon, x_\varepsilon, y_\varepsilon)$ for small ε. Thus $\limsup (1/\varepsilon)(\psi_y)^+(t_\varepsilon, x_\varepsilon, y_\varepsilon) \leqq 0$ so $\psi_y(t_0, x_0, y_0) \leqq 0$. Finally, if ψ is only C^1, approximate by C^2 functions.

In the next theorem we will prove that a solution to (QVI) can be obtained as the limit of a first order p.d.e. This is the penalty method for variational inequalities. The p.d.e. is here interpreted as the Hamilton-Jacobi-Bellman equation for the optimal control problem in which the controls must be monotone *and* Lipschitz with a fixed Lipschitz constant M. We will use the result obtained here to characterize viscosity solutions of (QVI) and to prove uniqueness. Define

$$\bar{f}: [0, T] \times R^m \times R^1 \to R^1 \quad \text{and} \quad \bar{h}: [0, T] \times R^m \times R^1 \to R^1$$

by $\bar{f}(t, x, y) = f(t, x, \pi y)$, $\bar{h}(t, x, y) = h(t, x, \pi y)$, where π is the projection of y on $[0, 1]$. Then $\bar{f}$, $\bar{h}$ are uniformly Lipschitz with the same Lipschitz constants as f, h respectively. If $\bar{V}$ is the value of the monotone problem with f, h replaced by $\bar{f}$, $\bar{h}$, then $\bar{V}(t, x, y) = V(t, x, y)$ if $y \in [0, 1]$ and $\bar{V}$ is an extension of V to all of $[0, T] \times R^m \times R^1$.

Let $M > 0$ and consider the first order p.d.e. for $V^M(t, x, y)$, $(t, x, y) \in [0, T) \times R^m \times R^1$:

$$(\text{PDE})_M \qquad \begin{aligned} &\frac{\partial V^M}{\partial t} + \bar{f} \cdot \nabla_x V^M + \bar{h} + M\left(\frac{\partial V^M}{\partial y}\right)^+ = 0, \\ &V^M(T, x, y) = g(x), \quad x \in R^m, \quad y \in R^1. \end{aligned}$$

Then, by Lions [14], there is a unique viscosity solution of $(\text{PDE})_M$, say V^M; $V^M \in W^{1,\infty}$ and V^M satisfies

$$V^M(t, x, y) = \sup\{\bar{P}_{t,x,y}(u); 0 \leqq u(\tau) \leqq M\}$$

(with $\bar{P}$ the same as P except $h = \bar{h}$) and on $t < \tau \leqq T$

$$\frac{d\bar{\xi}}{d\tau} = \bar{f}(\tau, \bar{\xi}, \bar{\eta}), \qquad \bar{\xi}(t) = x \in R^m$$

$$\frac{d\bar{\eta}}{d\tau} = u(\tau), \qquad \bar{\eta}(t) = y \in R^1.$$

Note that $V^M(t, x, 1) = r(t, x)$ on $[0, T] \times R^m$. (See Barron and Jensen [4].)

Remark. By "viscosity solution" of $(\text{PDE})_M$ is meant the following:

(i) $\forall \psi \in C^1$ if $V^M - \psi$ has a local max at (t_0, x_0, y_0) then

$$\psi_t + \bar{f} \cdot \nabla_x \psi + \bar{h} + M(\psi_y)^+ \geqq 0 \quad \text{at } (t_0, x_0, y_0)$$

and

(ii) $\forall \psi \in C^1$ with $V^M - \psi$ attaining a local min at (t_0, x_0, y_0) we have

$$\psi_t + \bar{f} \cdot \nabla_x \psi_t + \bar{h} + M(\psi_y)^+ \leqq 0 \quad \text{at } (t_0, x_0, y_0)$$

and $V^M \in \text{BUC}([0, T] \times R^m \times R^1)$, $V^M(T, x, y) = g(x)$.

The following lemma follows directly by optimal control methods (Lions [14]) or by p.d.e. methods (cf. Barron, Evans and Jensen [6]).

LEMMA. *$|D_t V^M|$, $|D_x V^M|$, $|D_y V^M| \leqq K$ with the constant K independent of M.*

THEOREM 4. *Let $V^M(t, x, y)$ be the unique viscosity solution of $(\text{PDE})_M$. Then $V(t, x, y) \equiv \lim_{M \to \infty} V^M(t, x, y)$ is a viscosity solution of*

$$\max(LV(t, x, y), V_y(t, x, y)) = 0 \quad \text{on } [0, T) \times R^m \times [0, 1),$$

$$V(T, x, y) = g(x), \qquad V(t, x, 1) = r(t, x), \tag{QVI}$$

$$LV \equiv V_t + f \cdot \nabla_x V + h.$$

Proof. Since V^M is bounded and Lipschitz independently of M, there is a sequence $M_j \to \infty$ and a uniformly bounded and Lipschitz function V such that $V^{M_j} \to V$ locally uniformly. Note that if $M \leqq M^1$, $V^M \leqq V^{M^1}$ by Lions [14, Thm. 1.4]. From the equation which V^M satisfies $(\text{PDE})_M$ we get

$$M(V_y^M)^+ \leqq |V_t^M| + |f||\nabla_x V^M| + |h|$$

so from the lemma

$$0 \leqq (V_y^{M_j})^+ \leqq \frac{C}{M_j}.$$

Letting $M_j \to \infty$, we must have that $V_y(t, x, y) \leqq 0$ a.e.

We will prove that V is a viscosity solution of (QVI).

Suppose $V - \psi$ has a local max at (t_0, x_0, y_0) and that ψ is C^1 with $\psi_y(t_0, x_0, y_0) < 0$. In fact, we may assume (t_0, x_0, y_0) is a strict local max. Since $V^{M_j} \to V$, there is a sequence (t_j, x_j, y_j), so that for large j, $V^{M_j} - \psi$ has a local max at (t_j, x_j, y_j) and $(t_j, x_j, y_j) \to (t_0, x_0, y_0)$.

Now, V^{M_j} is the viscosity solution of $(\text{PDE})_{M_j}$, so, by definition we must have

$$\psi_t + f \cdot \nabla_x \psi + M_j(\psi_y)^+ + h \geqq 0 \quad \text{at } (t_j, x_j, y_j).$$

Also, if $\psi_y(t_0, x_0, y_0) < 0$, then $\psi_y(t_j, x_j, y_j) \leqq 0$ for large enough j since ψ is C^1. Therefore $(\psi_y)^+ = 0$ at (t_j, x_j, y_j) and we have

$$(\psi_t + f \cdot \nabla_x \psi + h)(t_0, x_0, y_0) = \lim_{j \to \infty} (\psi_t + f \cdot \nabla_x \psi + h)(t_j, x_j, y_j) \geqq 0.$$

On the other hand, suppose $V - \psi$ has a (strict) local min at (t_0, x_0, y_0) with $\psi \in C^1$. Then, for large j, $V^{M_j} - \psi$ has a local min at $(t_j, x_j, y_j) \to (t_0, x_0, y_0)$. Since V^{M_j} is the viscosity solution of $(\text{PDE})_{M_j}$, this implies that

$$\psi_t + f \cdot \nabla_x \psi + M_j(\psi_y)^+ + h \leqq 0 \quad \text{at } (t_j, x_j, y_j).$$

Since $M_j(\psi_y)^+ \geqq 0$, we have

$$(\psi_t + f \cdot \nabla_x \psi + h)(t_0, x_0, y_0) \leqq \lim_{j \to \infty} (\psi_t + f \cdot \nabla_x \psi + h)(t_j, x_j, y_j) \leqq 0.$$

We must also show that $\psi_y(t_0, x_0, y_0) \leqq 0$. To see this, note that

$$(\psi_t + f \cdot \nabla_x \psi + h)(t_0, x_0, y_0) \leqq \overline{\lim_j} (\psi_t + f \cdot \nabla_x \psi + M_j(\psi_y)^+ + h)(t_j, x_j, y_j) \leqq 0.$$

Therefore

$$M_j \psi_y^+(t_j, x_j, y_j) \leqq C \quad \text{for large } j.$$

We conclude by letting $j \to \infty$ that $[\psi_y(t_0, x_0, y_0)]^+ = 0$ and we are done.

Remark. The proof used here is based on ideas from [6].

Remark. It is easy to prove that $\lim_{M\to\infty} V^M(t, x, y)$ is the monotone value function (on $[0, T] \times R^M \times [0, 1]$) directly by optimal control methods. See Barron and Jensen [4, Thm. 4.7].

THEOREM 5. *Let $u(t, x, y)$ be a viscosity solution of*

$$\max\left\{\frac{\partial u}{\partial t} + \bar{f} \cdot \nabla_x u + \bar{h}, \frac{\partial u}{\partial y}\right\} \leqq 0$$

and $u(T, x, y) = g(x)$. Then $\bar{V}(t, x, y) = \lim_{M\to\infty} V^M(t, x, y)$ on $[0, T] \times R^m \times R^1$ satisfies $\bar{V}(t, x, y) \leqq u(t, x, y)$ on $[0, T] \times R^m \times R^1$.

Proof. By definition, u is also a viscosity solution of

$$\frac{\partial u}{\partial t} + \bar{f} \cdot \nabla_x u + M\left(\frac{\partial u}{\partial y}\right)^+ + \bar{h} \leqq 0$$

for any $M \geqq 0$. Hence, by the uniqueness result of Crandall and Lions [8, Thm. V.2] for p.d.e.s. we have

$$\|(V^M - u)^+\|_\infty \leqq 0 \qquad (\text{since } V^M(T, x, y) = u(T, x, y)).$$

That is $V^M \leqq u$. Now let $M \to \infty$ and use the preceding theorem and the fact that $\bar{V}$ is an extension of V.

Remark. A viscosity solution of $\max\{Lu, u_y\} \geqq 0$ is a viscosity supersolution of $\max\{Lu, u_y\} = 0$. A viscosity solution of $\max\{Lu, u_y\} = 0$ is both a viscosity subsolution and supersolution.

4. Uniqueness of viscosity solutions to (QVI). In this section we will prove that V, the monotone value function is the only viscosity solution to (QVI). The proof will be based on ideas from Crandall and Lions [8] and Crandall, Evans and Lions [9], but it is substantially simpler due to the linear nature of the operators $L = \partial/\partial t + f \cdot \nabla_x + h$ and $\partial/\partial y$, the convex nonlinearity in (QVI) and the ability to draw on the Crandall and Lions theorem used here in Theorem 5.

THEOREM 6. *Let u be any viscosity solution of* (QVI) *on $[0, T) \times R^m \times [0, 1)$ with terminal condition $u(T, x, y) = g(x)$ and boundary condition $u(t, x, 1) = r(t, x)$, $(t, x) \in [0, T] \times R^m$. Then $u(t, x, y) \equiv V(t, x, y)$ on $[0, T] \times R^m \times [0, 1]$.*

Proof. First suppose V is a smooth function.

Since u is a viscosity solution, it is also a supersolution. By Theorem 5 we have $u(t, x, y) \geqq V(t, x, y)$ on $[0, T] \times R^M \times [0, 1]$. Suppose

$$\sup\{u(t, x, y) - V(t, x, y); (t, x, y) \in [0, T] \times R^m \times [0, 1]\} \equiv \sigma > 0.$$

Define $F: [0, T] \times R^m \times [0, 1] \to R^1$ by

$$F(t, x, y) = u(t, x, y) - V(t, x, y) - \lambda(T - t) - \mu(1 - y)$$

where $\lambda, \mu > 0$ will be chosen below. Let (t_1, x_1, y_1) be a point for which

$$F(t_1, x_1, y_1) \geqq \sup F - \varepsilon$$

for given $\sigma > \varepsilon > 0$.

We claim that $t_1 \neq T$. Indeed, if $t_1 = T$,

$$F(T, x_1, y_1) = g(x_1) - g(x_1) - \mu(1 - y_1) = -\mu(1 - y_1).$$

But $F(t_1, x_1, y_1) \geqq \sigma - \lambda T - \mu - \varepsilon$ so that if $t_1 = T$, then $0 \geqq \sigma - \lambda T - \mu y_1 - \varepsilon$ which can be made positive for λ, μ sufficiently small. Similarly $y_1 \neq 1$.

Let $0 < \gamma < (T - t_1)$, $(1 - y_1)$ and choose a smooth function $\zeta : [0, T] \times R^m \times [0, 1] \to R^1$ satisfying the properties

$$0 \leqq \zeta \leqq 1, \quad \zeta(t_1, x_1, y_1) = 1, \quad \zeta < 1 \quad \text{if } (t, x, y) \neq (t_1, x_1, y_1),$$
$$\zeta = 0 \quad \text{if } |t - t_1| + |x - x_1| + |y - y_1| > \gamma/2.$$

Finally, set

$$G(t, x, y) = F(t, x, y) + 2\varepsilon\zeta(t, x, y).$$

Then, since

$$G(t_1, x_1, y_1) \geqq \sup F + \varepsilon$$

and

$$G(t, x, y) \leqq \sup F \text{ on } |t - t_1| + |x - x_1| + |y - y_1| > \gamma/2$$

there exists (t_0, x_0, y_0), $T - t_0 > \gamma/2$, $1 - y_0 > \gamma/2$ so that

$$G(t_0, x_0, y_0) = \sup G(t, x, y).$$

Thus $u - \psi$ has a max at (t_0, x_0, y_0) for

$$\psi(t, x, y) = V(t, x, y) + \lambda(T - t) + \mu(1 - y) \in C^1.$$

Since u is a viscosity solution, this implies that

$$\max\{\psi_t + f \cdot \nabla_x \psi + h, \psi_y\} \geqq 0 \quad \text{at } (t_0, x_0, y_0).$$

But

$$\psi_y(t_0, x_0, y_0) = V_y(t_0, x_0, y_0) - \mu < 0 \qquad (\text{since } V_y \leqq 0)$$

so that

$$0 \leqq (\psi_t + f \cdot \nabla_x \psi + h)(t_0, x_0, y_0) = (V_t - \lambda + f \cdot \nabla_x V + h)(t_0, x_0, y_0).$$

Accordingly $V_t + f \cdot \nabla_x V + h \geqq \lambda > 0$ at (t_0, x_0, y_0), a contradiction. Finally, if V is only $W^{1,\infty}$ we approximate V uniformly as follows:

Let $J_\varepsilon k \equiv k^\varepsilon$ be a C^∞ mollifier of k. Then V^ε satisfies

$$V^\varepsilon(T, x, y) = g^\varepsilon(x), \qquad V^\varepsilon(t, x, 1) = r^\varepsilon(t, x)$$

and

$$V_t^\varepsilon + f \cdot \nabla_x V^\varepsilon + h^\varepsilon + (f \cdot \nabla_x V)^\varepsilon - f \cdot \nabla_x V^\varepsilon \leqq 0,$$

and

$$V_y^\varepsilon \leqq 0.$$

Since $\|k^\varepsilon - k\|_{L^\infty} \leqq C\varepsilon$, $k = g^\varepsilon$, r^ε, h^ε, V^ε and

$$\|(f \cdot \nabla_x V)^\varepsilon - f \cdot \nabla_x V^\varepsilon\|_{L^\infty} \leqq C\varepsilon$$

for some constant C independent of ε (see Lions [14, p. 32]), we can readily complete the proof.

Acknowledgment. The author is happy to acknowledge L. C. Evans for reading (and pointing out an error in) an early version of this paper.

REFERENCES

[1] G. BARLES, *Thèse de 3ème cycle*, Paris IX-Dauphine, 1982.
[2] E. N. BARRON, *Differential games with Lipschitz control functions and applications to games with partial differential equations*, Trans. Amer. Math. Soc., 219 (1976), pp. 39-76.
[3] ———, *Differential games with Lipschitz control functions and fixed initial control positions*, J. Differential Equations, 26 (1977), pp. 161-180.
[4] E. N. BARRON AND R. JENSEN, *Optimal control problems with no turning back*, J. Differential Equations, 36 (1980), pp. 223-248.
[5] ———, *A nonlinear evolution system with two subdifferentials and monotone differential games*, J. Math. Anal. Appls., 97 (1983), pp. 65-80.
[6] E. N. BARRON, L. C. EVANS AND R. JENSEN, *Viscosity solutions of Isaacs' equations and differential games with Lipschitz controls*, J. Differential Equations, to appear.
[7] I. CAPUZZO DOLCETTA AND L. C. EVANS, *Optimal switching for ordinary differential equations*, this Journal, 22 (1984), pp. 143-151.
[8] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-41.
[9] M. G. CRANDALL, L. C. EVANS AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487-502.
[10] L. C. EVANS AND P. E. SOUGANIDES, *Differential games and representation formulas for solutions of Hamilton-Jacobi equations*, to appear.
[11] A. FRIEDMAN, *Stochastic Differential Equations, Vol.* I, II, Academic Press, New York, 1976.
[12] ———, *The Cauchy problem for first order partial differential equations*, Indiana Univ. Math. J., 23 (1973), pp. 27-40.
[13] R. JENSEN AND P. L. LIONS, *Some asymptotic problems in fully nonlinear elliptic equations and stochastic control*, unpublished manuscript.
[14] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Research Notes in Mathematics, 69, Pitman, London, 1982.
[15] J. L. MENALDI AND M. ROBIN, *On some cheap control problems for diffusion processes*, Trans. Amer. Math. Soc., 278 (1983), pp. 771-802.
[16] P. E. SOUGANIDES, Ph.D. Thesis, Univ. Wisconsin, Madison, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

002

THE EXISTENCE OF VALUE AND SADDLE POINT IN GAMES OF FIXED DURATION*

LEONARD D. BERKOVITZ†

Abstract. Differential games of fixed duration are defined. The definition of strategy follows that of Friedman, while the definition of payoff follows that of Krasovskii and Subbotin. It is shown by relatively elementary methods that games of fixed duration which satisfy the Isaacs condition have values and saddle points. It is also shown under appropriate hypotheses on the data of the problem that if the Isaacs condition holds, then the value is uniformly Lipschitz continuous on bounded sets and satisfies the Isaacs equation at all points of differentiability. The relationship of the value as defined here to other values is studied.

Key words. differential games, strategy, value, saddle point, Isaacs equation

1. Introduction. A somewhat imprecise formulation of a zero-sum two-person differential game is the following. The state $x(t)$ of the game at time t is a vector in E^n determined by a system of differential equations

$$\frac{dx}{dt} = f(t, x, y, z), \qquad x(t_0) = x_0, \tag{1.1}$$

where $y = y(t)$ is a vector in E^r chosen from some set Y at each instant of time by Player I and $z = z(t)$ is a vector in E^s chosen from some set Z at each instant of time by Player II. The choice of $y(t)$ is governed by a "strategy" U selected by Player I prior to the start of play. The choice of $z(t)$ is governed by a strategy V selected by Player II prior to the start of play. Play proceeds from an initial point (t_0, x_0) until the point $(t, \phi(t))$, where ϕ is the solution of (1.1), reaches some preassigned terminal set $\mathcal{T}$. The point at which $(t, \phi(t))$ reaches $\mathcal{T}$ is called the terminal point and is denoted by $(t_f, \phi(t_f))$ or (t_f, x_f). The payoff is

$$P(t_0, x_0, U, V) = g(t_f, x_f) + \int_{t_0}^{t_f} f^0(s, \phi(s), y(s), z(s))\, ds. \tag{1.2}$$

Player I wishes to choose U so as to maximize P while Player II wishes to choose V so as to minimize P. Following Friedman [5], we call such games, games of survival.

If the terminal set $\mathcal{T}$ is the hyperplane $t = T$, the game is said to be a game of fixed duration. The payoff (1.2) then becomes

$$P(t_0, x_0, U, V) = g(x_f) + \int_{t_0}^{T} f^0(s, \phi(s), y(s), z(s))\, ds. \tag{1.3}$$

If we adjoin a new coordinate x^0 to $x = (x^1, \cdots, x^n)$ and adjoin the differential equation

$$\frac{dx^0}{dt} = f^0(t, x, y, z), \qquad x^0(t_0) = 0 \tag{1.4}$$

to (1.1), then we can write (1.2) as

$$P(t_0, x_0, U, V) = g(t_f, x_f) + x_f^0. \tag{1.2'}$$

* Received by the editors September 14, 1983, and in revised form March 16, 1984. This research was supported by the National Science Foundation under grant 7927137.

† Department of Mathematics, Purdue University, West Lafayette, Indiana 47907.

Here t_f and x_f have the same meaning as before and $x_f^0 = \phi^0(t_f)$, where (ϕ^0, ϕ) is now the solution of the system (1.1), (1.4). Note that ϕ^0 is obtained from (1.4) by an integration, once the solution ϕ of (1.1) is known. Thus there is no loss of generality in assuming that $f^0 \equiv 0$ and

$$P(t_0, x_0, U, V) = g(t_f, \phi(t_f)). \tag{1.5}$$

In games of fixed duration, which we shall study in this paper, (1.5) becomes

$$P(t_0, x_0, U, V) = g(\phi(T)). \tag{1.6}$$

If

$$\sup_U \inf_V P(t_0, x_0, U, V) = \inf_V \sup_U P(t_0, x_0, U, V),$$

then we denote this number by $W(t_0, x_0)$ and say that the game has value equal to $W(t_0, x_0)$. A pair of strategies (U^*, V^*) is said to be a saddle point if for all U and V over which the sup and inf are taken,

$$P(t_0, x_0, U, V^*) \leqq P(t_0, x_0, U^*, V^*) \leqq P(t_0, x_0, U^*, V).$$

The strategies U^* and V^* are called optimal strategies.

The study of differential games was initiated by Rufus Isaacs in a series of RAND Corporation memoranda [7], which were later expanded into a book [8]. Differential games originally arose in the study of pursuit and evasion problems and in the study of military tactical problems. They are models of zero sum conflict in which the state of the system is governed by differential equations controlled by two antagonists. Differential games can also be used to model control problems in which the system is subject to unknown disturbances, and the disturbances are viewed as being chosen by a malevolent nature, who is assigned the role of Player II. Player I wishes to choose controls that guarantee the best possible performance in the event of the worst possible disturbance.

The principal problems associated with differential games are the following:

1) Define the notion of strategy precisely.

2) For the given definition of strategy does the game have a value?

3) If the game has a value, does it have a saddle point?

4) If the game has a saddle point, find the optimal strategies U^* and V^* that constitute the saddle point.

The work of Isaacs was concerned with the last question. He assumed that the game had a twice continuously differentiable value and a saddle point. Under these assumptions he derived what is now known as the Isaacs equation and developed solution methods (see [8]). Many interesting examples, however, fail to have twice continuously differentiable values. Later, this author [1] treated a class of games in which it was assumed that saddle points of a certain type existed and obtained necessary conditions that must be satisfied by the saddle points. The assumptions made were much less restrictive than Isaacs'.

Investigation of questions one and two was begun by Fleming in [3] and [4]. A more comprehensive investigation of these questions was later carried out by Friedman [5] and was carried still further by Elliott and Kalton and by Friedman again. This work is summarized in [6], where complete references to the work are also given. In this work the direct definition of a game along the lines indicated in the opening paragraphs of this paper is abandoned. Instead, indirect definitions are used which involve choices of y and z at discrete times and passage to the limit as the lengths of the discrete time intervals tend to zero. It is shown that if a certain minimax condition,

known as the Isaacs condition, is satisfied, and other natural conditions hold, then the game has a value for all initial points (t_0, x_0). Moreover, the value is Lipschitz continuous and satisfies a Hamilton-Jacobi equation almost everywhere (see [6]). These results are established by adding a small noise term to the right-hand side of (1.1), obtaining appropriate results for the resulting game, and then passing to the limit as the noise tends to zero. The arguments are nonelementary and involve probability theory and the theory of partial differential equations.

Krasovskii and Subbotin [9] have defined the notion of strategy in a way that is different from that of Friedman. Their definition also involves limiting behavior of games with choices made at discrete times. Under their definition, games of fixed duration that satisfy the Isaacs condition have values and saddle points. The proof in [9] is intuitively appealing and is relatively elementary. It does not involve probabilistic arguments or the use of partial differential equations.

In this paper we shall combine what we consider to be the best features of Friedman's approach with the best features of the Krasovskii-Subbotin approach. We shall adopt the Friedman definition of strategy and the Krasovskii-Subbotin definitions of payoff and saddle point. We shall show, by relatively elementary methods, that games of fixed duration which satisfy the Isaacs condition have values and saddle points. The basic idea of the argument is due to Krasovskii and Subbotin. The details, however, are different in many important respects, and, we feel, simpler. For the benefit of readers familiar with the Krasovskii-Subbotin approach and with the Friedman approach, we have tried to use notations similar to those used by these authors for related or identical concepts.

We show, again by relatively elementary methods, that under appropriate conditions on the data of the problem, if the Isaacs condition holds, then the value is uniformly Lipschitz continuous on bounded sets and satisfies the Isaacs equation at all points of differentiability. We also show that if the Isaacs condition holds, then the value in our sense is equal to the values of Friedman, Krasovskii-Subbotin, Fleming and Elliott-Kalton.

Finally, we point out that except for § 13, which compares our upper and lower values with others, no knowledge of the differential game literature is required to read this paper.

2. Assumptions and notation. Let t denote time, let $x=(x^1, \cdots, x^n)$ denote a vector in n-dimensional real Euclidean space R^n, let $y=(y^1, \cdots, y^r)$ denote a vector in R^r and let $z=(z^1, \cdots, z^s)$ denote a vector in R^s. We shall use superscripts to denote components of vectors and we shall use subscripts to distinguish vectors. We shall denote the inner product of two vectors x and y by $\langle x, y\rangle$ and the Euclidean norm of a vector x by $|x|$. Thus, $|x|^2=\langle x, x\rangle$.

Let Y be a compact subset of R^r, let Z be a compact subset of R^s and let T_0 and T_1 be two real numbers satisfying $T_0<T_1$. Let $f^1, \cdots, f^n$ be real valued functions of (t, x, y, z) defined on the set

$$\mathcal{D} \equiv [T_0, T_1]\times R^n \times Y \times Z, \tag{2.1}$$

and let $f=(f^1, \cdots, f^n)$. Let $T_0<T<T_1$ and let $\mathcal{T}$ be the closed set in (t, x) space, $[T, \infty]\times R^n$. Let g be a real valued function of (t, x) defined on $[T_0, T_1]\times R^n$ and bounded on bounded subsets of its domain.

The function f will govern the dynamics, as suggested by (1.1), and the function g will enter the payoff as in (1.6). The precise interpretation of (1.1) and (1.3) will be given in § 4.

Concerning f and g we assume the following.

Assumption I. (i) The function f is continuous on $\mathcal{D} = [T_0, T_1] \times R^n \times Y \times Z$. (ii) There exists a function k that is integrable on $[T_0, T_1]$ such that

$$\langle x, f(t, x, y, z)\rangle \leqq k(t)(1+|x|^2) \tag{2.2}$$

for all (t, x, y, z) in $\mathcal{D}$. (iii) There exists a constant $K>0$ such that for all t in $[T_0, T_1]$, y in Y, z in Z and $x, \bar{x}$ in R^n

$$|f(t, x, y, z)-f(t, \bar{x}, y, z)| \leqq K|x-\bar{x}|. \tag{2.3}$$

(iv) The function g is continuous on R^n.

In some situations we shall need the following slightly stronger version of Assumption I.

Assumption I′. Statements (i)′ and (ii)′ are as (i) and (ii) in Assumption I. (iii)′ There exists a $K>0$ such that for all $x, \bar{x}$ in R^n and $t, \bar{t}$ in $[T_0, T_1]$,

$$|f(t, x, y, z)-f(\bar{t}, \bar{x}, y, z)| \leqq K(|t-\bar{t}|+|x-\bar{x}|).$$

(iv)′ There exists a constant $K'>0$ such that for $x, \bar{x}$ in R^n

$$|g(x)-g(\bar{x})| \leqq K'|x-\bar{x}|. \tag{2.4}$$

Using standard arguments in the theory of ordinary differential equations, we obtain the following consequences of Assumption I, which will be needed in our formulation of the differential game and in our analysis. Let $T_1 < t_0 < T$. Let u and v be two measurable functions defined on $[t_0, T]$ and satisfying $u(t) \in Y$ a.e. and $v(t) \in Z$ a.e. Such functions will be called *control functions* or *controls*. Then for any (t_0, x_0) with $x_0 \in R^n$, there exists a unique solution ϕ of the differential equation

$$x' = f(t, x, u(t), v(t)), \qquad x(t_0) = x_0, \tag{2.5}$$

and this solution is defined for all $t_0 \leqq t \leqq T$. Let X be a compact set in R^n. Then there exists a constant K_0 such that any solution ϕ of (2.5) with $x_0 \in X$ satisfies $|\phi(t)| \leqq K_0$ for $t_0 \leqq t \leqq T$, independent of the choice of controls u and v. Since f is continuous, this implies that there exists a constant K such that all such solutions satisfy

$$\begin{aligned} &|\phi'(t)| \leqq K \quad \text{a.e. on } [t_0, T], \\ &|\phi(t)-x_0| \leqq K|t-t_0| \quad \text{on } [t_0, T]. \end{aligned} \tag{2.6}$$

From (2.6) it is clear that the set of solutions of (2.5) obtained as x_0 ranges over X and u and v range over all possible controls is uniformly bounded and equi-absolutely continuous.

Now let u and v be any pair of controls, let $T_0 \leqq t_0 < t_1 \leqq T_1$ and let x_0 and x_1 belong to X. Let ϕ_0 denote the solution of (2.5) satisfying the initial condition $\phi_0(t_0) = x_0$ and let ϕ denote the solution of (2.5) satisfying the initial condition $\phi(t_1) = x_1$. It follows from (2.6), the Lipschitz condition (2.3) and Gronwall's lemma that there exists a constant K_1, independent of t_0, t_1, x_0, x_1, u, v, such that for $t > t_1$

$$|\phi_1(t)-\phi_0(t)| \leqq K_1(|t_1-t_0|+|x_1-x_0|).$$

Note that the preceding conclusions remain valid if we replace (iii) by a Lipschitz condition with Lipschitz constant $K_R>0$, valid for $|x|<R$, $|\bar{x}|<R$, for each $R>0$. In particular, the solutions of (2.5) will be uniformly bounded for all x_0 in X. Hence there is no loss of generality in assuming a global Lipschitz condition.

3. Definition of the game. Let $T_0 \leqq t_0 < T$. Let π_n be a partition of $[t_0, T]$ by points $t_0 < t_1 < \cdots < t_n = T$. Define the norm of π_n, denoted by $\|\pi_n\|$, as follows: $\|\pi_n\| = \max\{i: t_i - t_{i-1}, i = 1, \cdots, n\}$. Let $\{\pi_n\}_{n=1}^{\infty}$ be a sequence of partitions such that for each $n = 1, 2, 3, \cdots$, $\|\pi_n\| \leqq K/n$, where K is a constant independent of n. Let

$$I_j = \{t: t_{j-1} \leqq t < t_j\} = [t_{j-1}, t_j), \qquad j = 1, \cdots, n.$$

Let $\mathcal{Y}_j$ denote the set of controls u defined on I_j and let $\mathcal{Z}_j$ denote the set of controls v defined on I_j. Note that I_j and hence $\mathcal{Y}_j$ and $\mathcal{Z}_j$ depend on π_n. We shall not indicate this dependence in the notation.

We can now define the notion of strategy for each player. Let $\Gamma_{n,1}$ denote a choice of function in $\mathcal{Y}_1$. For $2 \leqq j \leqq n$ let $\Gamma_{n,j}$ denote a map from $\mathcal{Y}_1 \times \mathcal{Z}_1 \times \cdots \times \mathcal{Y}_{j-1} \times \mathcal{Z}_{j-1}$ to $\mathcal{Y}_j$. Let

$$\Gamma_n = (\Gamma_{n,1}, \cdots, \Gamma_{n,n}), \qquad n = 1, 2, \cdots.$$

We call Γ_n an nth stage strategy for Player I. Similarly, let $\Delta_{n,1}$ denote a choice of function in $\mathcal{Z}_1$. For $2 \leqq j \leqq n$ let $\Delta_{n,j}$ denote a map from $\mathcal{Y}_1 \times \mathcal{Z}_1 \times \cdots \times \mathcal{Y}_{j-1} \times \mathcal{Z}_{j-1}$ to $\mathcal{Z}_j$. Let

$$\Delta_n = (\Delta_{n,1}, \cdots, \Delta_{n,n}), \qquad n = 1, 2, \cdots.$$

We call Δ_n an nth *stage strategy* for Player II. The nth stage strategies Γ_n and Δ_n correspond to the lower δ strategies of Friedman. We shall not use upper strategies in our work.

By a strategy Γ for Player I we mean a sequence $\{\Gamma_n\}$ of nth stage strategies. Similarly, by a strategy Δ for Player II we mean a sequence $\{\Delta_n\}$ of nth stage strategies.

Let $\{u_n\}$ be a sequence of controls. Let $\Gamma^c = \{\Gamma_n^c\}$ be the strategy such that for $n = 1, 2, 3, \cdots$

$$(\Gamma_{n,1}^c)(t) = u_n(t) \qquad \text{for } t \in I_1$$

and

$$\Gamma_{n,j}^c: \mathcal{Y}_1 \times \mathcal{Z}_1 \times \cdots \times \mathcal{Y}_{j-1} \times \mathcal{Z}_{j-1} \to u_n(t) \quad \text{for } t \in I_j,$$

for $j = 2, \cdots, n$, $n = 1, 2, 3, \cdots$. Thus Γ_n^c always selects u_n. The strategy Γ^c will be called the *constant component strategy* corresponding to u_n. If $u_n = u$ for all n, then we denote the constant component strategy Γ^c corresponding to $\{u_n\}$ by $\bar{\Gamma} = \{\bar{\Gamma}_n\}$ and call it the *constant strategy corresponding* to u. The strategy $\bar{\Gamma} = \{\bar{\Gamma}_n\}$ is such that $\bar{\Gamma}_{n,j}$ always selects the fixed control u. Constant component strategies Δ^c and constant strategies $\bar{\Delta}$ for Player II are defined similarly.

We next define the concept of a motion in the game. A pair of nth stage strategies (Γ_n, Δ_n) determine control functions (u_n, v_n) on $[t_0, T]$, where

$$u_n(t) = (\Gamma_{n,1})(t), \quad v_n(t) = (\Delta_{n,1})(t), \quad t \in I_1,$$

$$u_n(t) = (\Gamma_{n,j}(u_1, v_1, u_2, v_2, \cdots, u_{j-1}, v_{j-1}))(t), \qquad t \in I_j,$$

$$v_n(t) = (\Delta_{n,j}(u_1, v_1, u_2, v_2, \cdots, u_{j-1}, v_{j-1}))(t), \qquad 2 \leqq j \leqq n.$$

The controls (u_n, v_n) determined this way are called the nth stage outcomes of (Γ_n, Δ_n).

In equation (1.1), if we replace y by $u_n(t)$, z by $v_n(t)$, and x_0 by x_{0n} we obtain the system of differential equations

$$\frac{dx}{dt} = f(t, x, u_n(t), v_n(t)), \qquad x(t_0) = x_{0n}. \tag{3.1}$$

As pointed out in the discussion following Assumption I, equation (3.1) has a unique solution $\phi_n(\ , t_0, x_{0n}, u_n, v_n)$ defined on $[t_0, T]$. We call the solution $\phi_n(\ , t_0, x_{0,n}, u_n, v_n)$ an nth stage trajectory.

Any uniform limit of a subsequence of the nth stage trajectories $\phi_n(\ , t_0, x_{0,n}, u_n, v_n)$, $n = 1, 2, 3, \cdots$, where $x_{0,n} \to x_0$ and (u_n, v_n) is the outcome of (Γ_n, Δ_n), will be called a *motion* or *motion of the game* corresponding to strategies $\Gamma = \{\Gamma_n\}$ and $\Delta = \{\Delta_n\}$. Following Krasovskii and Subbotin [9] we denote a motion corresponding to (Γ, Δ) by $\phi[\ , t_0, x_0, \Gamma, \Delta]$.

It follows from Assumption I that corresponding to a pair of strategies (Γ, Δ) and a sequence of initial conditions $\{x_{0n}\}$ with $x_{0n} \to x_0$, the corresponding sequence of nth stage trajectories $\{\phi_n(\ , t_0, x_{0,n}, u_n, v_n)\}$ is uniformly bounded and equicontinuous. Hence there do exist motions $\phi[\ , t_0, x_0, \Gamma, \Delta]$. We shall denote the set of all motions corresponding to (Γ, Δ) by $\Phi[\ , t_0, x_0, \Gamma, \Delta]$.

By $\phi[t, t_0, x_0, \Gamma, \Delta]$ we shall mean, as usual, the motion evaluated at the point t. By $\Phi[t, t_0, x_0, \Gamma, \Delta]$ we shall mean the set of all values $\phi[t, t_0, x_0, \Gamma, \Delta]$, where $\phi[\ , t_0, x_0, \Gamma, \Delta]$ ranges over $\Phi[\ , t_0, x_0, \Gamma, \Delta]$. We point out that if $\Gamma \equiv \{\Gamma_n\}$ and $\Gamma' \equiv \{\Gamma'_n\}$ differ only in a finite number of elements Γ_n and Γ'_n, and if the same is true for $\Delta \equiv \{\Delta_n\}$ and $\Delta' \equiv \{\Delta'_n\}$, then

$$\Phi[\ , t_0, x_0, \Gamma, \Delta] = \Phi[\ , t_0, x_0, \Gamma', \Delta'].$$

We shall sometimes write $\phi[\ ]$ for $\phi[\ , t_0, x_0, \Gamma, \Delta]$ and $\Phi[\ ]$ for $\Phi[\ , t_0, x_0, \Gamma, \Delta]$.

In conclusion, we point out that, in general, a motion $\phi[\ ]$ will not be obtained as a solution of

$$\frac{dx}{dt} = f(t, x, u(t), v(t)), \qquad x(t_0) = x_0$$

for an appropriate choice of controls u, v.

Having defined the strategy spaces for the two players, it only remains to define the payoff in order to define the game. In the game of fixed duration with initial point (t_0, x_0) the payoff $P(t_0, x_0, \Gamma, \Delta)$ corresponding to a pair of strategies (Γ, Δ) is set valued and is defined as follows:

$$P(t_0, x_0, \Gamma, \Delta) = g(\Phi[T, t_0, x_0, \Gamma, \Delta]). \tag{3.2}$$

Player I's objective is to choose Γ so as to maximize $P(t_0, x_0, \Gamma, \Delta)$, while Player II's objective is to minimize $P(t_0, x_0, \Gamma, \Delta)$. That is, Player I wishes to choose Γ so as to make all the elements of $P(t_0, x_0, \Gamma, \Delta)$ large, while Player II wishes to choose Δ so as to make all the elements of $P(t_0, x_0, \Gamma, \Delta)$ small.

4. Value and saddle point. Let

$$\begin{aligned} W^-(t_0, x_0) &= \sup_{\Gamma} \inf_{\Delta} P(t_0, x_0, \Gamma, \Delta), \\ W^+(t_0, x_0) &= \inf_{\Delta} \sup_{\Gamma} P(t_0, x_0, \Gamma, \Delta). \end{aligned} \tag{4.1}$$

Then

$$W^-(t_0, x_0) \leqq W^+(t_0, x_0). \tag{4.2}$$

If $W^-(t_0, x_0) = W^+(t_0, x_0)$, we denote this common value by $W(t_0, x_0)$ and say that the game has a value equal to $W(t_0, x_0)$.

Let A and B be two sets of real numbers. We say that $A \geqq B$, or $B \leqq A$, if for every a in A and every b in B the inequality $a \geqq b$ holds. Also, if μ is a real number and A is a set, by $\mu \geqq A$ we mean that $\mu \geqq a$ for all $a \in A$. A similar meaning holds for $\mu \leqq A$.

A pair of strategies (Γ^*, Δ^*) will be called *optimal strategies*, or will be said to constitute a *saddle point* of the game, if

$$P(t_0, x_0, \Gamma, \Delta^*) \leqq P(t_0, x_0, \Gamma^*, \Delta^*) \leqq P(t_0, x_0, \Gamma^*, \Delta) \tag{4.3}$$

for all strategies Γ for Player I and all strategies Δ for Player II.

Since

$$\inf_{\Delta} P(t_0, x_0, \Gamma^*, \Delta) \leqq \sup_{\Gamma} \inf_{\Delta} P(t_0, x_0, \Gamma, \Delta) = W^-(t_0, x_0),$$

$$\sup_{\Gamma} P(t_0, x_0, \Gamma, \Delta^*) \geqq \inf_{\Delta} \sup_{\Gamma} P(t_0, x_0, \Gamma, \Delta) = W^+(t_0, x_0),$$

it follows from (4.3) and (4.2) that if a saddle point exists, then the game has a value and

$$W(t_0, x_0) = P(t_0, x_0, \Gamma^*, \Delta^*). \tag{4.4}$$

We point out that our definition of saddle point is different from Friedman's definition of saddle point and generalized saddle point. We do not require that the payoff be evaluated along a trajectory of the system.

5. Relaxed controls and trajectories. In what follows we shall utilize relaxed controls. We therefore summarize some of the essential properties of relaxed controls for the convenience of the reader. For details and further discussion see Warga [11, Chap. IV].

Let Ω be a compact set in R^m and let g be a mapping from $[T_0, T_1] \times R^n \times \Omega$ to R^n having the following properties: (i) g is continuous on $R^n \times \Omega$ for each $t \in [T_0, T_1]$ and is measurable on $[T_0, T_1]$ for each (x, w) in $R^n \times \Omega$; (ii) for each compact subset C of $R^n \times \Omega$ there is an integrable function L defined on $[T_0, T_1]$ such that for all (t, x, w) in $[T_0, T] \times C$, $|g(t, x, w)| \leqq L(t)$. We shall consider trajectories for control systems of the form

$$x' = g(t, x, \omega(t)), \qquad x(t_0) = x_0, \tag{5.1}$$

where ω is a measurable function defined on $[T_0, T_1]$ with values $\omega(t) \in \Omega$ a.e.

A relaxed control on a fixed interval $I \subseteq [T_0, T_1]$ is a mapping $\mu: t \to \mu(t) = \mu(t, \cdot)$ from $[T_0, T_1]$ to the probability measures on Ω such that for every polynomial p the function P defined by

$$P(t) = \int_{\Omega} p(z)\, d\mu\ (t; z)$$

is Lebesgue measurable on I. An arbitrary control ω can be identified with the relaxed control μ_ω that assigns to t the atomic measure concentrated at $\omega(t)$. A *relaxed trajectory* ψ corresponding to the relaxed control μ is an absolutely continuous function ψ defined on I and satisfying $\psi(t_0) = x_0$, $t_0 \in I$, and

$$\psi'(t) = \int_{\Omega} g(t, \psi(t), w)\, d\mu\ (t, w) \quad \text{a.e.} \tag{5.2}$$

To simplify notation we shall define

$$g(t, \psi(t), \mu(t)) \equiv \int_{\Omega} g(t, \psi(t), w)\, d\mu(t, w)$$

and shall write (5.2) as

$$\psi'(t) = g(t, \psi(t), \mu(t)).$$

It can be shown that an equivalent definition of relaxed trajectory ψ is an absolutely continuous function defined on I satisfying $\psi(t_0) = x_0$ and $\psi'(t) \in \operatorname{co}\{g(t, \psi(t), \Omega)\}$, where co denotes "convex hull" and

$$g(t, x, \Omega) = \{y: y = g(t, x, w),\ w \in \Omega\}.$$

It is useful to view the set of relaxed controls as a subset of a space of linear functionals. Let $\mathscr{L}$ denote the set of functions θ defined on $I \times \Omega$ such that $\theta(\ \ , w)$ is measurable on I for each w in Ω, $\theta(t, \ \ )$ is continuous on Ω for each t in I and the function $\max\{|\theta(t, w)|: w \in \Omega\}$ is integrable over I. The set $\mathscr{L}$ is a Banach space with norm

$$\|\theta\| = \int_I \left(\max_w |\theta(t, w)| \right) dt.$$

Let $\tilde{\mu}$ be a mapping $t \to \tilde{\mu}(t) = \tilde{\mu}(t, \ \ )$ from I to the set of finite Radon measures on Ω. Let $\mathscr{N}$ denote the set of such mappings with the following two properties. (i) For every polynomial p the function $t \to \int_\Omega p(w)\, d\tilde{\mu}\ (t, w)$ is measurable. (ii) If $|\tilde{\mu}(t)|$ denotes the total variation measure of $\tilde{\mu}(t)$, then *ess* $\sup\{|\tilde{\mu}(t)|(\Omega): t \in I\}$ is finite. Then $\mathscr{N}$ can be identified with the dual space $\mathscr{L}^*$ of $\mathscr{L}$, where each $\tilde{\mu}$ in $\mathscr{N}$ is identified with the functional which we also denote by $\tilde{\mu}$, as follows:

$$\tilde{\mu}(\theta) = \int_I \left(\int_\Omega \theta(t, w)\, d\tilde{\mu}\ (t, w) \right) dt, \qquad \theta \in \mathscr{L}.$$

The norm of an element of $\mathscr{N}$ viewed as an element of $\mathscr{L}^*$ is denoted by $\|\tilde{\mu}\|_L$ and is defined by $\|\tilde{\mu}\|_L = \sup\{|\tilde{\mu}(\theta)|: |\theta| = 1\}$. It is also given by $\|\tilde{\mu}\|_L =$ *ess* $\sup\{|\tilde{\mu}(t)|(\Omega): t \in I\}$.

The set of relaxed controls is clearly a subset of $\mathscr{L}^*$. An important property for our purposes is that the relaxed controls constitute a convex subset of $\mathscr{L}^*$ that is compact and sequentially compact in the weak-star topology of $\mathscr{L}^*$.

Another important fact is that the ordinary trajectories of the system (7.1) are dense in the set of relaxed trajectories in the uniform topology on I. Thus, for any relaxed trajectory ψ on I there is a sequence of controls $\{\omega_n\}$ and a sequence of corresponding trajectories $\{\phi_n\}$ such that $\phi_n(t)$ converges to $\psi(t)$ uniformly on I.

The following lemma will be used frequently.

LEMMA 5.1. *Let $\{\mu_n\}$ be a sequence of relaxed controls on an interval $I = [t_1, t_2] \subseteq [T_0, T_1]$ and let $\{\psi_n\}$ be a sequence of corresponding relaxed trajectories. Let ψ_n converge uniformly to a function ψ on I and let μ_n converge weak star to a relaxed control μ. Then ψ is a relaxed trajectory corresponding to μ; i.e.*

$$\psi'(t) = g(t, \psi(t), \mu(t)), \qquad \psi(t_1) = x_1,$$

where $x_1 = \lim_{n \to \infty} \psi_n(t_1)$.

Lemma 5.1 is a corollary of the following known result.

LEMMA 5.2. *Let $\{\psi_n\}$ be a sequence of continuous functions converging uniformly on I to a function ψ. Let $\{\mu_n\}$ be a sequence of relaxed controls converging weak-star to a relaxed control μ. Then for any function X in $L_\infty[I]$,*

$$\lim_{n\to\infty} \int_I X(t)g(t, \psi_n(t), \mu_n(t))\, dt = \int_I X(t)g(t, \psi(t), \mu(t))\, dt. \tag{5.3}$$

6. Properties of motions. In this and subsequent sections we shall consider games, trajectories and motions with varying initial points. We shall denote these initial points by (τ, ξ), where $T_0 \leqq \tau < T$ and $\xi \in R^n$. Strategies Γ and Δ will be strategies over the interval $[\tau, T]$.

Our first property is that motions resulting from strategy pairs (Γ, Δ) with at least one of Γ or Δ a constant strategy can be obtained as appropriate relaxed trajectories.

LEMMA 6.1. *Let (τ, ξ) be a point in $[T_0, T) \times R^n$. Let $\bar{\Gamma}$ be the constant strategy with value $u(t)$ on $[\tau, T]$, and let Δ be a strategy on $[\tau, T]$. Then for any motion $\phi[\ , \tau, \xi, \bar{\Gamma}, \Delta]$ there exist a relaxed control ζ and corresponding relaxed trajectory $\psi(\cdot) = \psi(\cdot, \tau, \xi, u, \zeta)$ satisfying*

$$\psi'(t) = \int_z f(t, \psi(t), u(t), z)\, d\zeta(t, z), \qquad x(\tau) = \xi \tag{6.1}$$

for a.e. t in $[\tau, T]$ such that

$$\phi[t, \tau, \xi, \bar{\Gamma}, \Delta] = \psi(t, \tau, \xi, u, \zeta). \tag{6.2}$$

Conversely, given any solution ψ of (6.1), *there exist a strategy Δ for Player* II *and a motion $\phi[\ , \tau, \xi, \bar{\Gamma}, \Delta]$ such that* (6.2) *holds.*

Proof. The motion $\phi[\ , \tau, \xi, \bar{\Gamma}, \Delta]$ is the uniform limit of n_kth stage trajectories $\phi_{n_k}(\ ) \equiv \phi_{n_k}(\ , \tau, \xi_{n_k}, u(t), v_{n_k}(t))$ where (u, v_{n_k}) is the outcome of $(\bar{\Gamma}_{n_k}, \Delta_{n_k})$ and $\xi_{n_k} \to \xi$. We relabel the sequence ϕ_{n_k} as ϕ_n, ξ_{n_k} as ξ_n, and the controls as v_n. Thus for a.e. t in $[\tau, T]$,

$$\phi_n'(t) = f(t, \phi_n(t), u(t), \zeta_n(t)), \qquad \phi_n(\tau) = \xi_n,$$

where $\zeta_n(t)$ is the delta measure over Z that is concentrated at $v_n(t)$. There exists a subsequence $\{\zeta_n\}$ and a probability measure ζ on Z such that $\zeta_n \to \zeta$ weak star. Hence by Lemma 5.1, the corresponding subsequence $\{\phi_n\}$ converges uniformly to a solution $\psi(\ ) = \psi(\ , \tau, \xi, u, \zeta)$ of (6.1). But the original sequence $\{\phi_n\}$ converged uniformly to $\phi[\ , \tau, \xi, \bar{\Gamma}, \Delta]$, and so (6.2) is established.

To establish the converse we note that since ordinary trajectories are dense in the relaxed trajectories in $C^n[\tau, T]$, the following holds. There exists a sequence of controls $\{v_n\}$ and corresponding trajectories ϕ_n such that

$$\phi_n'(t) = f(t, \phi_n(t), u(t), v_n(t)), \qquad \phi_n(\tau) = \xi_n$$

and $\phi_n(t) \to \psi(t)$ uniformly on $[\tau, T]$. Let $\Delta^c = \{\Delta_n^c\}$ be the constant component strategy for Player II corresponding to $\{v_n\}$. Then $\psi(t, \tau, \xi, u, \zeta) = \phi[t, \tau, \xi, \bar{\Gamma}, \Delta^c]$, and the lemma is proved.

For fixed Γ, let

$$\Phi[\ , \tau, \xi, \Gamma] = \bigcup_{\Delta} \Phi[\ , \tau, \xi, \Gamma, \Delta]. \tag{6.3}$$

For fixed Δ let $\Phi[\ , \tau, \xi, \Delta]$ be defined similarly.

LEMMA 6.2. *Let X be a compact set in R^n and let Γ be a strategy for Player* I. *Then the union of all sets $\Phi[\cdot, \tau, \xi, \Gamma]$ as ξ ranges over X is compact in $C^n[\tau, T]$. A similar statement holds for $\Phi[\cdot, \tau, \xi, \Delta]$.*

Proof. We carry out the proof for $\Phi[\ \ , \tau, \xi, \Gamma]$. From the discussion in § 2 following Assumption I, it follows that the set of all nth stage trajectories, $n = 1, 2, 3, \cdots$ with initial point ξ in X resulting from the fixed strategy Γ and any strategy Δ is uniformly bounded and equiabsolutely continuous. Thus, the set of all motions $\Phi[\ \ , \tau, \xi, \Gamma]$, $\xi \in X$, is uniformly bounded and equiabsolutely continuous. It now follows from Ascoli's theorem that to prove that $\Phi[\ \ , \tau, \xi, \Gamma]$, $\xi \in X$, is compact in $C^n[\tau, T]$ we need only show that it is closed in $C^n[\tau, T]$.

Let $\{\phi_n[\ \ , \tau, \xi_n, \Gamma, \Delta(n)]\}$ be a sequence of motions converging uniformly on $[\tau, T]$ to a function ϕ. Let $\phi(\tau) = \xi$. Then $\xi_n \to \xi$ and $\xi \in X$. Also, for each positive integer n there exists an integer $k = k(n)$ such that $k(n+1) > k(n)$ and a $k(n)$th stage trajectory $\phi_{n,k}(\ \ , \tau, \xi_{n,k} u_{n,k}, v_{n,k})$, where $(u_{n,k}, v_{n,k})$ is the $k(n)$th stage outcome corresponding to $(\Gamma_{k(n)}, \Delta_{k(n)}(n))$, such that

$$
\begin{aligned}
&|\xi_{n,k} - \xi_n| < 1/2^n, \\
&|\phi_n[t, \tau, \xi_n, \Gamma, \Delta(n)] - \phi_{n,k}(t, \pi, \xi_{n,k}, u_{n,k}, v_{n,k})| < 1/2^n
\end{aligned} \tag{6.4}
$$

for all t in $[\tau, T]$.

Let v be an arbitrary control on $[\tau, T]$. Let $\hat{\Delta} = \{\hat{\Delta}_m\}$, where if $m \neq k(n)$, then $\hat{\Delta}_{m,j}$ assigns the control v to I_j, $j = 1, 2, \cdots, m$. If $m = k(n)$, $\hat{\Delta}_m = \hat{\Delta}_{k(n)}(n)$. Let $\{\xi_m\}$ be a sequence defined by $\xi_m = \xi$ if $m \neq k(n)$ and $\xi_m = \xi_{n,k}$ if $m = k(n)$. Corresponding to $(\Gamma, \hat{\Delta})$ and $\{\xi_m\}$ we obtain a sequence of outcomes $(\hat{u}_m, \hat{v}_m)$ and mth stage trajectories $\phi_m(\ \ , \tau, \xi_m, \hat{u}_m, \hat{v}_m)$. By construction, the subsequence $\{\phi_{k(n)}(\ \ , \tau, \xi_{k(n)}, \hat{u}_{k(n)}, \hat{v}_{k(n)})\}$ is such that $\phi_{k(n)} = \phi_{n,k}$, where $\phi_{n,k}$ is as in (6.4), $n = 1, 2, 3, \cdots$. It now follows from (6.4) and the uniform convergence of $\phi_n[\ \ , \tau, \xi_n, \Gamma, \Delta(n)]$ to ϕ that $\phi_{k(n)}$ converges uniformly to ϕ. Thus ϕ is a motion $\phi[\ \ , \tau, \xi, \Gamma, \hat{\Delta}]$ and $\Phi[\ \ , \tau, \xi, \Gamma]$ is closed.

COROLLARY. *For each t in $[\tau, T]$, the sets $\Phi[t, \tau, \xi, \Gamma]$ and $\Phi[t, \tau, \xi, \Delta]$, ξ in X, are compact in R^n.*

Our next result states that any terminal segment of a motion is again a motion.

LEMMA 6.3. *Let $\phi[\ \ , \tau, \xi, \Gamma, \Delta]$ be an arbitrary motion, let $\tau < t_1 < T$ and let $x_1 = \phi[t_1, \tau, \xi, \Gamma, \Delta]$. Then there exist strategies Γ' and Δ' for the game with initial point (t_1, x_1) and a motion $\phi[\ \ , t_1, x_1, \Gamma', \Delta']$ such that for all $t_1 \leqq t \leqq T$*

$$\phi[t, t_1, x_1, \Gamma', \Delta'] = \phi[t, \tau, \xi, \Gamma, \Delta].$$

We leave the proof to the reader.

In studying the properties of the functions W^+ and W^- it will be necessary for us to compare the set of all possible motions having initial point (τ, ξ) with the set of all possible motions having initial point (τ', ξ'). Therefore, we consider two intervals $[\tau, T]$ and $[\tau', T]$. Let

$$s = s(t) = \tau' + \frac{T - \tau'}{T - \tau}(t - \tau). \tag{6.5}$$

We next define a transplant mapping θ from the control functions on $[\tau, T]$ to the control functions on $[\tau', T]$. If w is a control function on $[\tau, T]$, we assign to w the control function $\bar{w} = \theta(\mathbf{w})$ on $[\tau', T]$ by the formula

$$\bar{w}(s) = (\theta(w))(s) = w(t), \tag{6.6}$$

where $s = s(t)$ is given by (6.5). This correspondence is a one-to-one map of the space of all control functions on $[\tau, T]$ to the space of all control functions on $[\tau', T]$. Thus the inverse mapping θ^{-1} exists.

The mapping θ induces a one-to-one mapping of the strategies Γ, Δ on $[\tau, T]$ onto the strategies $\tilde{\Gamma}, \tilde{\Delta}$ on $[\tau', T]$. We shall denote this mapping also by θ. Let π_n be the nth partition of $[\tau, T]$ with partition points $\tau = \tau_0 < \tau_1 < \cdots < \tau_n = T$, and with corresponding intervals $I_1, \cdots, I_n$. Let

$$\tau_i' = \tau' + \frac{T-\tau'}{T-\tau}(\tau_i - \tau), \qquad i = 1, \cdots, n.$$

Then the points $\tau' = \tau_0' < \tau_1' < \cdots < \tau_n' = T$ induce a partition π_n' of $[\tau', T]$ of norm not exceeding K'/n, where K' is a constant independent of n. Let $I_1', \cdots, I_n'$ denote the intervals comprising the partition π_n' on $[\tau', T]$. By means of (6.5) and (6.6) with τ' replaced by τ_{j-1}' and T replaced by τ_j, the controls u on I_j are put into one-to-one correspondence with the controls $\bar{u}$ on I_j'. A similar statement holds for controls v and $\bar{v}$.

We now define $\tilde{\Gamma} = \theta\Gamma$ by defining

$$\tilde{\Gamma}_n = (\theta\Gamma)_n = ((\theta\Gamma)_{n,1} \cdots (\theta\Gamma)_{n,n}) = (\tilde{\Gamma}_{n,1}, \cdots, \tilde{\Gamma}_{n,n}).$$

For $j = 1$,

$$\text{if} \quad \Gamma_{n,1}(t) = u_1(t), \quad \text{then} \quad \tilde{\Gamma}_{n,1}(s) = \bar{u}_1(s) \equiv (\theta u_1)(s).$$

For $2 \leqq j \leqq n$, if

$$(\Gamma_{n,j}(u_1, v_1, \cdots, u_{j-1}, v_{j-1}))(t) = u_j(t),$$

then

$$(\tilde{\Gamma}_{n,j}(\bar{u}_1, \bar{v}_1, \cdots, \bar{u}_{j-1}, \bar{v}_{j-1}))(t) = \bar{u}_j(s) = (\theta u_j)(s),$$

where $\bar{u}_i = \theta u_i$, $\bar{v}_i = \theta v_i$, $i = 1, \cdots, j-1$. The mapping θ is clearly one-to-one and onto.

The mapping $\theta: \Delta \to \tilde{\Delta} = \theta\Delta$ is defined in a similar fashion. We leave the details to the reader.

Let u and v be control functions on $[\tau, T]$ for Players I and II and let ϕ denote the corresponding trajectory with initial value (τ, ξ), i.e.

$$\phi'(t) = f(t, \phi(t), u(t), v(t)), \qquad \phi(\tau) = \xi.$$

Let $\bar{\phi}(s) \equiv \bar{\phi}(s(t))$ denote the trajectory corresponding to $\bar{u} = \theta u$, $\bar{v} = \theta v$ and initial value (τ', ξ'); i.e.

$$\bar{\phi}'(s) = f(s, \bar{\phi}(s), \bar{u}(s), \bar{v}(s)), \qquad \phi(\tau') = \xi'.$$

Note that in general $\bar{\phi}(s(t)) \neq \phi(t)$.

LEMMA 6.4. *Let* (i)-(iii) *of Assumption* I *hold, and let* ξ *and* ξ' *lie in a bounded set* X. *Then there exists a nonnegative function* η, *defined for all* $\rho > 0$ *such that* $\eta(\rho) \to 0$ *as* $\rho \to 0$ *and such that for all* ξ, ξ' *in* X, τ, τ' *in* $[T_0, T]$

$$\max_{\tau \leqq t \leqq T} |\bar{\phi}(s(t)) - \phi(t)| \leqq \eta(\rho), \qquad \rho = |\tau - \tau'| + |\xi - \xi'|. \tag{6.7}$$

If (i)-(iii) *of Assumption* I′ *hold, then there exists a constant* K *such that for all* ξ, ξ' *in* X, *and all* τ, τ' *in* $[T_0, T]$

$$\max_{\tau < t < T} |\bar{\phi}(s(t)) - \phi(t)| \leqq K[|\tau - \tau'| + |\xi - \xi'|]. \tag{6.8}$$

The proof of the first conclusion is straightforward and can be found in Friedman [5, Lemma 2.6.1]. The proof of the second statement involves obvious modifications of the proof of the first statement.

The following corollary of Lemma 6.4 will be important for us.

LEMMA 6.5. *Let τ, τ' belong to $[T_0 T]$ and let ξ, ξ' lie in a bounded set X. Let* (i)-(iii) *of Assumption* I *hold. Then for every motion $\phi[\ , \tau, \xi, \Gamma, \Delta]$ there exists a motion $\phi[\ , \tau', \xi', \theta\Gamma, \theta\Delta]$ such that*

$$\max_{\tau \leqq t \leqq T} |\phi[t, \tau, \xi, \Gamma, \Delta] - \phi[s(t), \tau', \xi', \theta\Gamma, \theta\Delta]| \leqq \eta(\rho). \tag{6.9}$$

If (i)-(iii) *of Assumption* I′ *hold, then there exists a constant K such that*

$$\max_{\tau \leqq t \leqq T} |\phi[t, \tau, \xi, \Gamma, \Delta] - \phi[s(t), \tau', \xi', \theta\Gamma, \theta\Delta]| \leqq K(|\tau - \tau'| + |\xi - \xi'|). \tag{6.10}$$

LEMMA 6.6. *Let $\phi[\ , \tau, \xi, \Gamma, \Delta]$ be a set of motions such that $(\tau, \xi) \to (T, \xi_0)$. Then $\phi[T, \tau, \xi, \Gamma, \Delta] \to \xi_0$, uniformly with respect to Γ, Δ, ξ_0, for ξ_0 in a bounded set.*

For nth stage trajectories $\phi(\ , \tau, \xi_n, u_n, v_n)$

$$\phi_n(T) - \xi_0 = (\xi_n - \xi) + (\xi - \xi_0) + \int_\tau^T f(s, \phi_n(s), u_n(s), v_n(s))\, ds.$$

Letting $n \to \infty$, we get

$$|\phi[T, \tau, \xi, \Gamma, \Delta] - \xi_0| \leqq |\xi - \xi_0| + K|T - \tau|,$$

and the result follows.

7. Continuity properties of W^+ and W^-. In § 4 we defined the upper value $W^+(t_0, x_0)$ and the lower value $W^-(t_0, x_0)$ for a game with initial point (t_0, x_0). In this section we shall study the continuity properties of W^+ and W^- as functions of the initial point, which we now designate as (t, x) rather than (t_0, x_0).

THEOREM 7.1. *Let Assumption* I *hold. Then W^+ and W^- are continuous on $[T_0, T) \times R^n$. If Ω is a bounded set in R^n, then W^+ and W^- are uniformly continuous on $[T_0, T) \times \Omega$. If we set $W^-(T, x) = g(x)$ and $W^+(T, x) = g(x)$, then W^+ and W^- are continuous on $[T_0, T] \times R^n$.*

Proof. We first prove that W^- is uniformly continuous on $[T_0, T) \times \Omega$. The proof for W^+ is similar. Let (τ, ξ) and (τ', ξ') be two points in $[T_0, T) \times \Omega$. Then

$$W^-(\tau, \xi) = \sup_\Gamma \inf_\Delta g(\Phi[T, \tau, \xi, \Gamma, \Delta])$$

and

$$W^-(\tau', \xi') = \sup_{\Gamma'} \inf_{\Delta'} g(\Phi[T, \tau', \xi', \Gamma, \Delta])$$

where Γ', Δ' denote strategies over $[\tau', T]$ and Γ, Δ denote strategies over $[\tau, T]$. Since the mapping θ defined in § 6 is a one-to-one mapping from the strategies on $[\tau, T]$ to the strategies on $[\tau', T]$, we may write

$$W^-(\tau', \xi') = \sup_\Gamma \inf_\Delta g(\Phi[T, \tau', \xi', \theta\Gamma, \theta\Delta]).$$

From (6.9) of Lemma 6.5, from the corollary to Lemma 6.2, and from the continuity of g, we get that for fixed Γ,

$$\inf_\Delta g(\phi[T, \tau', \xi', \theta\Gamma, \theta\Delta]) = \inf_\Delta g(\phi[T, \tau, \xi, \Gamma, \Delta]) + E(\tau, \xi, \tau', \xi', \Gamma),$$

where $|E| < \bar{\eta}(|\tau - \tau'| + |\xi - \xi'|)$, $\bar{\eta}(\rho) \to 0$ as $\rho \to 0$ and $\bar{\eta}$ depends on Ω. Hence,

$$|W^-(\tau, \xi) - W^-(\tau', \xi')| \leqq \eta(|\tau - \tau'| + |\xi - \xi'|),$$

and the uniform continuity of W^- on $[t_0, T) \times \Omega$ is established.

Since W^+ and W^- are uniformly continuous on all sets of the form $[0, T) \times A$, where $A = \{x: |x| \leqq a\}$, it follows that W^+ and W^- are continuous on $[0, T) \times R^n$.

To prove the last statement we must show that if $(\tau, \xi) \to (T, \xi_0)$, then $W^-(\tau, \xi) \to g(\xi_0)$. From the definition of W^- and Lemma 6.6 we have that

$$W^-(\tau, \xi) = \sup_{\Gamma} \inf_{\Delta} g(\Phi[T, \tau, \xi, \Gamma, \Delta])$$

$$= \sup_{\Gamma} \inf_{\Delta} g(\xi_0 + \varepsilon(\tau, \xi, \xi_0, \Gamma, \Delta)),$$

where $\varepsilon \to 0$ as $(\tau, \xi) \to (T, \xi_0)$, uniformly with respect to (Γ, Δ). The result now follows from the continuity of g.

THEOREM 7.2. *Let Assumption* I' *hold. Then* W^+ *and* W^- *are uniformly Lipschitz continuous on bounded subsets of* $[T_0, T) \times R^n$.

The proof is similar to that of Theorem 7.1, except that we use (6.10) of Lemma 6.5 and utilize the Lipschitz continuity of g.

8. The sets $C(v_0)$ and $C(v^0)$. Let (t_0, x_0) be the initial point of the game. Let

$$v_0 = W^-(t_0, x_0), \qquad v^0 = W^+(t_0, x_0). \tag{8.1}$$

Let $W^+(T, x) = g(x)$ and let $W^-(T, x) = g(x)$. Let

$$\begin{aligned} C(v_0) &= \{(\tau, \xi): t_0 \leqq \tau \leqq T, \xi \in R^n, W^-(\tau, \xi) \leqq v_0\} \\ C(v^0) &= \{(\tau, \xi): t_0 \leqq \tau \leqq T, \xi \in R^n, W^+(\tau, \xi) \geqq v^0\}. \end{aligned} \tag{8.2}$$

Sets equivalent to $C(v_0)$ and $C(v^0)$ were introduced by Krasovskii and Subbotin in [9]. The properties of these sets given in Lemmas 8.1 to 8.3 were also first stated in [9]. Our proofs will be different from those in [9].

LEMMA 8.1. *Let Assumption* I *hold. Then the sets* $C(v_0)$ *and* $C(v^0)$ *are closed.*

This lemma is an immediate consequence of the definitions of $C(v_0)$ and $C(v^0)$ and Theorem 8.1.

LEMMA 8.2. *A point* (τ, ξ) *belongs to* $C(v_0)$ *if and only if for every* $\varepsilon > 0$ *and every strategy* Γ *there exists a strategy* $\Delta(\Gamma)$ *and a motion* $\phi[\ , \tau, \xi, \Gamma, \Delta(\Gamma)]$ *such that* $g(\phi[T, \tau, \xi, \Gamma, \Delta(\Gamma)]) < v_0 + \varepsilon$.

The "if" statement follows from the definition of $W^-(\tau, \xi)$, since for each Γ we have $\inf\{\Delta: P(\tau, \xi, \Gamma, \Delta)\} \leqq v_0$. If the "only if" statement were false, there would exist a Γ_0 and an $\varepsilon_0 > 0$ such that for all Δ and all motions $\phi[\ , \tau, \xi, \Gamma_0, \Delta]$, we would have $g(\phi[T, \tau, \xi, \Gamma_0, \Delta]) > v_0 + \varepsilon_0 > v_0$. This implies that $W^-(\tau, \xi) > v_0$, and the "only if" statement is proved.

In the Krasovskii-Subbotin terminology, Lemma 8.2 states that "escape" from $C(v_0)$ by Player I is not possible.

Remark 8.2. A result analogous to Lemma 8.2 holds for $C(v^0)$.

LEMMA 8.3. *Let* (τ, ξ) *be a point of* $C(v_0)$. *Let* t_1 *satisfy* $\tau < t_1 < T$ *and let* u *be any control for Player* I *on* $[\tau, t_1]$. *Then there exists a relaxed control* $\zeta = \zeta(u)$ *such that the relaxed trajectory* $\psi(\ , \tau, \xi, u, \zeta)$ *has the property that* $(t_1, \psi(t_1)) \in C(v_0)$.

Remark 8.3. Similarly, if (τ, ξ) belongs to $C(v^0)$, then for every $\tau < t_1 < T$ and control v for Player II on $[\tau, t_1]$, there exists a relaxed control η such that the relaxed trajectory $\psi(\ , \tau, \xi, \eta, v)$ has the property that $(t_1, \psi(t_1)) \in C(v^0)$.

In the terminology of Krasovskii and Subbotin [9], Lemma 8.3 says that $C(v_0)$ is v-stable.

Proof. Suppose that the result were false. Then there would exist a t_1 in (τ, T) and a control u defined on $[\tau, t_1]$ such that for any ζ, the relaxed trajectory $\psi(\ , \tau, \xi, u, \zeta)$ would have the property that $(t_1, \psi(t_1)) \notin C(v_0)$. We shall show that this leads to a contradiction of the assumption that $(\tau, \xi) \in C(v_0)$.

Let Γ_T be an arbitrary strategy for Player I in the game over the time interval $[t_1, T]$. We shall associate to Γ_T a strategy $\hat{\Gamma} = \hat{\Gamma}(\Gamma_T)$ in the game over the time interval $[\tau, T]$ by defining $\hat{\Gamma}_n = (\hat{\Gamma}_{n,1}, \cdots, \hat{\Gamma}_{n,n})$ for any sequence of partitions $\{\pi_n\}$ of $[\tau, T]$ with $\|\pi_n\| \leqq K/n$, where K is a constant independent of n. Let $\{\pi_n\}$ be such a sequence of partitions with partition points of π_n, $\tau = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_n = T$. Let $I_{i+1} = [\tau_i, \tau_{i+1})$, $i = 0, 1, \cdots, n-1$. Let $j = j(n)$ be the integer such that $t_1 \in I_{j+1}$. The partition π_n of $[\tau, T]$ induces a partition π'_m of $[t_1, T]$ with partition points $t_1 = \tau'_0 < \tau'_1 < \cdots < \tau'_m = T$, where $\tau'_1 = \tau_{j+1}$, $\tau'_2 = \tau_{j+2}, \cdots, \tau'_m = \tau_{j+m} = \tau_n = T$. Note that m depends on n. Since $m < n$, $\|\pi'_m\| \leqq K/m$. For $i > j+1$, the interval I_i of the partition π_n is the interval I'_k of the partition π'_m, where $k = i - j$. The interval I'_1 is the interval $[t_1, \tau_{j+1})$.

For each n, $\hat{\Gamma}_n$ will be the concatenation of the constant nth stage strategy $u(t)$ on $[\tau, t_1]$ and an mth stage strategy $\Gamma_{T,m}$ on $[t_1, T]$. More precisely, we define $\hat{\Gamma}_n = (\hat{\Gamma}_{n,1}, \cdots, \hat{\Gamma}_{n,n})$ as follows. Let $\hat{\Gamma}_{n,1} = u$. For $2 \leqq i \leqq j$ let

$$\hat{\Gamma}_{n,i}(\mathcal{Y}_1 \times \mathcal{Z}_1 \times \cdots \times \mathcal{Y}_{i-1} \times \mathcal{Z}_{i-1}) = u.$$

For $i = j+1$, if $t_1 = \tau_j$, let

$$\hat{\Gamma}_{n,j+1}(\mathcal{Y}_1 \times \mathcal{Z}_1 \times \cdots \times \mathcal{Y}_j \times \mathcal{Z}_j) = \Gamma_{T,m,1};$$

if $\tau_j < t_1 < \tau_{j+1}$ let,

$$(\hat{\Gamma}_{n,j+1}(\mathcal{Y}_1 \times \mathcal{Z}_1 \times \cdots \times \mathcal{Y}_j \times \mathcal{Z}_j))(t) = u(t), \qquad \tau_j \leqq t < t_1,$$

$$(\hat{\Gamma}_{n,j+1}(\mathcal{Y}_1 \times \mathcal{Z}_1 \times \cdots \times \mathcal{Y}_j \times \mathcal{Z}_j))(t) = \Gamma_{T,m,1}(t), \qquad t_1 \leqq t < \tau_{j+1}.$$

For $i = j+k$, $k > 1$ let

$$\hat{\Gamma}_{n,j+k}(\mathcal{Y}_1 \times \mathcal{Z}_1 \times \cdots \times \mathcal{Y}_{j+k-1} \times \mathcal{Z}_{j+k-1}) = \Gamma_{T,m,k}(\mathcal{Y}_j \times \mathcal{Z}_j \times \cdots \times \mathcal{Y}_{j+k-1} \times \mathcal{Z}_{j+k-1}).$$

Let $\Delta = \{\Delta_n\}$ be an arbitrary strategy for Player II in the game over $[\tau, T]$. Let $\phi[\ , \tau, \xi, \hat{\Gamma}, \Delta]$ be a motion in the game over $[\tau, T]$ resulting from $(\hat{\Gamma}, \Delta)$ and a sequence of partitions $\{\pi_n\}$. Let $\{(u_n, v_n)\}$ be the sequence of outcomes of $(\hat{\Gamma}, \Delta)$. Let $\{\phi_{n_k}(\ , \tau, \xi_{n_k}, u_{n_k}, v_{n_k})\}$ be the sequence of trajectories converging uniformly to $\phi[\ , \tau, \xi, \hat{\Gamma}, \Delta]$. Let $x_{1n_k} = \phi_{n_k}(t_1, \tau, \xi_{n_k}, u_{n_k}, v_{n_k})$ and let $x_1 = \lim_{k \to \infty} x_{1n_k}$. Then $x_1 = \phi[t_1, \tau, \xi, \hat{\Gamma}, \Delta]$.

We now define a strategy $\Delta_T = \Delta_T(\Delta)$ such that there exists a motion $\phi[\ , t_1, x_1, \Gamma_T, \Delta_T]$ satisfying

$$\phi[t, t_1, x_1, \Gamma_T, \Delta_T] = \phi[t, \tau, \xi, \hat{\Gamma}, \Delta] \tag{8.3}$$

for all $t_1 \leqq t \leqq T$.

As noted before, a partition π_n of $[\tau, T]$ induces a partition π'_m of $[t_1, T]$, where the integer m depends on n. There exists a subsequence $\{n_{k_j}\}$ of $\{n_k\}$, which for typographical convenience we relabel as $\{n_p\}$, such that $m(n_{p+1}) > m(n_p)$. Let $\{m_p\} \equiv \{m(n_p)\}$. Let $\{\pi_m\}$ be a sequence of partitions of $[t_1, T]$ into intervals $I_1, \cdots, I_m$ such that if $m = m_p$ for some p, then $\pi_{m_p} = \pi'_{m_p}$. Recall that the outcome of $(\hat{\Gamma}_{n_p}, \Delta_{n_p})$ on $[t_1, T]$ is (u_{n_p}, v_{n_p}).

We now define $\Delta_T = \Delta_T(\Delta)$ on $[\tau_1, T]$. Let $\bar{v}$ be a fixed, but arbitrary, control for Player II on $[\tau_1, T]$. If $m \neq m_p$ for all p, let $v_m \equiv \bar{v}$; if $m = m_p$ for some p, let $v_m \equiv v_{m_p} \equiv v_{n_p}$. Let Δ_T^c be the constant component strategy corresponding to the sequence $\{v_m\}$.

By our construction we have that for $p = 1, 2, 3, \cdots$ the sequence of outcomes $\{(u_{m_p}, v_{m_p})\}$ on $[\tau_1, T]$ resulting from $\{(\Gamma_{T,m_p}, \Delta_{T,m_p})\}$ will be the sequence of outcomes $\{(u_{n_p}, v_{n_p})\}$. Hence for $p = 1, 2, 3, \cdots$ and $t_1 \leqq t \leqq T$ we have

$$\phi_{m_p}(t, t_1, x_{1m_p}, u_{m_p}, v_{m_p}) = \phi_{n_p}(t, \tau, \xi_{n_p}, u_{n_p}, v_{n_p}),$$

where $x_{1m_p} = x_{1n_p}$. If we let $p \to \infty$, we obtain (8.3).

Since on the interval $[\tau, t_1]$, $\hat{\Gamma}$ is the constant strategy with value u, it follows from Lemma 8.1 that there is a relaxed control ζ and a corresponding relaxed trajectory $\psi(\ , \tau, \xi, u, \zeta)$ such that for $t \in [\tau, t_1]$, we have $\phi[t, \tau, \xi, \hat{\Gamma}, \Delta] = \psi(t, \tau, \xi, u, \zeta)$. Hence by our assumption concerning u, the point (t_1, x_1) where $x_1 = \phi[t_1, \tau, \xi, \hat{\Gamma}, \Delta]$, does not belong to $C(v_0)$.

The last result can be stated in the following way. If H_{t_1} denotes the hyperplane $t = t_1$ in R^{n+1}, then

$$(t_1 X \Phi[t_1, \tau, \xi, \hat{\Gamma}]) \cap (C(v_0) \cap H_{t_1}) = \phi. \tag{8.4}$$

From (8.3) we have that for any motion $\phi[\ , \tau, \xi, \hat{\Gamma}, \Delta]$ over the interval $[\tau, T]$ there exists a motion $\phi[\ , t_1, x_1, \Gamma_T, \Delta_T]$ such that $x_1 = \phi[t_1, \tau, \xi, \hat{\Gamma}, \Delta]$ and

$$g(\phi[T, \tau, \xi, \hat{\Gamma}, \Delta]) = g(\phi[T, t_1, x_1, \Gamma_T, \Delta_T]).$$

Recall that $\hat{\Gamma} = \hat{\Gamma}(\Gamma_T)$ and $\Delta_T = \Delta_T(\Delta)$. Hence for fixed Γ_T and corresponding fixed $\hat{\Gamma}$, we have

$$\inf_{\Delta} g(\Phi[T, \tau, \xi, \hat{\Gamma}, \Delta]) \geqq \inf_{(x_1, \Delta_*)} g(\Phi[T, t_1, x_1, \Gamma_T, \Delta_*]), \tag{8.5}$$

where by the infimum over (x_1, Δ_*) we mean that we take the infimum over all initial positions $x_1 \in \Phi[t_1, \tau, \xi, \hat{\Gamma}]$ and all strategies Δ_* for Player II in the game over $[t_1, T]$.

Let

$$\mu \equiv \inf_{(x_1, \Delta_*)} g(\Phi[T, t_1, x_1, \Gamma_T, \Delta_*]). \tag{8.6}$$

Then there exists sequences $\{x_{1n}\}$ and $\{\Delta_*(n)\}$ and a corresponding sequence of motions $\phi[\ , t_1, x_{1n}, \Gamma_T, \Delta_*(n)]$ such that $g(\phi[T, t_1, x_{1n}, \Gamma_T, \Delta_*(n)]) \to \mu$. Since all of the points x_{1n} lie in the compact set $\Phi[t_1, \tau, \xi, \hat{\Gamma}]$, it follows from Lemma 6.2 that there is a subsequence of the sequence of motions, which we again label as $\phi[\ , t_1, x_{1n}, \Gamma_T, \Delta_*(n)]$, that converges to some motion $\phi[\ , t_1, \bar{x}_1, \Gamma_T, \bar{\Delta}_*]$, where $\bar{x}_1$ is in $\Phi[t_1, \tau, \xi, \hat{\Gamma}]$. Since $g(\phi[T, t_1, x_{1n}, \Gamma_T, \Delta_*(n)]) \to g(\phi[T, t_1, \bar{x}_1, \Gamma_T, \bar{\Delta}_*])$, we have that $g(\phi[T, t_1, \bar{x}_1, \Gamma_T, \bar{\Delta}_*]) = \mu$. Hence $\inf_{\Delta_*} g(\Phi[T, t_1, \bar{x}_1, \Gamma_T, \Delta_*]) \leqq \mu$. But by (8.6) $\inf_{\Delta_*} g(\Phi[T, t_1, \bar{x}_1, \Gamma_T, \Delta_*]) \geqq \mu$. Hence

$$\mu = \inf_{\Delta_*} g(\Phi[T, t_1, \bar{x}_1, \Gamma_T, \Delta_*]).$$

Combining this with (8.5) and (8.6) gives

$$\inf_{\Delta} g(\Phi[T, \tau, \xi, \hat{\Gamma}, \Delta]) \geqq \inf_{\Delta_*} g(\Phi[T, t_1, \bar{x}_1, \Gamma_T, \Delta_*]),$$

where Δ ranges over all strategies for Player II in the game over $[\tau, T]$ and Δ_* ranges over all strategies for Player II in the game over $[t_1, T]$. Thus

$$\sup_{\hat{\Gamma}} \inf_{\Delta} g(\Phi[T, \tau, \xi, \hat{\Gamma}, \Delta]) \geqq \sup_{\Gamma_T} \inf_{\Delta_*} g(\Phi[T, t_1, \bar{x}_1, \Gamma_T, \Delta_*]), \tag{8.7}$$

where Γ_T ranges over all strategies for Player I in the game over the interval $[t_1, T]$.

The right-hand side of (8.7) is equal to $W^-(t_1, \bar{x}_1)$. Since $\bar{x}_1 \in \Phi[t_1, \tau, \xi, \hat{\Gamma}]$ it follows from (8.4) that $W^-(t_1, \bar{x}_1) > v_0$. Hence

$$\sup_{\Gamma} \inf_{\Delta} g(\Phi[T, \tau, \xi, \Gamma, \Delta]) \geqq \sup_{\hat{\Gamma}} \inf_{\Delta} g(\Phi[T, \tau, \xi, \hat{\Gamma}, \Delta]) > v_0, \tag{8.8}$$

where Γ ranges over all strategies for Player I in the game over $[\tau, T]$. But this contradicts the assumption that (τ, ξ) belongs to $C(v_0)$, and the lemma is proved.

If the graph of an nth stage trajectory, or relaxed trajectory or motion lies in $C(v_0)$ then we shall say that the trajectory or the motion lies in $C(v_0)$.

LEMMA 8.4. *Let (τ, ξ) belong to $C(v_0)$. Let there exist a strategy Δ_0 such that for any strategy Γ all motions $\phi[\ , \tau, \xi, \Gamma, \Delta_0]$ on $[\tau, T]$ lie in $C(v_0)$. Then for any strategy Γ and all motions $\phi[\ , \tau, \xi, \Gamma, \Delta_0]$, the inequality $g(\phi[T, \tau, \xi, \Gamma, \Delta_0]) \leqq v_0$ holds.*

If the lemma were false, there would exist a strategy Γ_0 and a motion $\phi[\ , \tau, \xi, \Gamma_0, \Delta_0]$ such that $c \equiv g(\phi[T, \tau, \xi, \Gamma_0, \Delta_0]) > v_0$. Let $\gamma = c - v_0$. It follows from the continuity of g, from (2.6) and the definition of motion that there exists a $\delta > 0$ such that for any two pairs of strategies (Γ, Δ) and (Γ', Δ') and any motions $\phi[\ , t_1, x_1, \Gamma, \Delta]$, $\phi[\ , t_1, x_1, \Gamma', \Delta']$ the inequality

$$|g(\phi[T, t_1, x_1, \Gamma, \Delta]) - g(\phi[T, t_1, x_1, \Gamma', \Delta'])| < \frac{\gamma}{2} \tag{8.9}$$

holds whenever $|T - t_1| < \delta$. Select a t_1 satisfying $|T - t_1| < \delta$. Let $x_1 = \phi[t_1, \tau, \xi, \Gamma_0, \Delta_0]$. By Lemma 6.3 the segment of the motion $\phi[\ , \tau, \xi, \Gamma_0, \Delta_0]$ on the interval $[t_1, T]$ is again a motion, say $\phi[\ , t_1, x_1, \Gamma', \Delta']$. Therefore $g(\phi[T, t_1, x_1, \Gamma', \Delta']) = c$. But $(t_1, x_1) \in C(v_0)$ so by Lemma 8.2 there exists a strategy $\Delta(\Gamma')$ and a motion $\phi[\ , t_1, x_1, \Gamma', \Delta(\Gamma')]$ such that $g(\phi[T, t_1, x_1, \Gamma', \Delta(\Gamma')]) < v_0 + \gamma/2$. This contradicts (8.9), and the lemma is proved.

Remark 8.5. A result similar to Lemma 8.4 holds for the set $C(v^0)$.

9. A comparison of trajectories. In this section we compare two trajectories. This comparison was introduced by Krasovskii and Subbotin [9, § 14] and is crucial for both their development and ours.

DEFINITION 9.1. The *Isaacs condition* holds at a point (t, x) in $[T_0, T_1] \times R^n$, if for all vectors s in R^n

$$\max_y \min_z \langle s, f(t, x, y, z) \rangle = \min_z \max_y \langle s, f(t, x, y, z) \rangle, \tag{9.1}$$

where y ranges over the set Y and z ranges over the set Z. We say that the Isaacs condition holds on $[T_0, T_1] \times R^n$ if it holds at all points (t, x) of $[T_0, T_1] \times R^n$.

We note that we can write max and min instead of sup and inf because f is continuous in (y, z) and Y and Z are compact.

Remark 9.1. If the problem originally is one with integral payoff, i.e. $f^0 \not\equiv 0$, then the Isaacs condition (9.1) when written in terms of the original problem would read

$$\max_y \min_z \langle \tilde{s}, \tilde{f}(t, x, y, z) \rangle = \min_z \max_y \langle \tilde{s}, \tilde{f}(t, x, y, z) \rangle, \tag{9.1}'$$

where $\tilde{s}=(s^0, s^1, \cdots, s^n)$ and $\tilde{f}=(f^0, f^1, \cdots, f^n)$. We shall show later that we need (9.1)′ to hold only for vectors $\tilde{s}$ with $s^0>0$. With this restriction (9.1)′ is equivalent to the Isaacs condition

$$(9.1)'' \quad \max_y \min_z [f^0(t, x, y, z)+\langle s, f(t, x, y, z)\rangle] = \min_z \max_y [f^0(t, x, y, z)+\langle s, f(t, x, y, z)\rangle]$$

used by Friedman and Elliott and Kalton.

The Isaacs condition was introduced by Isaacs in his early studies of differential games.

An equivalent formulation of the Isaacs condition is that for each s in R^n the zero sum game with payoff $\langle s, f(t, x, y, z)\rangle$ in which Player I, the maximizer, chooses an element $y \in Y$ and Player II, the minimizer, chooses an element $z \in Z$ has a saddle point (y^*, z^*). We then have

$$\langle s, f(t, x, y^*, z^*)\rangle = \max_y \min_z \langle s, f(t, x, y, z)\rangle = \min_z \max_y \langle s, f(t, x, y, z)\rangle,$$

and $\langle s, f(t, x, y^*, z^*)\rangle$ is the value of the game. We shall call this game over Y Z, the *local game* at (t, x, s). We also have that

$$\langle s, f(t, x, y, z^*)\rangle \leqq \langle s, f(t, x, y^*, z^*)\rangle \leqq \langle s, f(t, x, y^*, z)\rangle \tag{9.2}$$

for all y in Y and z in Z.

LEMMA 9.1. *Let Assumption* I *hold. Let* (τ_1, ξ_1) and (τ_1, x_1) *be points in a fixed bounded region* B *contained in* $[T_0, T_1]\times R^n$, *and let the Isaacs condition hold at* (τ_1, ξ_1). *Let* $s^*=\xi_1-x_1$. *Let* (y^*, z^*) *be a saddle point for the local game at* (τ_1, ξ_1, s^*) *with payoff* $\langle s^*, f(\tau_1, \xi_1, y, z)\rangle$. *Let* u *be a control for Player* I *on* $[\tau_1, T]$. *Let* ϕ *be an absolutely continuous function on* $[\tau_1, T]$ *satisfying*

$$\phi'(t)=f(t, \phi(t), u(t), z^*), \qquad \phi(\tau_1)=\xi_1, \tag{9.3}$$

and let ψ *be an absolutely continuous function on* $[\tau_1, T]$ *satisfying*

$$\begin{aligned}\psi'(t) &= f(t, \psi(t), y^*, \zeta(t)) \\ &= \int_Z f(t, \psi(t), y^*, z)\, d\zeta\,(t), \qquad \psi(\tau_1)=x_1.\end{aligned} \tag{9.4}$$

Let

$$\rho(t)=|\phi(t)-\psi(t)|, \qquad \tau_1 \leqq t \leqq T.$$

Then there exists a nondecreasing function E *defined on* $[0, T-\tau_1]$ *such that* $E(\delta)\to 0$ *as* $\delta\to 0$ *and a positive constant* β *such that for all* $0\leqq\delta\leqq T-\tau_1$,

$$\rho^2(\tau_1+\delta)\leqq\rho^2(\tau_1)(1+\beta\delta)+E(\delta)\delta,$$

for all (τ_1, ξ_1) *and* (τ_1, x_1) *in* B.

This lemma is proved in [9, § 14].

10. Extremal strategies. In this section we shall use the "extremal aiming" strategies introduced by Krasovskii and Subbotin [9] to define our "extremal" strategies. We shall show that if Player I uses a strategy Γ_e extremal to $C(v^0)$, then all motions $\phi[\ , t_0, x_0, \Gamma_e, \Delta]$ will lie in $C(v^0)$. If Player II uses a strategy Δ_e extremal to $C(v_0)$, then all motions $\phi[\ , t_0, x_0, \Gamma, \Delta_e]$ will lie in $C(v_0)$.

Let U be a function defined on $[t_0, T]\times R^n$ with values $U(t, x)$ in Y. The function U determines a strategy $\Gamma=\Gamma(U)$ in the game with initial point (t_0, x_0), as follows.

Let π_n be the nth partition of $[t_0, T]$ and let $I_1, \cdots, I_n$ denote the partition intervals, $I_j = [\tau_{j-1}, \tau_j)$ $j = 1, \cdots, n$. Define

$$\Gamma_{n,1}(t) = U(t_0, x_0) \quad \text{for } t \in I_1.$$

To define $\Gamma_{n,j}, j = 2, \cdots, n$ we must assign a control u_j on I_j for Player II for each $(u_1, v_1, \cdots, u_{j-1}, v_{j-1})$, where for $i = 1, \cdots, j-1$, u_i is a control for Player I on I_i and v_i is a control for Player II on I_i. The function u defined on $[t_0, \tau_{j-1})$ by $u(t) = u_i(t)$ for $t \in I_i$, $i = 1, \cdots, j-1$ is a control for Player I on $[t_0, \tau_{j-1})$. Similarly, the function v defined on $[t_0, \tau_{j-1})$ by $v(t) = v_i(t)$ for $t \in I_i$, $i = 1, \cdots, j-1$, is a control for Player II on $[t_0, \tau_j)$. Under Assumption II, the differential equation

$$x' = f(t, x, u(t), v(t)), \qquad x(t_0) = x_0$$

has a unique solution $\phi(\cdot) = \phi(\cdot, t_0, x_0, u, v)$ on the interval $[t_0, \tau_{j-1}]$. We define

$$(\Gamma_{n,j}(u_1, v_1, \cdots, u_{j-1}, v_{j-1}))(t) = U(\tau_{j-1}, \phi(\tau_{j-1})), \qquad \tau_{j-1} \leqq t < \tau_j.$$

We shall call a strategy $\Gamma = \Gamma(U)$ determined this way a *feedback strategy* for Player I. We shall also call a function U defined on $[t_0, T] \times R^n$ with values in Y a *feedback strategy.*

Given a function V defined on $[t_0, T]$ with values in Z, we can associate to V a strategy $\Delta(V)$ in a way similar to that used to define $\Gamma(U)$. We shall also call both the function V and the strategy $\Delta(V)$, *feedback strategies* for Player II.

We now shall follow Krasovskii and Subbotin [9] and shall define a feedback strategy V_e extremal to the set $C(v_0)$. A similar procedure will give a feedback strategy U_e extremal to $C(v^0)$. From U_e and V_e we can obtain feedback strategies Γ_e and Δ_e in the manner outlined above. Note that Γ_e and Δ_e will depend on the initial point (t_0, x_0). In order to define V_e and U_e we must assume that the Isaacs condition (9.1) holds at all points of $[t_0, T] \times R^n$.

If (t, x) belongs to $C(v_0)$, define $V_e(t, x) = z$ where z is any vector in Z. If $(t^*, x^*) \notin C(v_0)$, $t^* < T$, then by Lemma 8.3, with $(\tau, \xi) = (t_0, x_0)$, the set

$$S(t^*) = H(t^*) \cap C(v_0) \tag{10.1}$$

is not empty, where $H(t^*)$ is the hyperplane $t = t^*$. It follows from Lemma 8.1 that $S(t^*)$ is closed relative to $H(t^*)$. Hence there is at least one point w^* in $S(t^*)$ that is at minimum distance from x^* in $S(t^*)$. If w^* is not unique, select any w^*. Let $s^* = x^* - w^*$. Define

$$V_e(t^*, x^*) = z^*,$$

where (y^*, z^*) is any saddle point of the local game (t^*, x^*, s^*), with payoff $\langle s^*, f(t^*, x^*, y, z)\rangle$.

Remark 10.1. It can be shown that if the problem originally is one with integral payoff ($f^0 \not\equiv 0$) and the state variable is $\tilde{x} = (x^0, x)$ in R^{n+1}, then $s^{*0} \geqq 0$. (See Remark 9.1.)

LEMMA 10.1. *Let Assumption* I *hold. Let V_e be extremal to $C(v_0)$ and let $\Delta_e = \Delta_e(V_e)$ be the corresponding feedback strategy for the game with initial point (t_0, x_0). Then every motion $\phi[\ , t_0, x_0, \Gamma, \Delta_e]$ lies entirely in $C(v_0)$; i.e. $\{(t, x)\colon t_0 \leqq t \leqq T;\ x \in \phi[t, t_0, x_0, \Delta_e]\} \subseteq C(v_0)$.*

In the terminology of Krasovskii-Subbotin [9], the lemma asserts that Δ_e is a "railing" for the "stable bridge" $C(v_0)$. Our proof is simpler than that of the similar result in [9].

Suppose the lemma were false. Then there would exist a strategy Γ and a motion $\phi[\cdot]=\phi[\cdot, t_0, x_0, \Gamma, \Delta_\varepsilon]$ that did not lie entirely in $C(v_0)$. Since $(t_0, x_0)\in C(v_0)$, the set $C_t=\{t: t\in[t_0, T], (t, \phi[t, t_0, x_0, \Gamma, \Delta_0])\in C(v_0)\}$ is not empty. By Lemma 8.1, the set C_t is closed. Since $\phi[\ \ , t_0, x_0, \Gamma, \Delta_\varepsilon]$ does not lie entirely in $C(v_0)$, the complement of C_t relative to $[t_0, T]$ is not empty. Let $t_1=\inf\{t: t\in[t_0, T], t\notin C_t\}$. Since C_t is closed, $t_1\in C_t$ and $t_1>t_0$. The point t_1 can be considered as the first time at which $\phi[\ \ ]$ leaves $C(v_0)$.

For each $t_0\leqq t\leqq T$, let $S(t)$ be defined as in (10.1). Let $\varepsilon(t)$ denote the distance from the point $(t, \phi[t])$ to the set $S(t)$. Then $\varepsilon(t)$ is also the distance in the hyperplane $H(t)$ from $\phi[t]$ to $S(t)$. Let $\bar{\varepsilon}(t)$ denote the distance from $(t, \phi[t])$ to the set $C(v_0)$. Then $\bar{\varepsilon}(t)\leqq\varepsilon(t)$. Clearly, if $\varepsilon(t)=0$, then $\bar{\varepsilon}(t)=0$ and $(t, \phi[t])\in S(t)\subseteq C(v_0)$. If $\bar{\varepsilon}(t)=0$, then since $C(v_0)$ is closed, $(t, \phi[t])\in C(v_0)$. Hence $(t, \phi[t])\in S(t)$, and so $\varepsilon(t)=0$. Thus $\varepsilon(t)=0$ if and only if $\bar{\varepsilon}(t)=0$, and so $\varepsilon(t)=0$ if and only if $t\in C_t$. We shall arrive at a contradiction by showing that $\varepsilon(t)=0$ for all t in $[t_0, T)$. For $t\leqq t_1$, we already have $\varepsilon(t)=0$. Therefore, we need only consider $t>t_1$.

Let $\phi_n(\ \ )=\phi_n(\ \ , t_0, x_0, u_n, v_n)$, be a sequence of nth stage trajectories converging uniformly to $\phi[\ \ , t_0, x_0, \Gamma, \Delta_\varepsilon]$. For $t\geqq t_1$, let $\varepsilon_n(t)$ denote the distance between $(t, \phi_n(t))$ and $S(t)$. Clearly, $\lim \varepsilon_n(t)=\varepsilon(t)$. Therefore to establish the desired contradiction it suffices to show that

$$\lim_{n\to\infty}\varepsilon_n(t)=0 \quad \text{for } t_1<t<T. \tag{10.2}$$

Let $I_1, \cdots, I_n$, where $I_j=[\tau_{j-1}, \tau_j)$, $j=0, 1, \cdots, n$, be the intervals of the nth partition π_n. Let $\|\pi_n\|\leqq K/n$, where K is a constant independent of n. Let $k=k(n)$ denote the integer such that $t_1\in I_k=[\tau_{k-1}, \tau_k)$. We emphasize that k depends on n; for notational convenience, however, we shall write k instead of $k(n)$. As $n\to\infty$, $\tau_k\to t_1$. Since $\phi_n(\ \ )$ converges uniformly on $[t_0, T]$ to $\phi[\ \ ]$, $(\tau_k, \phi_n(\tau_k))\to(t_1, \phi[t_1])\in C(v_0)$, as $n\to\infty$.

Let $x_1=\phi[t_1]$. By Lemma 8.3, for each integer n, there exists a relaxed control ζ_n such that the relaxed trajectory $\psi_n(\ \ , t_1, x_1, u_n, \zeta_n)$ has the property that $(\tau_k, \psi_n(\tau_k))\in C(v_0)$. Hence

$$\varepsilon_n(\tau_k)\leqq\left|(\tau_k, \psi_n(\tau_k))-(\tau_k, \phi_n(\tau_k))\right|.$$

By the triangle inequality we get

$$\varepsilon_n(\tau_k)\leqq\left|(\tau_k, \psi_n(\tau_k))-(t_1 x_1)\right|+\left|(t_1, x_1)-(\tau_k, \phi_n(\tau_k))\right|.$$

From this we get,

$$\lim_{n\to\infty}\varepsilon_n(\tau_k)=0. \tag{10.3}$$

Let t be any point in (t_1, T). Then there exists an integer $n_0=n_0(t)$ such that for $n\geqq n_0$, $t\geqq\tau_k$. Let $n>n_0$. We shall estimate $\varepsilon_n(t)$.

Suppose $t\in[\tau_k, \tau_{k+1}]$. Let $x^*=\phi_n(\tau_k)$, let $t^*=\tau_k$ and let w^* be the point in $S(\tau_k)$ selected as being at minimum distance from x^* in $S(\tau_k)$ in the definition of $V_\varepsilon(t^*, x^*)=V_\varepsilon(\tau_k, \phi_n(\tau_k))$. Let $s^*=x^*-w^*$ and let y^* be any point in Y such that (y^*, z^*) is a saddle point for the local game $(t^*, x^*, s^*)=(\tau_k, \phi(\tau_k), s^*)$ with payoff $\langle s^*, f(\tau_k, \phi_n(\tau_k), y, z)\rangle$. By Lemma 8.3 there exists a relaxed control ζ such that the corresponding relaxed trajectory $\psi(\ \ )=\psi(\ \ , \tau_k, w^*, y^*, \zeta)$, has the property that $\psi(t, \tau_k, w^*, y^*, \zeta)\in C(v_0)$. Since for $s\geqq\tau_k$, $\phi_n(s, t_0, x_0, u_n, v_n)=\phi_n(s, \tau_k, \phi_n(\tau_k), u_n, v_n)$, we have that

$$\varepsilon_n(t)\leqq\left|\phi_n(t, \tau_k, \phi_n(\tau_k), u_n, v_n)-\psi(t, \tau_k, w^*, y^*, \zeta)\right|.$$

Note that for $t=\tau_k$, equality holds and that for $t\in I_k$, $v_n(t)=z^*$. Hence by Lemma 9.1 with $\tau_1=\tau_k$, $\xi_1=\phi_n(\tau_k)$, $x_1=w^*$, we have

$$\varepsilon_n^2(t)\leqq\varepsilon_n^2(\tau_k)(1+\beta(t-\tau_k))+E(t-\tau_k)(t-\tau_k).$$

If we set $\delta_n=K/n$, we have that for all t in $[\tau_k, \tau_{k+1}]$

$$\varepsilon_n^2(t)\leqq\varepsilon_n^2(\tau_k)(1+\beta\delta_n)+E(\delta_n)\delta_n. \tag{10.4}$$

By a similar argument, we get that for all t in $[\tau_{k+j}, \tau_{k+j+1}]$, $j=1,\cdots,n-1-k$,

$$\varepsilon_n^2(t)\leqq\varepsilon_n^2(\tau_{k+j})(1+\beta\delta_n)+E(\delta_n)\delta_n.$$

It now follows by induction that for t in $[\tau_{k+j}, \tau_{k+j+1}]$, $j=1,\cdots,n-1-k$,

$$\varepsilon_n^2(t)\leqq\varepsilon_n^2(\tau_k)(1+\beta\delta_n)^j+E(\delta_n)\delta_n\left(\sum_{i=0}^{j-1}(1+\beta\delta_n)^i\right).$$

Recall that $\delta_n=K/n$. Then for $n>n_0$,

$$\begin{aligned}\varepsilon_n^2(t)&\leqq\varepsilon_n^2(\tau_k)(1+\beta\delta_n)^n+E(\delta_n)\delta_n\sum_{i=0}^{n-1}(1+\beta\delta_n)^i\\&=\varepsilon_n^2(\tau_k)(1+\beta\delta_n)^n+E(\delta_n)((1+\beta\delta_n)^n-1)/\beta\\&\leqq\varepsilon_n^2(\tau_k)\,e^{\beta K}+E(\delta_n)(e^{\beta K}-1)/\beta.\end{aligned}$$

If we let $n\to\infty$, we get from (10.3) and $E(\delta_n)\to 0$ that $\varepsilon_n^2(t)\to 0$. Since $t>t_1$ is arbitrary, (10.2) is established and the lemma is proved.

Remark 10.2. We can define a feedback strategy U_e extremal to the set $C(v^0)$ in a manner analogous to that used to define V_e. We can then define a feedback strategy $\Gamma_e=\Gamma_e(U_e)$ for the game with initial point (t_0, x_0). This strategy will have the property that every motion $\phi[\ \ , t_0, x_0, \Gamma_e, \Delta]$ will lie entirely in $C(v^0)$.

We now point out a major advantage in using the Friedman definition of strategy over that of Krasovskii and Subbotin. In § 1 we showed that if the original problem has an integral payoff as in (1.2), then the problem can be transformed into a problem with state vector $\tilde{x}$ in R^{n+1} with terminal payoff $g(\tilde{x})$. The Krasovskii–Subbotin extremal strategies are functions V_e and U_e defined on $(t, \tilde{x})$-space. These functions must be replaced by functions V_e' and U_e' of (t, x), since the definition of strategy in [9] requires that a strategy in a game with integral payoff be a function of (t, x) and not of $(t, \tilde{x})=(t, x^0, x)$. This leads to difficulties, which in this writer's opinion were not adequately addressed in [9], but which were later taken up in [10].

For us also, V_e and U_e will be functions of $(t, \tilde{x})$, but for us this is not a problem. We convert these strategies into instructions for choosing a control on some interval, given the past history of the controls. The fact that the instructions are arrived at utilizing the x^0th coordinate is irrelevant.

11. The existence of value and saddle points. The principal result of this paper is Theorem 11.1, below. Its proof will be an easy consequence of the results in §§ 7 and 9.

THEOREM 11.1. *Let Assumption* I *hold and let the Isaacs condition* (9.1) *hold on* $[t_0, T]\times R^n$. *Then the game of fixed duration with initial point* (t_0, x_0) *and with payoff*

$$P(t_0, x_0, \Gamma, \Delta)=g(\phi[T, t_0, x_0, \Gamma, \Delta]), \tag{11.1}$$

has a value $W(t_0, x_0)$ *and a saddle point. The value* W *is continuous on* $[T_0, T)\times R^n$ *and* $W(t, x)\to g(x)$ *as* $t\to T$, *for all* x *in* R^n. *If* Ω *is a bounded set in* R^n, *then* W *is*

uniformly continuous on $[T_0, T) \times \Omega$. *If Assumption* I′ *holds, then* W *is uniformly Lipschitz continuous on bounded subsets of* $[T_0, T) \times R^n$.

Since (4.2) always holds, in order to prove the existence of a value it suffices to show that

$$W^+(t_0, x_0) \leqq W^-(t_0, x_0). \tag{11.2}$$

The statements concerning the continuity and Lipschitz continuity of the value will then follow from Theorems 7.1 and 7.2, as will the statement that $W(t, x) \to W(T, x)$ as $t \to T$.

We now establish (11.2). By Lemma 10.1 there exists a strategy Δ_e such that every motion $\phi[\ \ , t_0, x_0, \Gamma, \Delta_e]$ lies in $C(v_0)$. By Lemma 8.4 this implies that

$$g(\Phi[T, t_0, x_0, \Gamma, \Delta_e]) \leqq v_0 \quad \text{for all } \Gamma. \tag{11.3}$$

Hence

$$\sup_{\Gamma} g(\Phi[T, t_0, x_0, \Gamma, \Delta_e]) \leqq v_0 = W^-(t_0, x_0),$$

and so

$$W^+(t_0, x_0) = \inf_{\Delta} \sup_{\Gamma} g(\Phi[T, t_0, x_0, \Gamma, \Delta]) \leqq W^-(t_0, x_0).$$

This establishes (11.2).

To complete the proof of the theorem we must show the existence of a saddle point.

By virtue of Remarks 10.2 and 8.5, there exists a strategy Γ_e such that for all Δ

$$g(\Phi[T, t_0, x_0, \Gamma_e, \Delta]) \geqq W^+(t_0, x_0). \tag{11.4}$$

If we take $\Gamma = \Gamma_e$ in (11.3) and use the fact that we have already established the existence of the value, we get that $g(\Phi[T, t_0, x_0, \Gamma_e, \Delta_e]) \leqq W(t_0, x_0)$. If we take $\Delta = \Delta_e$ in (11.4), we get that $g(\Phi[T, t_0, x_0, \Gamma_e, \Delta_e]) \geqq W(t_0, x_0)$. Hence

$$g(\Phi[T, t_0, x_0, \Gamma_e, \Delta_e]) = W(t_0, x_0). \tag{11.5}$$

This equality, the equality $W(t_0, x_0) = W^+(t_0, x_0) = W^-(t_0, x_0)$ and (11.3) and (11.4) show that the pair (Γ_e, Δ_e) is a saddle point for the game.

12. Relàtionships among various definitions of value. In this section we assume that the reader is familiar with the Friedman definition of a differential game (see [5, §§ 1.3 to 1.6]). Friedman partitions the interval $[t_0, T]$ into n equal subintervals of length $\delta = (T - t_0)/n$, but shows that the same results would be obtained if one were to deal with sequences of arbitrary partitions π_δ, whose norms tend to zero.

We shall denote an upper δ strategy for Player I, as defined by Friedman, by

$$\Gamma_F^\delta = (\Gamma_F^{\delta,1}, \cdots, \Gamma_F^{\delta,n}).$$

Similarly, an upper δ strategy for Player II, as defined by Friedman, will be denoted by

$$\Delta_F^\delta = (\Delta_F^{\delta,1}, \cdots, \Delta_F^{\delta,n}).$$

Lower strategies in the sense of Friedman will be denoted by

$$\Gamma_{F,\delta} = (\Gamma_{F,\delta,1}, \cdots, \Gamma_{F,\delta,n}),$$

$$\Delta_{F,\delta} = (\Delta_{F,\delta,1}, \cdots, \Delta_{F,\delta,n}).$$

The payoffs $P[\Delta_{F,\delta}, \Gamma_F^\delta]$ and $P[\Delta_F^\delta, \Gamma_{F,\delta}]$ and the numbers V^δ, V_δ, V^+ and V^- are as defined in [5].

The following relationships hold between the Friedman upper value and our upper value and the Friedman lower value and our lower value.

THEOREM 12.1. *Let Assumption* I *hold. Then*

$$W^+(t_0, x_0) \geqq V^+(t_0, x_0), \qquad W^-(t_0, x_0) \leqq V^-(t_0, x_0).$$

Proof. The payoff is given by (11.1) for our game and by

$$P(y, z) = g(x(t, y, z))$$

for the Friedman game.

Let $\varepsilon > 0$ be given and let $W^- = W^-(t_0, x_0)$. Then by the definition of $W^-(t_0, x_0)$ there exists a strategy Γ_ε such that

$$\inf_{\Delta} P(t_0, x_0, \Gamma_\varepsilon, \Delta) > W^- - \varepsilon.$$

Hence for all Δ,

$$P(t_0, x_0, \Gamma_\varepsilon, \Delta) > W^- - \varepsilon. \tag{12.1}$$

For $n = 1, 2, 3, \cdots$, the nth stage strategy $\Gamma_{\varepsilon,n} = (\Gamma_{\varepsilon,n,1}, \cdots, \Gamma_{\varepsilon,n,n})$ is clearly a lower δ strategy in the sense of Friedman for Player I. We denote the lower δ strategy by $\Gamma_{F,\delta}(\varepsilon)$. Then for any upper δ strategy Δ_F^δ for Player II,

$$P[\Gamma_{F,\delta}(\varepsilon), \Delta_F^\delta] \leqq \sup_{\Gamma_{F,\delta}} P[\Gamma_{F,\delta}, \Delta_F^\delta].$$

Hence

$$\inf_{\Delta_F^\delta} P[\Gamma_{F,\delta}(\varepsilon), \Delta_F^\delta] \leqq \inf_{\Delta_F^\delta} \sup_{\Gamma_{F,\delta}} P[\Gamma_{F,\delta}, \Delta_F^\delta] = V_\delta.$$

Also, there exists a $\Delta_F^\delta(\varepsilon)$ such that

$$P[\Gamma_{F,\delta}(\varepsilon), \Delta_F^\delta(\varepsilon)] < \inf_{\Delta_F^\delta} P[\Gamma_{F,\delta}(\varepsilon), \Delta_F^\delta] + \varepsilon.$$

Combining the last two chains of inequalities gives

$$P[\Gamma_{F,\delta}(\varepsilon), \Delta_F^\delta(\varepsilon)] < V_\delta + \varepsilon. \tag{12.2}$$

Let $(u_\delta, \tilde{v}^\delta)$ be the outcome of $(\Gamma_{F,\delta}(\varepsilon), \Delta_F^\delta(\varepsilon))$. Let $\Delta_\varepsilon = \{\Delta_{\varepsilon,n}\}$ be the strategy such that $\Delta_{\varepsilon,n}$ assigns $v_n(t) = \tilde{v}^\delta(t)$ to I_j, the jth interval of the partition, for $j = 1, \cdots, n$. Then the outcome of $(\Gamma_{n,\varepsilon}, \Delta_{n,\varepsilon})$ will be $(u_n, v_n) = (u_\delta, \tilde{v}^\delta)$. Since $\Gamma_{\varepsilon,n}$ and $\Delta_{\varepsilon,n}$ determine the outcome (u_n, v_n) uniquely, there is a unique nth stage trajectory $\phi_n(\ \ , t_0, x_0, u_n, v_n)$ resulting from $(\Gamma_{\varepsilon,n}, \Delta_{\varepsilon,n})$, $n = 1, 2, \cdots$. Thus,

$$P[\Gamma_{F,\delta}(\varepsilon), \Delta_F^\delta(\varepsilon)] = g(\phi_n(T, t_0, x_0, u_n, v_n)). \tag{12.3}$$

Combining (12.3) with (12.2) gives

$$g(\phi_n(T, t_0, x_0, u_n, v_n)) < V_\delta + \varepsilon.$$

Recalling that $V_\delta \to V^-$ as $n \to \infty$, we obtain

$$\overline{\lim_{n\to\infty}}\, g(\phi_n(T, t_0, x_0, u_n, v_n)) \leqq V^- + \varepsilon. \tag{12.4}$$

But

$$P(t_0, x_0, \Gamma_\varepsilon, \Delta_\varepsilon) = g(\Phi[T, t_0, x_0, \Gamma_\varepsilon, \Delta_\varepsilon]) \leqq \overline{\lim_{n\to\infty}}\, g(\phi_n(T, t_0, x_0, u_n, v_n)). \tag{12.5}$$

From (12.4), (12.5) and (12.1), we get

$$W^- - \varepsilon < P(t_0, x_0, \Gamma_\varepsilon, \Delta_\varepsilon) \leqq V^- + \varepsilon.$$

Thus

$$W^- \leqq V^- + 2\varepsilon.$$

If we let $\varepsilon \to 0$, we obtain

$$W^- \leqq V^-.$$

A similar argument shows that $W^+ \geqq V^+$.

From Theorems 11.1 and 12.1 we immediately obtain the following result concerning the Friedman value.

THEOREM 12.2. *Let Assumption* I *hold and let the Isaacs condition* (9.1) *hold on* $[t_0, T] \times R^{n+1}$. *Then the Friedman game of fixed duration with initial point* (t_0, x_0) *has value* $V(t_0, x_0)$ *and* $V(t_0, x_0) = W(t_0, x_0)$.

We point out that we have obtained the existence of the Friedman value without recourse to arguments involving probability theory or the theory of partial differential equations.

Other definitions of upper and lower value for a differential game of fixed duration have been given by Fleming [4] and by Elliott and Kalton [2]. Let $F^+(t_0, x_0)$ and $F^-(t_0, x_0)$ denote the upper and lower Fleming values and let $U^+(t_0, x_0)$ and $U^-(t_0, x_0)$ denote the upper and lower Elliott-Kalton values. Friedman has shown under hypotheses weaker than those of Assumption I that $F^+(t_0, x_0) = V^+(t_0, x_0)$ and $F^-(t_0, x_0) = V^-(t_0, x_0)$ (see [6, p. 24]). Elliott and Kalton [2] have shown under hypotheses weaker than those of Assumption I that $U^+(t_0, x_0) = V^+(t_0, x_0)$ and $U^-(t_0, x_0) = V^-(t_0, x_0)$. The arguments used to establish the equalities $F^\pm = V^\pm$ and $U^\pm = V^\pm$ are nonelementary.

A differential game of fixed duration has value $F(t_0, x_0)$ in the sense of Fleming if $F^+(t_0, x_0) = F^-(t_0, x_0)$, in which case $F(t_0, x_0)$ is the common value of $F^+(t_0, x_0)$ and $F^-(t_0, x_0)$. A similar definition holds for the value $U(t_0, x_0)$ in the sense of Elliott and Kalton.

If Assumption I holds and the Isaacs condition (9.1) holds, then the game of fixed duration has a value in our sense and in the sense of Krasovskii-Subbotin [9]. Moreover, a saddle point (Γ_e, Δ_e) exists in our sense and a saddle point (U_e, V_e) exists in the Krasovskii-Subbotin sense. If we denote the Krasovskii-Subbotin value by $K(t_0, x_0)$ ([9, Chap. 4]), then in our notation,

$$K(t_0, x_0) = g(\Phi[T, t_0, x_0, U_e, V_e]).$$

(We have reversed the roles of U of V from those in Krasovskii and Subbotin. They use U for the minimizer and V for the maximizer.) From (11.5) we get

$$g(\Phi[T, t_0, x_0, \Gamma_e, \Delta_e]) = V(t_0, x_0).$$

From the definitions of Γ_e and Δ_e and the definitions of U_e and V_e it is clear that

$$g(\Phi[T, t_0, x_0, \Gamma_e, \Delta_e]) = g(\Phi[T, t_0, x_0, U_e, V_e]).$$

From the preceding discussion we obtain the following result.

THEOREM 12.3. *Let Assumption* I *hold and let the Isaacs condition* (9.1) *hold. Then*

$$K(t_0, x_0) = W(t_0, x_0) = V(t_0, x_0) = F(t_0, x_0) = U(t_0, x_0).$$

13. The Isaacs equation. Let Assumption I′ hold. Then by Theorem 7.2 W^+ and W^- are uniformly Lipschitz continuous on bounded subsets of $[0, T]\times R^n$. Hence, by Rademacher's theorem, W^+ and W^- are differentiable at almost all points of any open bounded set. Thus W^+ and W^- are differentiable almost everywhere.

LEMMA 13.1. *Let* (t, x) *be a point of differentiability of* W^-. *Then*

$$\max_y \min_z [W_t^-(t, x)+\langle W_x^-(t, x), f(t, x, y, z)\rangle]\leqq 0, \tag{13.1}$$

where the max *is taken over all* $y\in Y$ *and the* min *is taken over all* z *in* Z. *If* (t, x) *is a point of differentiability of* W^+ *then*

$$\min_z \max_y [W_t^+(t, x)+\langle W_x^+(t, x), f(t, x, y, z)\rangle]\geqq 0. \tag{13.2}$$

We shall prove the lemma for W^-; the proof for W^+ is similar, with appropriate modifications. To establish (13.1) it suffices to show that for all y in Y,

$$\min_z [W_t^-(t, x)+\langle W_x^-(t, x), f(t, x, y, z)\rangle]\leqq 0, \tag{13.3}$$

at each point of differentiability of W^-. If (13.3) were not true, there would exist a point of differentiability (t_1, x_1) and a vector $\bar{y}$ in Y such that

$$\min_z [W_t^-(t_1, x_1)+\langle W_x^-(t_1, x_1), f(t_1, x_1, \bar{y}, z)\rangle]>0. \tag{13.4}$$

We shall show that (13.4) leads to a contradiction.

It follows from Lemma 9.3, applied to the set $C(W^-(t_1, x_1))$, that for every $t_1<t<T$, there exists a relaxed control ζ_t such that if the relaxed trajectory $\psi(\ \ , t_1, x_1, \bar{y}, \zeta_t)$ is the solution of

$$x' = f(s, x, \bar{y}, \zeta_t) = \int_Z f(s, x, \bar{y}, z)\, d\zeta_t(s), \qquad x(t_1) = x_1, \tag{13.5}$$

then $W^-(t, \psi(t))\leqq W^-(t_1, x_1)$. Thus if $\Delta W^-(t)\equiv W^-(t, \psi(t)) - W^-(t_1, x_1)$, we have that for all $t_1\leqq t<T$

$$0\geqq \Delta W^-(t) = W_t^-(t_1, x_1)\Delta t+\langle W_x^-(t_1, x_1), \Delta x\rangle+\eta_1(t), \tag{13.6}$$

where $\Delta t = t - t_1$, $\Delta x = \psi(t) - x_1$ and

$$\eta_1(t)/(\Delta t^2+|\Delta x|^2)^{1/2}\to 0 \quad \text{as } |(\Delta t, \Delta x)|\to 0.$$

From the continuity of f and the fact that all solutions of (13.5) lie in a compact set, independent of ζ_t, it follows that there is a constant K, independent of ζ_t and t such that $|\Delta x|\leqq K|\Delta t|$. Hence $\eta_1(t)/\Delta t\to 0$ as $\Delta t\to 0$. We also have

$$\Delta x = \int_{t_1}^t \left[\int_Z f(s, \psi(s), \bar{y}, z)\, d\zeta_t(s)\right] ds = \int_{t_1}^t \left[\int_Z \{f(t_1, x_1, \bar{y}, z)+\eta_2(s)\}\, d\zeta_t(s)\right] ds,$$

where $\eta_2(s)\to 0$ as $s\to t_1$, uniformly with respect to z in Z. Since for each s, $\zeta_t(s)$ is a probability measure on Z, we get that

$$\Delta x = \int_{t_1}^t \left[\int_Z f(t_1, x_1, \bar{y}, z)\, d\zeta_t(s)\right] ds+\eta_3(t)\Delta t,$$

where $\eta_3(t) \to 0$ as $\Delta t \to 0$. Substituting the last expression for Δx into (13.6) gives

(13.7)
$$\Delta W^-(t) = \int_{t_1}^{t} \int_Z [W_t^-(t_1, x_1) + \langle W_x^-(t_1, x_1), f(t_1, x_1, \bar{y}, z)\rangle]\, d\zeta_t(s)\, ds + \eta(t)\Delta t,$$

where $\eta(t) \to 0$ as $\Delta t \to 0$.

By (13.4), the term in square brackets in (13.7) equals a constant $c > 0$. Hence we get

$$\Delta W^-(t) = [c + \eta(t)]\Delta t \quad \text{for all } t_1 \leqq t < T.$$

This implies that for t sufficiently close to t_1, $\Delta W^-(t) > 0$, which contradicts (13.6), where we have $\Delta W^-(t) \leqq 0$ for all $t_1 \leqq t < T$. This contradiction proves the lemma.

If in addition to Assumption I′ we assume that the Isaacs condition (9.1) holds, then by Theorem 11.1, $W^+ = W^-$. From this and from Lemma 13.1 we get that the value function W satisfies the Isaacs equation almost everywhere. From Theorem 12.3 we also get that the Fleming value F, the Friedman value V and the Elliott-Kalton value U satisfy the Isaacs equation.

The preceding discussion has established the following result, among others.

THEOREM 13.1. *Let Assumption* I′ *hold and let the Isaacs condition* (9.1) *hold. At points of differentiability* W *satisfies*

$$\langle W_x(t, x), f(t, x, y^*, z^*)\rangle = \max_y \min_z \langle W_x(t, x), f(t, x, y, z)\rangle$$

(13.8)
$$= \min_z \max_y \langle W_x(t, x), f(t, x, y, z)\rangle$$
$$= -W_t(t, x),$$

where the max *is taken over* $y \in Y$ *and the* min *is taken over* $z \in Z$. *The function* W *also satisfies the boundary condition* $W(T, x) = g(x)$.

REFERENCES

[1] L. D. BERKOVITZ, *Necessary conditions for optimal strategies in a class of differential games and control problems*, SIAM J. Control, 5 (1967), pp. 1-24.

[2] R. J. ELLIOTT AND N. J. KALTON, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126 (1972).

[3] W. H. FLEMING, *The convergence problem for differential games*, J. Math. Anal. Appl., 3 (1961), pp. 102-116.

[4] ———, *The convergence problem for differential games* II, Annals of Mathematics Study 52, Princeton Univ. Press, Princeton, NJ, 1964, pp. 195-210.

[5] A. FRIEDMAN, *Differential Games*, John Wiley, New York, London, Sydney, 1971.

[6] ———, *Differential Games*, CBMS Regional Conference Series in Applied Mathematics 18, American Mathematical Society, Providence, RI, 1974.

[7] R. ISAACS, *Differential games* I, II, III, IV, RAND Corporation, Research Memoranda RM-1391, RM-1399, RM-1411, RM-1486 (1954-1955).

[8] ———, *Differential Games*, John Wiley, New York, London, Sydney, 1965.

[9] N. N. KRASOVSKII AND A. I. SUBBOTIN, *Positional Differential Games*, Nauka, Moscow, 1974. (In Russian.)

[10] A. N. KRASOVSKII, *A nonlinear differential game with value given by an integral*, Differencial'nye Uravnenija, 18, 8 (1982), pp. 1306-1312; Differential Equations 18, 8 (1982), no. 8, pp. 911-916.

[11] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

NONLINEAR OBSERVERS WITH LINEARIZABLE ERROR DYNAMICS*

ARTHUR J. KRENER† AND WITOLD RESPONDEK‡

Abstract. We present a new method for designing asymptotic observers for a class of nonlinear systems. The error between the state of the system and the state of the observer in appropriate coordinates evolves linearly and can be made to decay arbitrarily exponentially fast.

Key words. nonlinear estimation, nonlinear observer, observable and observer form

1. Introduction. The problem of approximating the state $x \in \mathbb{R}^n$ of a linear system

$$\dot{x} = Ax + Bu, \tag{1.1a}$$

$$y = Cx \tag{1.1b}$$

based on knowledge of the input $u \in \mathbb{R}^n$ and output $y \in \mathbb{R}^p$ has a well-known solution provided only that (C, A) be an observable pair, i.e.

$$\operatorname{rank} \begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix} = n. \tag{1.2}$$

We define $z(t)$, an estimate of $x(t)$, to evolve according to the dynamics

$$\dot{z} = (A + GC)z - Gy + Bu \tag{1.3}$$

where G is an $n \times p$ matrix to be chosen. Then the error $e = x - z$ satisfies

$$\dot{e} = (A + GC)e. \tag{1.4}$$

The observability hypothesis (1.2) ensures that for any set of n complex numbers invariant under complex conjugation there exists a G so that the spectrum of $(A + GC)$ is that set. In particular G can be chosen so that the spectrum is sufficiently to the left in the complex plane so that error decays arbitrarily exponentially fast. See [6] for details.

In this paper we identify a class of nonlinear systems of the form

$$\dot{\xi} = f(\xi, u), \tag{1.5a}$$

$$\psi = h(\xi) \tag{1.5b}$$

for which there exists observers with arbitrary exponential error decay at least locally. We give necessary and sufficient conditions in the form of a constructive algorithm for there to exist changes of coordinates

$$\xi = \xi(x), \tag{1.6a}$$

$$\psi = \psi(y) \tag{1.6b}$$

* Received by the editors November 8, 1983, and in revised form April 11, 1984. This research was supported in part by the National Science Foundation under grant MCS-8300884, by the National Aeronautics and Space Administration under grant N CA2-OY-18 0-251 and by the Department of Energy under grant DE-AC01-80RA 50421.

† Department of Mathematics, University of California, Davis, California 95616.

‡ Institute of Mathematics, Technical University of Warsaw, Pl. Jednosci Robotniczej 1, 00-661 Warsaw, Poland.

in a suitable domain transforming (1.5) into

$$\dot{x} = Ax + \gamma(y, u), \tag{1.7a}$$

$$y = Cx \tag{1.7b}$$

where (C, A) is an observable pair. If (C, A) is in dual Brunovsky canonical form we say that (1.7) is in observer form. A slight modification of (1.3) yields an observer for (1.6)

$$\dot{z} = (A + GC)z - Gy + \gamma(y, u) \tag{1.8}$$

with the same error dynamics (1.4) as before.

If we transform (1.8) back by (1.6), we obtain a differential equation for $\zeta(t) = \xi(z(t))$

$$\dot{\zeta} = \hat{f}(\zeta, \psi, u). \tag{1.9}$$

On a compact subdomain one can achieve arbitrary exponential decay of the error between $\xi(t)$ and $\zeta(t)$ by proper choice of G.

This paper grew out of earlier work of Krener-Isidori [1] who considered the above question when $p = 1$, $\psi = y$ and with no inputs. Essentially we shall reduce the more general question to the multi-output ($p \geqq 1$) version of that. In some loose sense the question which we address is the mathematical dual of that solved by Jakubcyzk-Respondek [2] and Hunt-Su [3]. They considered the problem of linearization of (1.5a) using change of coordinates in the state space and state dependent change of coordinates in the input space (nonlinear state feedback). We refer the reader to [1]-[3] for a fuller discussion of these points.

A referee called to our attention similar work of Bestle and Zeitz [7]. They assumed the existence of the linearizing transformations and showed how the observer could be constructed when $p = 1$.

The paper is organized as follows. Section 2 discusses the observability of a nonlinear system and § 3 develops a key necessary condition (Proposition 3.3), for the existence of an observer form. Section 4 is the heart of the paper, in which the two theorems which reduce the general problem to the multivariable version of [1] are presented. In § 5 the multivariable version of [1] is given. Sections 2-5 consider systems without inputs, while § 6 generalizes to systems with inputs. We close by a series of examples in § 7. The reader may wish to consult these immediately after reading the statements of Theorems 4.1, 4.2 and 5.1 and the associated remarks.

2. Observability. Consider the problem of estimating the current state $\xi(t)$ of the nonlinear system without inputs

$$\dot{\xi} = f(\xi), \qquad \xi \in \mathbb{R}^n, \tag{2.1a}$$

$$\psi = h(\xi), \qquad \psi \in \mathbb{R}^p, \tag{2.1b}$$

$$\xi^0 \approx \xi(0) \tag{2.1c}$$

from knowledge of the past outputs $\psi(s)$, $0 \leqq s \leqq t$, but with no knowledge of the initial state $\xi(0)$ except that it is near ξ^0. Later in § 6 we shall treat systems with inputs. We are not using the term "estimation" in a statistical sense although one could make additional assumptions about (2.1) and formulate the problem as such. Rather we desire that our estimate $\hat{\xi}(t)$ converge to $\xi(t)$ "sufficiently fast" as t increases. The initial displacement $\xi(0) - \xi^0$ represents an error in our current estimate of the state due to the accumulation of past disturbances. The estimate should converge fast enough

so that the error becomes negligible in a short length of time and future disturbances are dampened at a rate faster than they arrive. On the other hand if one attempts too high a rate of decay, the inaccuracies of the observations (2.1b) can play havoc with the estimate.

The mathematical extreme of this approach is to estimate $\xi(t)$ by differentiating the output $\psi(t)$ several times. While this is not a practical approach, it does set a limit on the observability inherent in the model (2.1). Of course this requires that (2.1) be sufficiently differentiable or $\mathscr{C}^\infty$, we shall implicitly assume this throughout this paper. If from the knowledge of $\psi(t)$ and its derivatives at t one can uniquely determine $\xi(t)$, then (2.1) is observable.

To make this mathematically precise we must introduce some terminology. Let $\psi_i^{(j)}(t)$ denote the jth time derivative of the ith output. This can be expressed using Lie differentiation of the functions h_i by the vector field f,

$$\psi_i^{(j)}(t) = L_f^j(h_i)(\xi(t)). \tag{2.2}$$

$L_f^j(h_i)(\xi)$ is the jth Lie derivative of h_i by f and a function of ξ defined inductively by

$$L_f^0(h_i)(\xi) = h_i(\xi), \tag{2.3a}$$

$$L_f^j(h_i)(\xi) = \frac{\partial}{\partial \xi}(L_f^{j-1}(h_i)(\xi))f(\xi). \tag{2.3b}$$

The symbol $(\partial/\partial\xi)(h_i)$ stands for the gradient of the function h_i and is a $1 \times n$ vector valued function of ξ. It is the local coordinate description of the one form dh_i, which can also be Lie differentiated by f. For our purposes the following suffices as a definition

$$L_f(dh_i) = d(L_f(h_i)). \tag{2.4}$$

DEFINITION 2.1. This system (2.1) is *observable* at ξ^0 if there exists a neighborhood $\mathscr{U}$ of ξ^0 and p-tuple of integers $(k_1, \cdots, k_p)$ such that

(i) $k_1 \geqq k_2 \geqq \cdots \geqq k_p \geqq 0$ and $\sum_{i=1}^p k_i = n$.

(ii) After suitable reordering of the h_i's, at each $\xi \in \mathscr{U}$ the n row vectors $\{L_f^{j-1}(dh_i)\colon i = 1, \cdots, p;\ j = 1, \cdots, k_i\}$ are linearly independent.

(iii) If $(l_1, \cdots, l_p)$ satisfies (i) and after suitable reordering the n row vectors $\{L_f^{j-1}(dh_i)(\xi)\colon i = 1, \cdots, p;\ j = 1, \cdots, k_i\}$ are linearly independent at some $\xi \in \mathscr{U}$ then $(l_1, \cdots, l_p) \geqq (k_1, \cdots, k_p)$ in the lexographic ordering [$(l_1 > k_1)$ or $(l_1 = k_1$ and $l_2 > k_2)$ or $(l_1 = k_1, l_2 = k_2$ and $l_3 > k_3)$ or $\cdots$ or $(l_1 = k_1, \cdots, l_p = k_p)$].

The integers $(k_1, \cdots, k_p)$ are called the observability indices at ξ^0.

This definition of observability is not the only one which has appeared in the literature. See [4] and [5] for alternatives. It is equivalent to being able to take the n functions $\{L_f^{j-1}(h_i)\colon i = 1, \cdots, p;\ j = 1, \cdots, k_i\}$ as coordinates in a neighborhood of ξ^0 where no set of lower derivatives would suffice. If we abuse notation by letting $\xi_{ij} = L_f^j(h_i)(\xi)$, then (2.1) becomes

$$\begin{array}{llll} \psi_1 = \xi_{11} & & \cdots & \psi_p = \xi_{p1} \\ \dot{\xi}_{11} = \xi_{12} & & & \dot{\xi}_{p1} = \xi_{p2} \\ \dot{\xi}_{12} = \xi_{13} & & & \vdots \\ \vdots & & & \dot{\xi}_{pk_p} = f_{p(\xi)} \\ \dot{\xi}_{1k_1} = f_{1(\xi)} & & \cdots & \end{array} \tag{2.5}$$

where $f_i = L_f^{k_i}(h_i)$.

Following Kailath [6] we refer to (2.5) as a system in *observable form.* It is not a canonical form relative to the pseudo-group of state and output coordinate changes because different output coordinates (or even different ordering of the outputs) lead to different f_i's.

LEMMA 2.2. *The system* (2.1) *admits an observable form around* ξ^0 *iff* (2.1) *is observable at* ξ^0. *The observability indices at* ξ^0 *are the same as the* k_i*'s of any observable form* (2.5) *at* ξ^0.

Let us consider how one might verify Definition 2.1 and Lemma 2.2 for a system (2.1). Define $\mathscr{E}^0 = \{0\}$ and

$$\mathscr{E}^k = \text{Span}\,\{L_f^{j-1}(dh_i)\colon i = 1, \cdots, p; j = 1, \cdots, k\}$$

where Span indicates all linear combinations over the $\mathscr{C}^\infty$ functions of ξ. Each $\mathscr{E}^k$ is a module of one forms over this ring of functions; such an object is called a *codistribution* or *Pfaffian system.* Let $E^k(\xi)$ denote the space of cotangent vectors obtained by evaluating the one forms of $\mathscr{E}^k$ at ξ. Each $E^k(\xi)$ can be thought of as a space of $1 \times n$ vectors. Clearly $\mathscr{E}^{k-1} \subset \mathscr{E}^k$ and $E^{k-1}(\xi) \subset E^k(\xi)$. Let $d_k(\xi)$ denote the codimension of $E^{k-1}(\xi)$ in $E^k(\xi)$.

LEMMA 2.3. *The system* (2.1) *is observable at* ξ^0 *with observability indices* $(k_1, \cdots, k_p)$ *iff* $d_i(\xi)$ *is constant in a neighborhood of* ξ^0 *for* $i = 1, \cdots, n$ *and* $d_n(\xi) = n$. *The relation between these sets of integers is given by*

$$d_k = \text{card}\,\{k_i\colon k_i \geqq k\}, \tag{2.6a}$$

$$k_i = \max\,\{k\colon d_k \geqq i\}. \tag{2.6b}$$

The proof amounts to an algorithm for transforming (2.1) to (2.5). It uses the fact that $\mathscr{E}^k$ is invariant under change of output and state coordinates.

Proof. Suppose (2.1) is observable with indices $(k_1, \cdots, k_p)$; then $E^k(\xi)$ has as a basis $\{L_f^j(dh_i)\colon i = 1, \cdots, p; j = 1, \cdots, \min(k_i, k)\}$ hence is of constant dimension.

On the other hand suppose $d_k(\xi)$ is constant for each k. After reordering the outputs we can assume that the first d_1 of the dh_i's are a basis for $E^1(\xi)$. We can reorder the first d_1 of the outputs so that $L_f(dh_i)$ $i = 1, \cdots, d_2$ and dh_i, $i = 1, \cdots, d_1$ are a basis for $E^2(\xi)$. We repeat the processes reordering the first d_2 of the outputs so that $L_f^2(dh_i)\colon i = 1, \cdots, d_3$ and the previous chosen basis for $E^2(\xi)$ forms a basis for $E^3(\xi)$. In this way we obtain n linearly independent exact one forms. The corresponding functions are the desired coordinates ξ_{ij}. Q.E.D.

3. Necessary conditions. While observable form is useful for deciding the observability of a system, it is not particularly helpful in constructing an observer. Suppose there exist changes of coordinates $x = x(\xi)$ and $y = y(\psi)$ around ξ^0 and $\psi^0 = h(\xi^0)$ which transform (2.1) into *observer form*

$$\begin{aligned}
&y_1 = x_1 && \cdots && y_p = x_{p1} \\
&\dot{x}_{11} = x_{12} + \alpha_{11}(y) && && \dot{x}_{p1} = x_{p2} + \alpha_{p1}(y) \\
&\dot{x}_{12} = x_{13} + \alpha_{12}(y) && && \vdots \\
&\quad\vdots && && \dot{x}_{pk_p} = \alpha_{pk_p}(y) \\
&\dot{x}_{1k_1} = \alpha_{1k_1}(y) && \cdots &&
\end{aligned} \tag{3.1}$$

The construction of an observer for (3.1) is straightforward. Let z_{ij} evolve according to

$$\begin{aligned}
&\dot{z}_{11} = z_{12} + \alpha_{11}(y) + q_{11}(y_1 - z_{11}) && \cdots && \dot{z}_{p1} = z_{p2} + \alpha_{p1}(y) + q_{p1}(y - z_{p1}) \\
&\dot{z}_{12} = z_{13} + \alpha_{12}(y) + q_{12}(y_1 - z_{11}) && && \vdots \\
&\quad\vdots && && \dot{z}_{pk_p} = \alpha_{pk_p}(y) + q_{pk_p}(y_p - z_{p1}) \\
&\dot{z}_{1k_1} = \alpha_{1k_1}(y) + q_{1k_1}(y_1 - z_{11}) && \cdots &&
\end{aligned} \tag{3.2}$$

where q_{ij} are constants to be chosen.

If $e_{ij} = x_{ij} - z_{ij}$, then

$$\begin{aligned}
&\dot{e}_{11} = e_{12} - q_{11}e_{11} && \cdots && \dot{e}_{p1} = p_{p2} - q_{p1}e_{p1} \\
&\dot{e}_{12} = e_{13} - q_{12}e_{11} && && \vdots \\
&\quad\vdots && && \dot{e}_{pk_f} = -q_{pk_p}e_{p1} \\
&\dot{e}_{1k_1} = -q_{1k_1}e_{11} && \cdots &&
\end{aligned} \tag{3.3}$$

The characteristic polynomial of this linear system is

$$p(\lambda) = \prod_{i=1}^{p} \left(\sum_{j=0}^{k_i} q_{ij}\lambda^{k_i - j} \right) \tag{3.4}$$

where $q_{i0} = 1$. Clearly we can set the spectrum arbitrarily so that the error decays exponentially fast at any desired rate.

It is well known that every observable linear system can be transformed into observer form where α_{ij} is linear in y by linear coordinate changes in the state and output [6]. However, even if we allow nonlinear coordinate changes and nonlinear α_{ij}, the analogous result for nonlinear systems does not hold.

PROPOSITION 3.1. *If the system* (2.1) *admits an observer form* (3.1) *at* ξ^0, *it must be observable at* ξ^0 *with observability indices given by the* k_i*'s of* (3.1).

Proof. Let $\Psi = (\partial\psi_i/\partial y_j)$; then

$$d\psi = \begin{pmatrix} d\psi_1 \\ \vdots \\ d\psi_p \end{pmatrix} = \Psi\, dy = \Psi \begin{pmatrix} dy_1 \\ dy_p \end{pmatrix},$$

so

$$\mathscr{E}^1 = \text{Span}\{dy_1, \cdots, dy_p\} = \text{Span}\{dx_{11}, \cdots, dx_{p1}\}.$$

Assume by induction that

$$L_f^{j-2}(d\psi) \equiv \Psi L_f^{j-2}(dy) \qquad \text{mod } \mathscr{E}^{j-2};$$

then

$$\begin{aligned}
&L_f^{j-1}(d\psi) \equiv \Psi L_f^{j-1}(dy) + L_f(\Psi) L_f^{j-2}(dy) && \text{mod } L_f(\mathscr{E}^{j-2}) = \mathscr{E}^{j-1}, \\
&L_f^{j-1}(d\psi) \equiv \Psi L_f^{j-1}(dy) && \text{mod } \mathscr{E}^{j-1}.
\end{aligned}$$

But

$$L_f^{j-1}(dy_i) \equiv \begin{cases} dx_{ij+1} & \text{if } j+1 \leqq k_i, \\ 0 & \text{if } j+1 > k_i, \end{cases} \quad \text{mod } \mathscr{E}^{j-1}$$

so $\mathscr{E}^j$ is spanned by dx_{ij}, $j \leqq k_i$, mod $\mathscr{E}^{j-1}$. From this we see that the dimensions of $E^j(\xi)$ are constant. Q.E.D.

DEFINITION 3.2. Suppose the nonlinear system (2.1) is in observable form (2.5). We denote $\mathcal{P}(\xi)$ the ring of polynomials in ξ with coefficients that are $\mathcal{C}^\infty$ function of ψ. The degree of ξ_{ij} is defined to be $j-1$ and the degree of the monomial $\xi_{i_1 j_1} \cdot \cdots \cdot \xi_{i_n j_n}$ is the sum of the degrees of its factors, $(j_1-1)+\cdots+(j_r-1)$. $\mathcal{P}^k(\xi)$ denotes the polynomials of degree k or less and $\mathcal{P}_0^k(\xi)$ those polynomials of $\mathcal{P}^k(\xi)$ which are generated by elements of $\mathcal{P}^{k-1}(\xi)$. In particular, ξ_{ik+1} is in $\mathcal{P}^k(\xi)$ but not in $\mathcal{P}_0^k(\xi)$.

With this terminology available we can introduce our second necessary condition.

PROPOSITION 3.3. *Suppose the system* (2.1) *admits an observer form at* ξ^0; *then in any observable form* (2.5) *the functions* $f_i(\xi)$ *are in* $\mathcal{P}^{k_i}(\xi)$ *for* $i=1,\cdots,p$.

Proof. Let $\mathcal{P}(x)$ denote the polynomials in x with coefficients that are $\mathcal{C}^\infty$ functions of y; in a similar fashion we define the degree of x_{ij} to be $j-1$. $\mathcal{P}^k(x)$ are the polynomials of degree $\leqq k$ and $\mathcal{P}_0^k(x)$ the subset of $\mathcal{P}^k(x)$ generated by elements of $\mathcal{P}^{k-1}(x)$.

It is easy to see that $L_f(\mathcal{P}^{k-1}(x)) \subset \mathcal{P}^k(x)$ and $L_f(\mathcal{P}_0^{k-1}(x)) \subset \mathcal{P}_0^k(x)$. For example x_{ik}, $k \leqq k_i$, is of degree $k-1$ and

$$L_f(x_{ik}) = \dot{x}_{ik} = \begin{cases} x_{ik+1} + \alpha_{ik}(y), & k < k_i, \\ \alpha_{ik}(y), & k = k_i \end{cases}$$

is clearly of degree at most k. A similar calculation using the Leibniz rule shows that monomials of degree $k-1$ go into monomials of degree k under Lie differentiation by f.

Notice that the changes of coordinates transform $\mathcal{P}^0(\xi)$ onto $\mathcal{P}^0(x)$, i.e., $\mathcal{P}^0(\xi(x)) = \mathcal{P}^0(x)$ and $\mathcal{P}^0(\xi) = \mathcal{P}^0(\xi(x))$. Moreover $\mathcal{P}_0^1(\xi) = \mathcal{P}^0(\xi)$ and $\mathcal{P}_0^1(x) = \mathcal{P}^0(x)$ so $\mathcal{P}_0^1(\xi)$ is transformed into $\mathcal{P}_0^1(x)$.

We show induction that $\mathcal{P}^k(\xi)$ is transformed to $\mathcal{P}^k(x)$ and $\mathcal{P}_0^{k+1}(\xi)$ is transformed to $\mathcal{P}_0^{k+1}(x)$ for all k. If $k_i \geqq 2$, then

$$\xi_{i2} = \dot{\xi}_{i1} = \sum_{j=1}^{p} \frac{\partial \psi_i}{\partial y_j} \dot{x}_{j1},$$

$$\xi_{i2} = \sum_{k_j \geqq 2} \frac{\partial \psi_i}{\partial y_j} x_{j2} + p_{i2}(x) \tag{3.5}$$

where

$$p_{i2}(x) = \sum_{j=1}^{p} \frac{\partial \psi_i}{\partial y_j} \alpha_{j2}(y) \in \mathcal{P}_0^1(x) = \mathcal{P}^0(x).$$

This proves the above statement for $k=1$.

Suppose it is true for $k-1$ and suppose also that the generalization of (3.5) holds, i.e., if $k_i \geqq k$

$$\xi_{ik} = \sum_{k_j \geqq k} \frac{\partial \psi_i}{\partial y_j} x_{jk} + p_{ik}(x), \tag{3.6}$$

where $p_{ik}(x) \in \mathcal{P}_0^{k-1}(x)$. If $k_i \geqq k+1$, then

$$\xi_{ik+1} = L_f(\xi_{ik}) = \xi_{ik} = \sum_{k_j \geqq k+1} \frac{\partial \psi_i}{\partial y_j} x_{jk+1} + p_{ik+1}(x) \tag{3.7}$$

where

$$p_{ik+1}(x) = \sum_{k_j \geqq k} \left(\frac{\partial \psi_i}{\partial y_j} \alpha_{jk+1} + L_f\left(\frac{\partial \psi_i}{\partial y_j} \right) x_{jh} \right) + L_f(p_{ik}(x)), \tag{3.8}$$

and hence $p_{ik+1}(x) \in \mathcal{P}_0^k(x)$. From (3.7) the statement follows for $k+1$. Q.E.D.

Actually we can deduce a slightly stronger result from the above argument.

PROPOSITION 3.4. *If a system* (2.1) *admits an observer form* (3.1) *around* ξ^0, *then it admits an observable form* (2.5) *around* ξ^0 *which satisfies* $f_i(\xi) \in \mathscr{P}_0^{k_i}(\xi)$, $i = 1, \cdots, p$.

Proof. Suppose $y = \psi$. Differentiating (3.6) with $k = k_i$ yields

$$f_i(\xi) = \dot{\xi}_{ik_i} = L_f(\xi_{ik_i}) = \sum_{k_j \geqq k_i} \frac{\partial \psi_i}{\partial y_j} \dot{x}_{jk_i} + p_{ik+1}(x) \tag{3.9}$$

where $p_{ik+1}(x) \in \mathscr{P}_0^{k_i}(x)$ is given by (3.8). Since $\psi_i = y_i$ and $x_{ik_i} = \alpha_{ik_i} \in \mathscr{P}^0(x)$, the result follows. Q.E.D.

DEFINITION 3.5. A system (2.5) is in *special observable form* if $f_i(\xi) \in \mathscr{P}_0^{k_i}(\xi)$, $i = 1, \cdots, p$.

Of course a system need not have a special observable form and such forms are not always unique. As we shall see in the next section, they are a very useful intermediate step between the observable and observer forms. We will also give necessary and sufficient conditions for the existence of a special observable form. Notice that if $k_1 = \cdots = k_p$ (e.g. $p = 1$), then any observable form satisfying Proposition 3.3 is special. This is because $\mathscr{P}_0^{k_i}(\xi) = \mathscr{P}^{k_i}(\xi)$.

4. Change of output coordinates and prolongation. Consider a system satisfying the two necessary conditions of Propositions 3.1 and 3.3, namely, that it can be transformed to observable form (2.5) where $f_i(\xi) \in \mathscr{P}^{k_i}(\xi)$. Suppose we take the obvious approach and compare (2.5) and (3.1) to obtain differential equations for $\psi(y)$ and $\alpha(y) = (\alpha_{ij}(y))$. For simplicity assume $p = 1$ and hence $k_1 = n$. This approach yields

$$\begin{aligned}
&\xi_1 = \psi(x_1),\\
&\xi_2 = \dot{\xi}_1 = \frac{d\psi}{dx_1}(x_2 + \alpha_1),\\
&\xi_3 = \dot{\xi}_2 = \frac{d\psi}{dx_1}\left(x_3 + \alpha_2 + \frac{d\alpha_1}{dx_1}(x_2 + \alpha_1)\right) + \frac{d^2\psi}{dx_1}(x_2 + \alpha_1),\\
&\quad\vdots\\
&f_n(\xi) = \dot{\xi}_n = \frac{d\chi}{dx_1}(\alpha_n + \cdots) + \cdots + \frac{d^n\psi}{dx_1^n}(x_2 + \alpha_1)^n.
\end{aligned} \tag{4.1}$$

The result is an nth order system of nonlinear ordinary differential equations for the $1 + n$ unknowns $\psi, \alpha_1, \cdots, \alpha_n$. If $p > 1$ the situation is even worse, for we obtain a k_1th order system of nonlinear partial differential equations for the $p + n$ unknown ψ_i and α_{ij}. Clearly a better approach is needed for all but the smallest value of p and n.

Our approach will be to separate the problem into two parts. The first step is to derive a first order linear differential equation which essentially determines the change of output coordinates $\psi(y)$ if it exists. Once we have this, then we can use the method of Krener and Isidori [1] to decide if the system can be transformed into observer form and to compute the change of coordinates and $\alpha(y)$. This latter task we postpone to § 5. The rest of this section will be devoted to proving the following.

THEOREM 4.1. *Consider a system in special observable form. If it can be transformed to observer form, then the Jacobian* $\Psi = (\Psi_i^j) = (\partial \psi_i / \partial y_j)$ *of the change of output coordinates must satisfy*

$$\Psi_i^j = 0 \quad \text{if } k_j > k_i \tag{4.2a}$$

and

$$\frac{\partial}{\partial \psi_l} \Psi_i^j = \frac{1}{k_i} \sum_{r=1}^{p} f_{i;lk_i;r2} \Psi_r^j. \tag{4.2b}$$

We can normalize $\Psi(\psi)$ by specifying that

$$\Psi(\psi^0) = I. \tag{4.2c}$$

A different initial condition amounts to a linear change of state and output coordinates.

Notation. We are using the semicolon to denote partial differentiation, e.g.,

$$f_{i;sk_i} = \frac{\partial f_i}{\partial \xi_{sk_i}}, \qquad f_{i;sk_i;l2} = \frac{\partial^2 f_i}{\partial \xi_{l2}\, \partial \xi_{sk_i}}.$$

Remark 4.1. Note that $f_{i;sk_i;l2} \in \mathcal{P}^0(\xi)$ since the system is assumed to be in special observable form. Therefore the differential equation (4.2b) lives not on the state space but on the output space as it should.

Remark 4.2. The equations (4.2b) are locally solvable iff the mixed partial conditions are satisfied

$$\frac{\partial}{\partial \psi_n}\frac{\partial}{\partial \psi_l}(\Psi_i^j) = \frac{\partial}{\partial \psi_l}\frac{\partial}{\partial \psi_n}(\Psi_i^j).$$

Since Ψ must be invertible, this reduces to

$$f_{i;lk_i;r2;m1} + \sum_{s=1}^{p} \frac{1}{k_2} f_{i;lk_i;s2} f_{s;mk_s;r2} = f_{i;mk_i;r2;l1} + \sum_{s=1}^{p} \frac{1}{k_s} f_{i;mk_i;s2} f_{s;lk_s;r2} \tag{4.3}$$

for $i, m, l, r = 1, \cdots, p$.

Moreover, a solution of (4.2b) need not automatically satisfy (4.2a). This imposes additional necessary conditions on the f_i's, namely that their partials have the same block upper triangular structure as Ψ, i.e.

$$f_{i;lk_i;j2} = 0 \quad \text{if } k_j > k_i. \tag{4.4}$$

Remark 4.3. Suppose Ψ is a solution to (4.2); then $\psi(y)$ satisfies

$$\frac{\partial \psi_i}{\partial y_j} = \Psi_i^j. \tag{4.5}$$

Equation (4.5) is integrable iff the mixed partials commute. This is equivalent to

$$f_{i;sk_i;r2} = f_{i;rk_i;s2}. \tag{4.6a}$$

It is useful to choose the solution so that ψ^0 transforms to $y^0 = 0$, i.e.,

$$\psi(0) = \psi^0 = h(\xi^0). \tag{4.6b}$$

Remark 4.4. Proposition 4.1 deals with a system in special observable form. Of course any system which can be transformed into observer form must satisfy $f_i(\xi) \in \mathcal{P}^{k_i}(\xi)$ in any observable form. To bring it to special observable form requires a change of output coordinates $\tilde{\psi} = \tilde{\psi}(\psi)$ as described below.

Let $Y^1(\psi), \cdots, Y^p(\psi)$ be vector fields defined on the output space whose coordinate descriptions relative to ψ are given by

$$L_Y j(\psi_i) = \begin{cases} 1 & \text{if } i = j, \\ f_{i;jk_i+1} & \text{if } k_j > k_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.7}$$

Notice that $f_{i;jk_i+1}$ is a function of ψ alone since $f_i(\xi) \in \mathcal{P}^{k_i}(\xi)$.

THEOREM 4.2. *Consider a system is observable form satisfying* $f_i(\xi) \in \mathcal{P}^{k_i}(\xi)$. *It is in special observable form relative to the transformed output coordinates* $\tilde{\psi} = \tilde{\psi}(\psi)$ *iff*

$$L_Y j(\tilde{\psi}_i) = 0 \quad \text{for } k_j > k_i. \tag{4.8}$$

There exists such a change of coordinates $\tilde{\psi}(\psi)$ *iff the distributions*

$$\mathcal{U}^i = \text{Span}\{Y^j(\psi): k_j > k_i\}$$

are involutive for $i = 1, \cdots, p$.

Recall that a *distribution* is a family of vector fields closed under addition and multiplication by $\mathcal{C}^\infty$ functions. It is *involutive* if the Lie bracket of any two vector fields from the distribution is again in the distribution. The Lie bracket is defined in local coordinates by

$$[Y^i, Y^j] = \frac{\partial Y^j}{\partial \psi} Y^i - \frac{\partial Y^i}{\partial \psi} Y^j;$$

note that it is again a vector field on the same space, in this case the output space.

Proof of Theorem 4.1. We start with a system in special observable form. For the time being assume all of the observability indices are the same, $k_1 = k_2 = \cdots = k_p = k$, so that $n = p \cdot k$. Let $g^j(\xi)$ be the vector field on the state space which is the unit vector in the ξ_{jk}-direction, for $j = 1, \cdots, p$. Equivalently these p vector fields are characterized by the equations

$$L_{g^j} L_f^l(\psi_i) = \begin{cases} 0, & 0 \leqq l < k-1, \\ \delta_i^j, & l = k-1, \end{cases} \tag{4.9a}$$

where δ_i^j is the Kronecker δ symbol.

We introduce the *ad*-notation for repeated Lie brackets

$$ad^0(-f)g^j = g^j, \qquad ad^{l+1}(-f)g^j = [-f, ad^l(-f)g^j]$$

and the pairing of a one-form $\omega(\xi)$ and vector field $X(\xi)$

$$\langle \omega, X \rangle(\xi) = \omega(\xi) X(\xi).$$

In local coordinates the right side is the product of $1 \times n$ and $n \times 1$ vector valued functions of ξ. The Leibniz formula holds for this pairing under Lie differentiation

$$L_f \langle \omega, X \rangle = \langle L_f(\omega), X \rangle + \langle \omega, [f, X] \rangle.$$

Therefore (4.9a) is equivalent to

$$\begin{aligned} \langle L_f^{l-r}(d\psi_i), ad^r(-f)g^j \rangle &= L_{ad^r(-f)g^j} L_f^{l-r}(d\psi_i) \\ &= \begin{cases} 0, & 0 \leqq r \leqq l < k-1, \\ \delta_i^j, & 0 \leqq r \leqq l = k-1. \end{cases} \end{aligned} \tag{4.9b}$$

Suppose the system can be transformed into observer form by change of state and output coordinates. Let B^j be the vector field on the state space which is the unit vector in the x_{jk} direction and $\bar{g}^j(\xi)$ be the representation of this vector field in ξ coordinates. Let $\bar{f}(\xi)$ and $\bar{\alpha}(\xi)$ be the representations in ξ coordinates of the vector fields represented in x coordinates by Ax and $\alpha(y)$ where Ax is the linear part of the right side of the

differential equation of (3.1). In other words if x is the x_{ij}'s in lexographic ordering, then

(4.10a)
$$A=\left(\begin{array}{ccc|c|ccc} 0 & 1 & & & & & \\ & \ddots & 1 & 0 & & 0 & \\ & & 0 & & & & \\ \hline & 0 & & \ddots & & 0 & \\ \hline & & & & 0 & 1 & \\ & 0 & & 0 & & \ddots & 1 \\ & & & & & & 0 \end{array}\right).$$

The output is given by $y=Cx$ and $B=(B^1,\cdots,B^p)$ where

(4.10b)
$$B=\left(\begin{array}{c|c|c} 0 & & \\ \vdots & 0 & 0 \\ 0 & & \\ 1 & & \\ \hline 0 & \ddots & \\ \hline & & 0 \\ 0 & 0 & \vdots \\ & & 0 \\ & & 1 \end{array}\right).$$

(4.10c)
$$C=\left(\begin{array}{ccc|c|ccc} 1 & 0 & \cdots & 0 & 0 & & 0 & \\ \hline & 0 & & & \ddots & & 0 & \\ \hline & 0 & & & & 1 & 0 \cdots & 0 \end{array}\right).$$

The ith diagonal blocks of A, B and C are of dimensions $k_i\times k_i$, $k_i\times 1$ and $1\times k_i$ respectively.

Viewing $y_i=y_i(\xi)$ as a function of ξ we have

$$\begin{aligned} L_{\bar g}jL_{\bar f}^l(y_i)&=\langle L_{\bar f}^{l-r}(dy_i),\, ad^r(-\bar f)\bar g^j\rangle \\ &=L_{ad^r(-\bar f)\bar g^j}L_{\bar f}^{l-r}(dy_i) \\ &=\begin{cases} 0, & 0\leqq r\leqq l<k-1, \\ \delta_i^j, & 0\leqq r\leqq l=k-1. \end{cases} \end{aligned}$$

From the proof of Proposition 3.1 we see that both $g(\xi)=(g^1(\xi),\cdots,g^p(\xi))$ and $\bar g(\xi)=(\bar g^1(\xi),\cdots,\bar g^p(\xi))$ annihilate the codistribution $\mathscr{C}^{k-1}$ which is of codimension p. Moreover

$$L_f^{k-1}(d\psi)\equiv\Psi L_{\bar f}^{k-1}(dy)\mod \mathscr{C}^{k-1}$$

where $\Psi=(\Psi_i^j)=(\partial\psi_i/\partial y_j)$, hence $\langle L_f^{k-1}(d\psi),\bar g(\xi)\rangle=\Psi$ and so

$$\bar g=g\Psi. \tag{4.10}$$

Next we show by induction that

$$ad^l(-f)\bar g^j=ad^l(-\bar f)\bar g^j\quad\text{for } j=1,\cdots,p,\quad l=0,\cdots,k-1. \tag{4.11}$$

Suppose (4.11) holds for $l-1$; then since $f(\xi)=\bar f(\xi)+\bar\alpha(\xi)$, we have

$$\begin{aligned} ad^l(-f)\bar g^j&=-[\bar f+\bar\alpha,\, ad^{l-1}(-\bar f)\bar g^j] \\ &=ad^l(-\bar f)\bar g^j-[\bar\alpha,\, ad^{l-1}(-\bar f)\bar g^j]. \end{aligned}$$

In x coordinates $ad^{l-1}(-\bar{f})\bar{g}^j = A^{l-1}B^j$ and $-[\bar{a}, ad^{l-1}(\bar{f})\bar{g}^j]$ is $(\partial\alpha/\partial y)CA^{l-1}B = 0$. Moreover since $ad^l(-f)\bar{g}^s(\xi)$ is the constant vector field A^lB^s in x coordinates for $s = 1, \cdots, p$ and $l = 0, \cdots, k-1$, it follows that

$$[ad^j(-f)\bar{g}^r, ad^l(-f)\bar{g}^s] = 0 \tag{4.12}$$

for $r, s = 1, \cdots, p$ and $j, l = 0, \cdots, k-1$. (Note: A^l is the lth power of A, B^s the sth column of B.)

The Leibniz rule applied to (4.10) yields

$$ad^{k-1}(-f)\bar{g}^s = \sum_{l=1}^{k-1}\sum_{\sigma=1}^{s}\binom{k-1}{l} ad^l(-f)g^\sigma L_{-f}^{k-1-l}(\Psi_\sigma^s), \tag{4.13a}$$

$$ad^{k-2}(-f)\bar{g}^r = \sum_{j=1}^{k-2}\sum_{\rho=1}^{r}\binom{k-2}{j} ad^j(-f)g^\rho L_{-f}^{k-2-j}(\Psi_\rho^r). \tag{4.13b}$$

From (4.12) we see that

$$0 = \langle L_f(d\psi_i), [ad^{k-2}(-f)\bar{g}^r, ad^{k-1}(-f)\bar{g}^s]\rangle \tag{4.14}$$

and if we expand the right side using (4.13), most of the terms drop out because of (4.9). We are left with

$$\begin{aligned} 0 = \sum_{\rho,\sigma=1}^{p} \{&\langle L_f(d\psi_i), [ad^{k-2}(-f)g^\rho, ad^{k-1}(-f)g^\sigma]\rangle \Psi_\rho^r \Psi_\sigma^s \\ &+ (k-1)\langle L_f(d\psi_i), ad^{k-2}(-f)g^\sigma\rangle L_{ad^{k-2}(-f)g^\rho} L_{-f}(\Psi_\sigma^s)\Psi_\rho^r \\ &- \langle L_f(d\psi_i), ad^{k-2}(-f)g^\rho\rangle L_{ad^{k-1}(-f)g^\sigma}(\Psi_\rho^r)\Psi_\sigma^s\}. \end{aligned} \tag{4.15}$$

From (4.13) and the identity $L_f^k(d\psi_i) = f_i$ we obtain

$$\begin{aligned} \sum_{\rho,\sigma=1}^{p} f_{i;\sigma k;\rho 2}\Psi_\rho^r\Psi_\sigma^s &= \sum_{\rho,\sigma=1}^{p}\left\{(k-1)\frac{\partial}{\partial\psi_\rho}(\Psi_\sigma^s)\Psi_\rho^r + \frac{\partial}{\partial\psi_\sigma}(\Psi_\rho^r)\Psi_\sigma^s\right\} \\ &= (k-1)\sum_{\rho=1}^{p}\frac{\partial^2\psi_\rho}{\partial y_r\,\partial y_s} + \sum_{\sigma=1}^{p}\frac{\partial^2\psi_\sigma}{\partial y_s\,\partial y_r} \\ &= k\sum_{\rho,\sigma=1}^{p}\frac{\partial\Psi_\rho^r}{\partial\psi_\sigma}\Psi_\sigma^s. \end{aligned} \tag{4.16}$$

Multiplication by Ψ^{-1} yields the desired result (4.2).

Now suppose the observability indices are not all the same $k_1 \geqq \cdots \geqq k_p$. By hypothesis the system is in special observable form relative to the output ψ. Moreover by the proof of Proposition 3.4 the system will also be in special observable form relative to the output y.

Theorem 4.2 implies that the change of coordinates $y = y(\psi)$ satisfies the equations (4.8) with $y = \tilde{\psi}$, i.e.,

$$\frac{\partial\bar{y}_i}{\partial\psi_j} = 0 \quad \text{for } k_j > k_i$$

because $f_{i;jk_i+1} = 0$. This implies (4.2a).

To show (4.2b) we prolong the system, i.e., define a new system similar to the old but with all observability indices equal to the largest index k_1 of the old. We do this in such a way that the new system is transformable to observer form by ψ and α iff

the original system is also. Moreover the form of the differential equations (4.2b) for Ψ is left invariant.

In order to simplify the exposition, we will restrict to the case where there are two distinct observability indices $k_1 = k$ and $k_2 = k-1$ of multiplicities p_1 and p_2. The general case follows by repeated application of this technique. Let y_1 denote the first p_1 ouptuts and y_2 the last p_2 outputs; each ξ_{ij} is a p_i vector, etc. The original system and its transformed version are

$$\psi_i = \xi_{i1}, \qquad y_i = x_{i1},$$

$$\dot{\xi}_{ij} = \begin{cases} \xi_{ij+1}, & 1 \leqq j < k_i, \\ f_i, & j = k_i, \end{cases} \qquad \dot{x}_{ij} = \begin{cases} x_{ij+1} + \alpha_{ij}(y), & 1 \leqq j \leqq k_i, \\ \alpha_{ik_i(y)}, & j = k_i, \end{cases} \tag{4.17}$$

for $i = 1$ and 2.

The prolonged system and its transformed version are in different variables but the same function $\psi(\cdot)$ and $\alpha(\cdot)$ should accomplish the transformation,

$$\bar{\psi}_i = \bar{\xi}_{i1}, \qquad \bar{y}_i = \bar{x}_{i1},$$

$$\dot{\bar{\xi}}_{ij} = \begin{cases} \bar{\xi}_{ij+1}, & 1 \leqq j < k_1, \\ \bar{f}_i, & j = k_1, \end{cases} \qquad \dot{\bar{x}}_{ij} = \begin{cases} \bar{x}_{ij} + \alpha_{ij}(\bar{y}), & 1 \leqq j \leqq k_1, \\ \alpha_{ik}(\bar{y}), & j = k_1, \end{cases} \tag{4.18}$$

where

$$\alpha_{2k} = 0, \tag{4.19a}$$

$$\bar{f}_1 = f_1, \tag{4.19b}$$

$$\bar{f}_2 = \frac{\partial}{\partial \psi_2}(\Psi_2^2)\bar{\xi}_{22}\Psi_2^{2^{-1}}(\bar{\xi}_{2k_1} - f_2) + \sum_{i=1}^{2}\sum_{j=1}^{k_i} f_{2,ij}\bar{\xi}_{ij+1}. \tag{4.19c}$$

Of course the functions on the right side of (4.19) are to be evaluated on the new (barred) variables. Recall also that by Proposition 4.2

$$\Psi_2^1 = \frac{\partial \psi_2}{\partial y_1} = 0.$$

The claim is that (4.17) holds if (4.18) does. To see this notice that the straightforward approach (4.1) described at the beginning of this section yields almost the same set of differential equations. The only difference occurs at the k_2 and $k_1 = k_2 + 1$ time derivatives of ψ_2 and $\bar{\psi}_2$. At the k_2th derivative we have

$$f_2(\xi) = \dot{\xi}_{2k_2} = \Psi_2^2(\alpha_{2k_2} + \cdots) + \cdots, \tag{4.20a}$$

$$\bar{\xi}_{2k_1} = \dot{\bar{\xi}}_{2k_2} = \Psi_2^2(\bar{x}_{2k_1} + \alpha_{2k_2} + \cdots). \tag{4.20b}$$

Assuming that (4.17) holds and (4.18) holds up to this equation, then comparing (4.20) and the earlier equations yields

$$\bar{\xi}_{2k_1} = \Psi_2^2 \bar{x}_{2k_1} + f_2(\bar{\xi}). \tag{4.21}$$

Now (4.18) will hold if the derivative of this is consistent with

$$\dot{\bar{\xi}}_{2k_1} = \bar{f}_2(\bar{\xi}). \tag{4.22}$$

But differentiating (4.21) yields

$$\dot{\bar{\xi}}_{2k_1} = \frac{\partial}{\partial \psi_2}(\Psi_2^2)\bar{\xi}_{22}x_{2k_1} + \sum_{i=1}^{2}\sum_{2j=1}^{k_i} f_{2;ij}\xi_{ij+1} \tag{4.23}$$

as desired.

On the other hand suppose (4.18) holds. The differential equation on the right restricts to the hyperplane given by $x_{2k_1}=0$. This transforms to the hypersurface given by $\bar{\xi}_{2k_1}=f_2(\bar{\xi})$. The restricted systems are precisely those of the original (4.17). Q.E.D.

This completes the proof of Theorem 4.1 in the course of which we have used Theorem 4.2.

Proof of Theorem 4.2. We start with a system in observable form satisfying $f_i(\xi)\in\mathscr{P}^{k_i}(\xi)$ and at least two distinct observability indices, (if the observability indices are all the same there is nothing to prove.) Similar to before we define vector fields $g^1(\xi),\cdots,g^p(\xi)$ to be the unit vectors in the $\xi_{1k_1},\cdots,\xi_{pk_p}$ directions, i.e.,

$$L_{g^j}L_f^l(\psi_i)=\begin{cases}0, & 0\leqq l<k_i-1,\\ \delta_i^j, & l=k_i-1.\end{cases}\tag{4.24}$$

As we noted before in § 2 the codistributions

$$\mathscr{E}^k=\operatorname{Span}\{L_f^l(dh_i)\colon 1\leqq i\leqq p;\ 0\leqq l<k\}$$

are invariant under changes of state and output coordinates. The vector fields $g^1,\cdots,g^p$ and their brackets under f span the dual distributions, $\mathscr{D}^k=\mathscr{E}^{k\perp}$, given by

$$\mathscr{D}^k=\operatorname{Span}\{ad^l(-f)g^j\colon k_j>k,\ 0\leqq l<k_j-k\};$$

hence these are also invariant. Moreover the distribution obtained by bracketing the elements of $\mathscr{D}^{k_i}$ with f up to k_i times is also invariant; we denote this by $\tilde{\mathscr{D}}^{k_i}$

$$\tilde{\mathscr{D}}^{k_i}=\{ad^l(-f)g^j\colon k_j>k_i;\ 0\leqq l<k_j\}.$$

It is straightforward to verify that

$$L_{ad^l(-f)g^j}(\psi_i)=\begin{cases}1 & \text{if } i=j \text{ and } l=k_j-1,\\ f_{i;jk_i+1} & \text{if } k_j>k_i \text{ and } l=k_j-1,\\ 0 & \text{otherwise,}\end{cases}$$

so the image of $\bar{\mathscr{D}}^{k_i}$ under $d\psi$ is precisely the distribution $\mathscr{Y}^i$ on the output space. This shows that $\mathscr{Y}^i$ is independent of the output coordinates.

Now suppose we wish to choose output coordinates $\tilde{\psi}$ so that relative to these coordinates we have the special observable form; then $d\tilde{\psi}_i$ must annihilate $\mathscr{Y}^i$. Hence $\tilde{\psi}_i$ must satisfy (4.8).

But $\mathscr{Y}^1\subseteq\mathscr{Y}^2\subseteq\cdots\subseteq\mathscr{Y}^p$ so $d\tilde{\psi}_i,\cdots,d\tilde{\psi}_p$ also must annihilate the $p-i$ dimensional distribution $\mathscr{Y}^i$; hence $\mathscr{Y}^i$ must be involutive.

On the other hand if each $\mathscr{Y}^i$ is involutive, then we can choose p independent functions $\tilde{\psi}_1,\cdots,\tilde{\psi}_p$ such that $d\tilde{\psi}_i\perp\mathscr{Y}^i$. This is the desired output coordinate change. Q.E.D.

5. Sufficient condition. Let us review the previous sections. We start with a nonlinear system (2.1) around some nominal operating point ξ^0 for which we desire to build an observer. We first check that it is observable at ξ^0 by attempting to transform it into observable form (2.5); then we check if $f_i(\xi)\in\mathscr{P}^{k_i}(\xi)$ as described in Proposition 3.3. Next we attempt to make a change of output coordinates to get it into special observable form as described in Theorem 4.2. If this can be achieved, then we attempt to solve equation (4.6) using Theorem 4.1 to find the output $y=y(\psi)$. If we are able to accomplish all of this, we have the system in the form

$$\dot{\xi}=f(\xi),\tag{5.1a}$$

$$y=h(\xi),\tag{5.1b}$$

$$\xi(0) \approx \xi^0 \tag{5.1c}$$

which we wish to transform by change of state coordinates $\xi = \xi(x)$ into

$$\dot{x} = Ax + \alpha(y), \tag{5.2a}$$

$$y = \mathcal{C}x, \tag{5.2b}$$

$$y(0) \approx 0 \tag{5.2c}$$

where A, C are as in (4.10) but with possibly varying block sizes determined by the observability indices $k_1, \cdots, k_p$. The diagonal blocks of A and C are $k_i \times k_i$ and $1 \times k_i$ respectively.

The scalar output ($p = 1$) version of this problem was solved by Krener and Isidori [1]; the following theorems are straightforward generalizations.

THEOREM 5.1. *Let $\bar{g}^1(\xi), \cdots, \bar{g}^p(\xi)$ be vector fields defined by the equations*

$$L_{\bar{g}^j} L_f(y_i) = \begin{cases} 0, & 0 \leqq l \leqq k_i - 1, \\ \delta_i^j, & l = k_i - 1. \end{cases} \tag{5.3}$$

There exists a change of coordinates transforming (5.1) *to* (5.2) *iff*

$$[ad^k(-f)\bar{g}^i, ad^l(-f)\bar{g}^j] = 0 \tag{5.4}$$

for $i, j = 1, \cdots, p$, $k = 0, \cdots, k_i - 1$; $l = 0, \cdots, k_j - 1$. The appropriate coordinates $x = (x_{ik})$ are such that the vector field $ad^{k_i - k}(-f)\bar{g}^i$ is the unit vector in the x_{ik} direction,

$$L_{ad^{k_i - k}(-f)\bar{g}^i}(x_{jl}) = \delta_j^i \delta_l^k. \tag{5.5}$$

The appropriate functions $\alpha = (\alpha_{jl})$ can be computed by applying the state coordinate transformation to (5.1) and comparing the result with (5.2) or by solving the equations

$$\frac{\partial \alpha_{jl}}{\partial y_i} = L_{ad^{k_i}(-f)\bar{g}^i}(x_{jl}). \tag{5.6}$$

These are always solvable if (5.4) holds.

Remark 5.1. The repeated Lie brackets of vector fields X^1, X^2 and X^3 satisfy the Jacobi identity

$$[X^1[X^2, X^3]] = [[X^1, X^2]X^3] + [X^2[X^1, X^3]]. \tag{5.7}$$

This leads to considerable redundancy in the conditions (5.4). Suppose the conditions hold for any $k < k_i$, $l < k_j$ and $k + l < r$.

If $k + l = r$, applying (5.7) we obtain

$$\begin{aligned} [ad^k(-f)g^i, ad^l(-f)g^j] = &-[ad^{k-1}(-f)g^i, ad^{l+1}(-f)g^j] \\ &+[f[ad^{k-1}(-f)g^i, ad^l(-f)g^j]] \end{aligned} \tag{5.8}$$

but the second on the right is zero by assumption. Hence for each i, j and r we need check (5.4) for only one value of k and l summing to r. Moreover because of the skewsymmetry of the bracket, (5.8) is skewsymmetric in i and j if r is even and symmetric if r is odd. Therefore for even r we need only check for $i < j$ and for odd r for $i = j$.

In particular if $p = 1$, (5.4) need only be checked for $k = l - 1$ and $l = 1, \cdots, n - 1$.

The condition that $f_i(\xi) \in \mathcal{P}^{k_i}(\xi)$ and the basic differential equation (4.2) are implied by (5.4) and as we shall see in the examples are sometimes equivalent to (5.4).

Remark 5.2. Suppose we have a system in special observable form relative to the output ψ and we have computed Ψ, the solution of (4.2). It is not necessary to compute $\psi(y)$ to verify (5.4).

If $g^1, \cdots, g^p$ are the vector fields defined by (4.24), then they are related to $\bar{g}^1, \cdots, \bar{g}^p$ by (4.10). With the help of (4.13) we can convert (5.4) into a family of differential equations which Ψ must satisfy.

Proof of Theorem 5.1. Suppose there exists a change of coordinates $\xi = \xi(x)$ transforming (5.1) to (5.2), then the vector fields $\bar{g}^1, \cdots, \bar{g}^p$ are transformed to constant vector fields in the x_{ik_i} direction. Let B be as in (4.10) with block sizes determined by the observability indices, the diagonal blocks are $k_i \times 1$.

Then $ad^l(-f)g^j = A^l B^j$ for $j = 1, \cdots, p$ and $l = 0, \cdots, k_{j-1}$ so clearly (5.4) holds. (Note: A^l is lth power of A, B^j the jth column of B.)

On the other hand suppose (5.4) holds. These are the integrability conditions for the set of partial differential equations (5.5) so there must exist coordinates x in which $ad^{k_j-1}(-f)\bar{g}^j$ is the unit vector in the x_{jl} direction.

We wish to compute $\dot{x}_{jl} = L_f(x_{jl})$.

If $1 \leqq k \leqq k_i$,

$$\frac{\partial \dot{x}_{jl}}{\partial x_{ik}} = L_{ad^{k_i-k}(-f)\bar{g}^i} L_f(x_{jl}) = L_{ad^{k_i-k+1}(-f)\bar{g}^i}(x_{jl}) + L_f L_{ad^{k_i-k}(-f)\bar{g}^i}(x_{jl}).$$

From (5.5) we see that

$$\frac{\partial \dot{x}_{jl}}{\partial x_{ik}} = \begin{cases} \delta^i_j \delta^{k+1}_l & \text{if } k > 1, \\ L_{ad^k(-f)\bar{g}^i}(x_{jl}) & \text{if } k = 1. \end{cases}$$

But $x_{i1} = y_i$ so

$$\dot{x}_{jl} = \begin{cases} x_{jl+1} + \alpha_{jl}(y), & l < k_j, \\ \alpha_{jk_j}(y), & l = k_j \end{cases}$$

where α_{jl} is the solution of (5.6).

These are first order partial differential equations so they are solvable if the mixed partials agree.

$$\frac{\partial}{\partial y_r}\frac{\partial \alpha_{jl}}{\partial y_i} = L_{ad^{k_r-1}(-f)\bar{g}^r} L_{ad^{k_i}(-f)\bar{g}^i}(x_{jl})$$

$$= L_{[ad^{k_i-1}(-f)\bar{g}^r,\, ad^{k_i}(-f)\bar{g}^i]}(x_{jl}) - L_{ad^{k_i}(-\bar{f})\bar{g}^i} L_{ad^{k_r-1}(-f)\bar{g}^r}(x_{jl}).$$

The second term on the right is zero by (5.5). Skew symmetry, the Jacobi identity (5.7) and (5.4) yield

$$[ad^{k_r-1}(-f)\bar{g}^r, ad^{k_i}(-f)\bar{g}^i] = [ad^{k_i-1}(-f)\bar{g}, ad^{k_r-1}(-f)\bar{g}^i]$$

so the mixed partials agree. Q.E.D.

6. Systems with inputs. The previous method can be easily generalized to handle systems with inputs

(6.1a) $$\dot{\xi} = f(\xi, u),$$

(6.1b) $$\psi = h(\xi),$$

(6.1c) $$\xi^0 \approx \xi(0),$$

(6.1d) $$\psi^0 = h(\xi^0).$$

We seek a change of output and state coordinates which transforms (6.1) into

$$\dot{x} = Ax + \gamma(y, u), \tag{6.2a}$$

$$y = Cx, \tag{6.2b}$$

$$x^0 = x(\xi^0) = 0, \tag{6.2c}$$

$$y^0 = u(\psi^0) = 0. \tag{6.2d}$$

If A, C is in dual Brunovsky form as given by (4.10), we say that (6.2) is in observer form. The system

$$\dot{z} = (A + GC)z + \gamma(y, u) - Gy, \tag{6.3a}$$

$$z(0) = 0 \tag{6.3b}$$

tracks (6.2) with the error $e = x - z$ having dynamics

$$\dot{e} = (A + GC)e, \tag{6.3c}$$

$$e(0) = x(\xi(0)) \approx 0. \tag{6.3d}$$

Once again by proper choice of G we can insure that $e(t)$ goes to zero with arbitrary exponential decay.

To reduce this problem to the one considered previously we first choose a nominal input, either a constant u^0 or a function of ψ, $u^0(\psi)$. From a mathematical point of view the choice is immaterial. But of course the mathematical model is never an exact description of the real world; to reduce the effect of modeling errors the nominal control should be typical or average in some sense of the controls that will be employed.

We then rewrite (6.1a) as

$$\dot{\xi} = f^0(\xi) + f^1(\xi, u) \tag{6.4}$$

where

$$f^0(\xi) = f(\xi, u^0(\psi(\xi))), \tag{6.5a}$$

$$f^1(\xi, u) = f(\xi, u) - f^0(\xi) \tag{6.5b}$$

and proceed as before with the unforced system (6.6) and (6.1b, c, d)

$$\dot{\xi} = f^0(\xi). \tag{6.6}$$

If this can be transformed into observer form (6.7) and (6.2b, c, d)

$$\dot{x} = Ax + \alpha(y) \tag{6.7}$$

by change of state and output coordinates, then all one need check is that $f^1(\xi, u)$ is transformed into a vector field of the form

$$\beta(y, u). \tag{6.8}$$

If this is possible, then $\gamma(y, u) = \alpha(y) + \beta(y, u)$ and the problem is solved. If unforced system (6.6) and (6.1b, c, d) cannot be transformed into observer form or $f^1(\xi, u)$ does not transform into (6.8) then the original system (6.1) cannot be transformed into observer form (6.2).

As we remarked before the choice of the nominal control $u^0(\psi)$ is immaterial; a system (6.1) can be transformed into observer form (6.2) iff every unforced closed loop version (6.6) can be transformed into observer form by the same changes of coordinates.

Two nonlinear coordinate changes which transform a system (with or without inputs) into observer form differ by a linear change of coordinates. For such systems the output feedback $u = u^0(\psi)$ affects neither the observability nor the observability indices. However it is possible that (6.1) does not have an observer form yet one or more unforced closed loop versions do. If there is more than one, then typically these will involve different coordinate changes and perhaps even different observability indices.

7. Examples. We consider several simple cases of the above method for transforming a system into observer form.

Example 7.1. $p = 1$, $n = k_1 = 2$. In observable form we have the system

$$\begin{aligned} &\psi = \xi_1, \\ &\dot{\xi}_1 = \xi_2, \qquad \xi^0 = \begin{pmatrix} \xi_1^0 \\ \xi_2^0 \end{pmatrix}, \quad \psi^0 = \xi_1^0. \\ &\dot{\xi}_2 = f_1(\xi), \end{aligned} \tag{7.1}$$

Proposition 3.3 requires that $f_2 \in \mathscr{P}^2(\xi)$; hence

$$f_2(\xi) = a(\xi_1) + b(\xi_1)\xi_2 + c(\xi_1)\frac{\xi_2^2}{2}. \tag{7.2}$$

If this holds then since there is only one output the system is in special observable form, i.e., $f_2 \in \mathscr{P}_0^2(\xi)$. The differential equation (4.2) for $\Psi = d\psi/dy$ is

$$\frac{d\Psi}{d\psi} = \frac{1}{2} f_{1;2;2}\Psi(\psi) = \frac{1}{2} c(\psi)\Psi(\psi); \tag{7.3}$$

the solution is

$$\Psi(\psi) = \exp\left(\int_{\psi^0}^{\psi} \frac{c}{2}(\nu)\, d\nu\right) \tag{7.4}$$

where we have normalized the constant of integration so that $\Psi(\psi^0) = 1$. Next we check condition (5.4) using the identities (4.13)

$$\begin{matrix} g & ad(-f)g & \bar{g} & ad(-f)\bar{g} & [\bar{g}, ad(-f)\bar{g}] \end{matrix}$$
$$\begin{pmatrix} 0 \\ 1 \end{pmatrix} \begin{pmatrix} 1 \\ f_{1;2} \end{pmatrix} \begin{pmatrix} 0 \\ \Psi \end{pmatrix} \begin{pmatrix} \Psi \\ f_{1;2}\Psi - \Psi^{(1)}\xi_2 \end{pmatrix} \begin{pmatrix} 0 \\ f_{1;2;2}\Psi - 2\Psi^{(1)} \end{pmatrix}. \tag{7.5}$$

Equation (7.3) implies that (5.4) holds and hence the system can be transformed into observer form.

The required change of output coordinates is obtained by integrating (7.4) to

$$y(\psi) = \int_{\psi^0}^{\psi} \exp\left(-\int_{\psi^0}^{\mu} \frac{c}{2}(\nu)\, d\nu\right) d\mu \tag{7.6}$$

where $\psi^0 = \xi_1^0$ and the limits of integration have been chosen so that $y(\psi^0) = 0$. Since $\psi = \xi_1$ and $y = x_1$ this gives half of the change of state coordinates. The other coordinate $x_2(\xi)$ must satisfy (5.5) which reduce to

$$L_{\bar{g}(x_2)} = 1, \; L_{ad(-f)\bar{g}(x_2)} = 0. \tag{7.7}$$

From (7.5) this becomes

$$\frac{\partial x_2}{\partial \xi_1} = -\Psi^{-1}\left(\frac{b + c\xi_2}{2}\right), \qquad \frac{\partial x_2}{\partial \xi_2} = \Psi^{-1}. \tag{7.8}$$

These are easily integrated to obtain

$$x_2(\xi) = \Psi^{-1}\xi_2 - \xi_2^0 - \int_{\xi^0}^{\xi_1} \Psi^{-1}(\nu) b(\nu)\, d\nu \tag{7.9}$$

where the constant of integration has been chosen so that $x_2(\xi^0) = 0$.

Finally we compute α. Comparing the time derivative of (7.6) with (7.9) yields

$$\alpha_1(\xi_1) = \xi_2^0 + \int_{\xi_1^0}^{\xi_1} \Psi^{-1}(\nu) b(\nu)\, d\nu. \tag{7.10}$$

Time differentiating (7.9) yields

$$\alpha_2(\xi_1) = \Psi^{-1}(\xi_1) a(\xi_1). \tag{7.11}$$

Notice $\alpha_i = \alpha_i(\psi) = \alpha_i(\psi(y))$ as desired.

Hence we have that an $n = 2, p = 1$ system can be transformed to observer form iff it satisfies Proposition 3.3.

Example 7.2. $n = 20, k_1 = \cdots = k_p = 2$. The analysis is very similar to previous example. In observable form we have for $i = 1, \cdots, p$.

$$\begin{aligned} \psi_i &= \xi_{i1}, \\ \dot{\xi}_{i1} &= \xi_{i2}, \\ \xi_{i2} &= f_i(\xi). \end{aligned} \tag{7.12}$$

Let

$$\xi_1 = \begin{pmatrix} \xi_{11} \\ \vdots \\ \xi_{1p} \end{pmatrix}, \quad g_2 = \begin{pmatrix} \xi_{12} \\ \vdots \\ \xi_{p2} \end{pmatrix}, \quad x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ z_{1p} \end{pmatrix}, \quad x_2 = \begin{pmatrix} x_{21} \\ \vdots \\ x_{2p} \end{pmatrix}.$$

Again f_i must be quadratic in ξ_{i2} so

$$f_i(\xi) = a_i(\xi_1) + b_i(\xi_1)\xi_2 + \tfrac{1}{2}\xi_2' c_i(\xi_1)\xi_2$$

where a_i is a scalar and $b_i = (b_{ij})$ and $c_i = (x_{ikl})$ are $1 \times p$ and symmetric $p \times p$ matrix valued functions. The partial differential equations (4.2) become

$$\frac{\partial}{\partial \psi_k} \Psi_i^j = \frac{1}{2} \sum_{l=1}^{p} c_{ikl}(\psi) \Psi_l^j(\psi). \tag{7.13}$$

It is convenient to define $p \times p$ matrix valued functions

$$\Gamma^k(\psi) = (\Gamma_{ij}^k(\psi)) = (\tfrac{1}{2} c_{ikj}(\psi))$$

for $k = 1, \cdots, p$. We rewrite (7.13) as

$$\frac{\partial}{\partial \psi_k} \Psi = \Gamma^k \Psi. \tag{7.14}$$

If the integrability conditions

$$\frac{\partial \Gamma^k}{\partial \psi_l} - \frac{\partial \Gamma^l}{\partial \psi_k} = \Gamma^l \Gamma^k - \Gamma^k \Gamma^l \tag{7.15}$$

are satisfied, we have the solution

$$\Psi(\psi) = \exp\left(\sum_{i=1}^{p} \int_{\psi_i^0}^{\psi_i} \Gamma^k(\nu_i)\, d\nu_i \right). \tag{7.16}$$

Again (7.14) insures us that (5.4) holds, so the changes of coordinates exist. The change of output is obtained by integrating (7.15) via the line integral

$$y(\psi) = \int_{\psi^0}^{\psi} \Psi^{-1}(\nu)\, d\nu. \tag{7.17}$$

This is also half of the state coordinate change; the other half $x_2(\xi)$ satisfies (5.5) which becomes similar to

$$\frac{\partial x_2}{\partial \xi_1} = -\Psi^{-1}\left(b + \sum_{k=1}^{p} \Gamma^k \xi_{k2}\right), \qquad \frac{\partial x_2}{\partial \xi_2} = \Psi^{-1} \tag{7.18}$$

where b is the $p \times p$ matrix whose i,jth entry is the jth entry of b_i. The solution (7.18) is

$$x_2(\xi) = \Psi^{-1}\xi_2 - \xi_2^0 - \int_{\xi_1^0}^{\xi_1} \Psi^{-1}(\nu) b(\nu)\, d\nu \tag{7.19}$$

where the last term is a line integral. The computation of α is as before and given by the vector versions of (7.10) and (7.11).

Hence a $2p = n$, $k_1 = \cdots = k_p$ system can be transformed to observer form iff it satisfies Proposition 3.3 and the integrability conditions (7.15).

Example 7.3. $p = 1$, $n = k_1 = 3$. In observable form we have the system

$$\begin{aligned} \psi &= \xi_1, \\ \dot{\xi}_1 &= \xi_2, \\ \dot{\xi}_2 &= \xi_3, \\ \dot{\xi}_3 &= f_1. \end{aligned} \tag{7.20}$$

Proposition 3.3 requires that f_1 be of the form

$$\begin{aligned} f_1(\xi) = a(\xi_1) + b(\xi_1)\xi_2 + c(\xi_1)\frac{\xi_2^2}{2} + d(\xi_1)\frac{\xi_2^3}{3} \\ + (\rho(\xi_1) + \sigma(\xi_1)\xi_2)\xi_3. \end{aligned} \tag{7.21}$$

The basic differential equation (4.2) is

$$\frac{d\Psi}{d\psi} = \frac{1}{3} f_{1;3;2}\Psi = \frac{1}{3}\sigma(\psi)\Psi \tag{7.22}$$

and the solution is

$$\Psi(\psi) = \exp \int_{\psi^0}^{\psi} \frac{\sigma(\nu)}{3}\, d\nu. \tag{7.23}$$

If we use (4.13), then after a laborious calculation (5.4) reduces to the two differential equations

$$\frac{d\sigma}{d\xi_1} = \frac{3}{2}d + \frac{2}{3}\sigma^2, \tag{7.24a}$$

$$\frac{d\rho}{d\xi_1} = c + \rho\sigma. \tag{7.24b}$$

Hence a $p = 1$, $n = 3$ system can be transformed to observer form iff Proposition 3.3 and equations (7.24) are satisfied. The rest of the calculations proceed as before.

Example 7.4. $p=2$, $n=3$, $k_1=2$, $k_2=1$. In observable form we have

$$\begin{aligned} &\psi_1=\xi_{11}, && \psi_2=\xi_{21}, \\ &\dot\xi_{11}=\xi_{12}, && \dot\xi_{21}=f_2(\xi), \qquad \psi=\xi_1=\begin{pmatrix}\xi_{11}\\ \xi_{21}\end{pmatrix}, \\ &\dot\xi_{12}=f_1(\xi), \end{aligned} \tag{7.25}$$

where by Proposition 3.3

$$f_1(\xi)=a(\xi_1)+b(\xi_1)\xi_{12}+c(\xi_1)\frac{\xi_{12}^2}{2}, \tag{7.26a}$$

$$f_2(\xi)=\rho(\xi_1)+\sigma(\xi_1)\xi_{12}. \tag{7.26b}$$

First we transform this to special observable form by Theorem 4.2. We seek $\tilde\psi_2(\psi)$ such that

$$L_{Y^1}(\tilde\psi_2)=0 \quad \text{where } Y^1=\begin{pmatrix}1\\ f_{2;12}\end{pmatrix}=\begin{pmatrix}1\\ \sigma(\xi_1)\end{pmatrix} \tag{7.27}$$

or

$$\frac{\partial\tilde\psi_2}{\partial\psi_1}+\sigma(\psi)\frac{\partial\tilde\psi_2}{\partial\psi_2}=0.$$

This is always solvable. In observable form (2.5) relative to the new outputs $\tilde\psi_1=\psi_1$ and $\tilde\psi_2$ we have

$$\tilde f_1(\tilde\xi)=\tilde a(\tilde\xi_1)+\tilde b(\tilde\xi_1)\tilde\xi_{12}+\tilde c(\tilde\xi_1)\frac{\tilde\xi_{12}^2}{2}, \tag{7.28a}$$

$$\tilde\phi_2(\tilde\xi)=\tilde\rho(\xi_1). \tag{7.28b}$$

At this point the presence of the second output is immaterial and we proceed essentially as in Example 7.1 carrying $\tilde\psi_2=\tilde\xi_{21}$ as a parameter. Hence a $p=2$, $n=3$, $k_1=2$, $k_2=1$ system is transformable to observer form iff it satisfies Proposition 3.3.

REFERENCES

[1] A. J. KRENER AND A. ISIDORI, *Linearization by output injection and nonlinar observers*, Systems and Control Letters, 3 (1983), pp. 47-52.

[2] B. JAKUBCYZK AND W. RESPONDEK, *On the linearization of control systems*, Bull. Acad. Poln. Sci. Ser. Sci. Math. Aston. Phys., 28 (1980), pp. 517-522.

[3] L. R. HUNT AND R. SU, *Linear equivalents of nonlinear time varying systems*, International Symposium on the Mathematical Theory of Networks and Systems, Santa Monica, CA, 1981, pp. 119-123.

[4] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 728-740.

[5] A. J. KRENER, *(Adf, g), (adf, g) and locally (adf, g) invariant and controllability distributions*, this Journal, 23 (1985), to appear.

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

[7] D. BESTLE AND M. ZEITZ, *Canonical form observer design for non-linear time variable systems*, Int. J. Control, 38 (1983), pp. 419-431.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

004

IDENTIFICATION OF PARAMETERS IN DISTRIBUTED PARAMETER SYSTEMS BY REGULARIZATION*

COSTAS KRAVARIS†‡ AND JOHN H. SEINFELD†

Abstract. Identification of spatially varying parameters in distributed parameter systems from noisy data is an ill-posed problem. The concept of regularization, widely used in solving linear Fredholm integral equations, is developed for the identification of parameters in distributed parameter systems. A general regularization identification theory is first presented and then applied to the identification of parabolic systems. The performance of the regularization identification method is evaluated by numerical experiments on the identification of a spatially varying diffusivity in the diffusion equation.

Key words. system identification, parameter estimation, distributed parameter systems, inverse problems, ill-posed problems, regularization

1. Introduction.

Consider the following distributed parameter dynamic system:

$$\frac{\partial u}{\partial t}+A(t)u=f \quad \text{in } \Omega\times]0,\ T[, \tag{1.1}$$

$$u(x,0)=u_0 \quad \text{in } \Omega, \tag{1.2}$$

$$B_j u=g_j, \qquad j=0,\cdots,m-1 \quad \text{on } \Gamma\times]0,\ T[, \tag{1.3}$$

where $\Omega\subset\mathbb{R}^n$ with boundary Γ and $0<T<\infty$ and where

$$A(t)u=\sum_{|p|,|q|\leqq m}(-1)^{|p|}D_x^p(a_{pq}(x,t)D_x^q u), \tag{1.4}$$

$$B_j u=\sum_{|h|\leqq m_j} b_{jh}(x,t)D_x^h u, \qquad j=0,\cdots,m-1, \tag{1.5}$$

with $0\leqq m_j=$ order of $B_j\leqq 2m-1$.

The parameter identification problem associated with (1.1)-(1.5) can be stated as follows: Assuming the input function f, the initial condition and the boundary condition(s) to be known, and given an observation of u, determine the system operator $A(t)$, i.e. the parameters $a_{pq}(x,t)$.[1]

A number of important physical identification problems fall within the above framework. For example, the partial differential equation

$$\frac{\partial u}{\partial t}-\frac{\partial}{\partial x}\left(\alpha(x,y)\frac{\partial u}{\partial x}\right)-\frac{\partial}{\partial y}\left(\alpha(x,y)\frac{\partial u}{\partial y}\right)=f(x,y,t) \tag{1.6}$$

governs the temperature distribution in an inhomogeneous solid or the pressure distribution in a fluid-containing porous medium. The local flux of energy or fluid is dependent on the value of the parameter α. For example, in the case of fluid flow in a porous medium, α is termed the transmissivity. For models of petroleum reservoirs and subsurface aquifers the transmissivity is generally inaccessible to direct measurement, and its value must be inferred from measurements of the pressure u at wells.

* Received by the editors December 14, 1982, and in revised form April 2, 1984.

† Department of Chemical Engineering, California Institute of Technology, Pasadena, California 91125.

‡ Present address: Department of Chemical Engineering, University of Michigan, Ann Arbor, Michigan 48109.

[1] The boundary condition parameters $b_{jh}(x,t)$ may also be unknown, although we do not consider that case here.

Because of the economic importance of knowing accurately the properties of subsurface aquifers and petroleum reservoirs, a great deal of effort has been expended in developing techniques for determining transmissivity from measurements of pressure [5], [9], [10], [12], [14], [24], [25], [32]. The determination of α from data on u is a special case of the general linear parabolic system identification problem introduced at the outset. Specifically, given f, the initial condition and appropriate boundary conditions, and given measurements z_{d_i} of $u(x_i, y_i, t)$ at a set of discrete spatial locations, $i = 1, 2, \cdots, \mu$, it is desired to determine, or identify, $\alpha(x, y)$.

The key difficulty in developing successful numerical techniques for identifying spatially-dependent parameters is the fact that such problems are ill-posed. To see this, consider (1.6) as a first order hyperbolic equation in α. One can easily show that the characteristics $\omega(x, y) = c$ are orthogonal to the lines of constant u. Thus, one can define a new curvilinear coordinate system $(\mathbf{e}_u, \mathbf{e}_\omega)$ so that $\mathbf{e}_u$ is unitary in $\mathbb{R}^2$ and the metric factor in the ω-coordinate is 1. Equation (1.6) can be written as

$$|\nabla u| \frac{\partial}{\partial u}(\alpha |\nabla u|) = \frac{\partial u}{\partial t} - f \quad \text{where } |\nabla u| = \left(\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 \right)^{1/2}. \tag{1.7}$$

Upon integration of (1.7), we obtain

$$\alpha(x, y) = \frac{\int (\partial u / \partial t - f)\, dl_\omega}{|\nabla u|} \tag{1.8}$$

where integration is performed along the characteristics and l_ω denotes Lebesgue measure along the characteristics. Ill-posedness follows from the fact that the differentiation operator is not continuous with respect to any physically meaningful observation topology. The fact that the identification problem associated with (1.6) is not well-posed can also be illustrated by counterexample [32]. In summary, the problem of identifying spatially-dependent coefficients appearing in the differential operator of a partial differential equation is, in general, both nonlinear and ill-posed [17], [18].

The customary way to approach the identification of α in (1.6) has been by least-squares, i.e., by minimizing the functional

$$J_{LS} = \int_0^T \sum_{i=1}^{\mu} [u(x_i, y_i, t) - z_{d_i}]^2 \, dt \tag{1.9}$$

subject to (1.6), initial and boundary conditions. There have been two ways of treating the unknown parameter α. In the first, α is considered as an element of an infinite-dimensional function space [9], [5], whereas in the second, the minimization is performed over a finite-dimensional subspace, reducing the problem to one of determining a finite number of constant parameters [10], [12]. When the number of parameters is kept small in this approach, a well-behaved solution results. However, the modeling error introduced is significant, since the corresponding subspace of α's is too restricted to provide a good approximation of an arbitrary α. As the number of parameters is increased, on the other hand, numerical instabilities appear, manifested by spatial oscillations in the estimated α, the frequency and amplitude of which are inconsistent with the expected smoothness of the true α. The symptoms usually also include a flat global minimum in J_{LS} [24], [25], [32]. The same instability phenomena characterize the minima of J_{LS} over an infinite-dimensional function space. One approach that has been somewhat successful in alleviating numerical instabilities involves the incorporation of a priori statistics concerning α into the minimization by adding a Bayesian term in the performance index (1.9) [14], [24]. The major drawback to this approach is that reliable a priori statistics for α are seldom available. Thus, there is a need to

develop a rigorously based approach for identifying parameters in partial differential equations from noisy data that is numerically stable and physically consistent with the expected character of the unknown parameters.

The numerical instabilities and ill-posed nature of the problem of interest strongly suggest a regularization approach. "Regularization" of a problem refers in general to solving a related problem, called the regularized problem, the solution of which is more regular, in a sense, than that of the original problem and approximates the solution of the original problem. When referring to ill-posed problems, regularization is an approach to circumvent lack of continuous dependence on the data. The regularized problem is a well-posed problem whose solution yields a physically meaningful answer to the given ill-posed problem.

The idea of regularization of ill-posed problems was first proposed by Tikhonov [27], [28] as a method of solving linear Fredholm integral equations of the first kind. Further development of the theory for ill-posed linear operator equations followed [23]. Modern practical numerical methods for the solution of linear Fredholm integral equations involve regularization [31].

The object of the present work is to develop a regularization theory for the identification of parameters in distributed parameter systems. In § 2 we define the parameter identification problem in an abstract manner that facilitates proof of the major theorems. The concept of identifiability is discussed in § 3. In §§ 4 and 5 a general regularization identification theory is presented. In §§ 6, 7 and 8 the theory is applied to the identification of parabolic systems from distributed and point observations. Finally, in § 9 numerical results are given for the identification of a spatially-varying diffusivity in the one-dimensional diffusion equation.

2. Problem statement. To develop a general identification theory, we introduce the following abstract problem.

Let $\mathcal{A}$, U and F be Banach spaces. Consider a system described by

$$\Psi(A, u) = f \tag{2.1}$$

where Ψ is a mapping, not necessarily linear, from $\mathcal{A} \times U$ into F. We assume:

(A1) Ψ is of C^k-class ($k \geqq 1$).

(A2) There is an open subset $\mathcal{A}_c$ of $\mathcal{A}$ and an open subset U_c of U such that $\forall A \in \mathcal{A}_c$ (2.1) admits a unique solution $u \in U_c$.

(A3) $\forall A \in \mathcal{A}_c$, $\forall u \in U_c$, $(\partial\Psi/\partial u)(A, u)$ is a linear homeomorphism of U onto F.

Furthermore, consider that A depends on a set of parameters λ belonging to the Banach space Λ. The set of physically admissible λ is $\Lambda_{ad} \subseteq \Lambda$. We assume:

(A4) $A: \Lambda \to \mathcal{A}$ is of C^k-class ($k \geqq 1$).

(A5) Λ_{ad} is a norm-closed convex subset of Λ.

(A6) $A(\Lambda_{ad}) \subseteq \mathcal{A}_c$.

Now from the implicit function theorem ([26, pp. 277–304]) we have:

PROPOSITION 2.1. *Assume* (A1)–(A3) *are valid. Then the implicit function,* $u = \Phi(A)$, *defined as the solution of* (2.1) *is of* C^k*-class from* $\mathcal{A}_c$ *into* U_c. *Its first derivative is given by*

$$\Phi'(A) = -\left[\frac{\partial\Psi}{\partial u}(A, u)\right]^{-1} \circ \left[\frac{\partial\Psi}{\partial A}(A, u)\right] \forall A \in \mathcal{A}_c. \tag{2.2}$$

Equivalently, $\Phi'(A)$ *associates* $\delta A \in \mathcal{A} \to \delta u \equiv \Phi'(A) \cdot \delta A \in U$, *where* δu *is the solution of*

$$\frac{\partial\Psi}{\partial u}(A, u) \cdot \delta u + \frac{\partial\Psi}{\partial A}(A, u) \cdot \delta A = 0. \tag{2.3}$$

As an immediate consequence, we have

PROPOSITION 2.2. *Assume that* (A1)-(A4) *and* (A6) *are valid. Then* $\Phi \circ A: \Lambda_{ad} \to U$ *is of* C^k*-class. Its first derivative* $(\Phi \circ A)'(\lambda)$ *associates* $\delta\lambda \in \Lambda_{ad} \to \delta u \in U$, *where* δu *is the solution of*

$$\frac{\partial \Psi}{\partial u}(A(\lambda), u) \cdot \delta u + \frac{\partial \Psi}{\partial A}(A(\lambda), u) \circ A'(\lambda) \cdot \delta\lambda = 0.$$

Now the identification problem can be posed as follows:

Knowing the mappings $\Psi: \mathscr{A} \times U \to F$ and $A: \Lambda \to \mathscr{A}$ and the element $f \in F$ and given an observation of u, find $\lambda \in \Lambda_{ad}$ to satisfy (2.1).

We need to be precise about the nature of the observation of u. Thus, consider a Hilbert space $\mathscr{H}$ (observation space). Denote by $\Lambda_{\mathscr{H}}$ the canonical isomorphism of $\mathscr{H}$ onto $\mathscr{H}'$. Also, consider an observation operator, not necessarily linear, $\mathscr{C}: U \to \mathscr{H}$ and assume

(A7) $\mathscr{C}$ is of C^k-class ($k \geqq 1$).

The situation is depicted in Fig. 1.

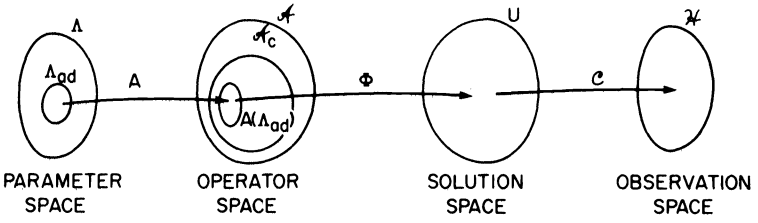


FIG. 1. *Function spaces for the abstract identification problem.*

3. Identifiability. The identification problem, as defined in § 2 can be viewed as solving in Λ_{ad} the (nonlinear) operator equation

$$(\mathscr{C} \circ \Phi \circ A)(\lambda) = z_d. \tag{3.1}$$

Before one develops an identification method, two key issues need to be examined:

(a) Whether equation (3.1) admits a unique solution (identifiability).

(b) Whether the solution of (3.1) depends continuously on the data z_d (stability).

The purpose of this section is twofold:

(i) To discuss the available concepts of identifiability and introduce two new concepts: conditional identifiability and pointwise identifiability. The latter is a special case of the former and will play an important role in §§ 4 and 5.

(ii) To stress the importance of stability and illustrate (through homogenization theory) why identification of spatially-varying parameters in distributed parameter systems is unstable. Also, to point out that stability is a necessary condition for output least-square identifiability. These considerations will motivate the treatment of identification problems as ill-posed problems.

DEFINITION 3.1 [4], [15], [7]. A parameter λ is said to be *identifiable* in Λ_{ad} for the observation operator $\mathscr{C}$, if the mapping (parameter → observation) is injective, i.e. if $\mathscr{C} \circ \Phi \circ A: \Lambda_{ad} \to \mathscr{H}$ has a unique inverse.

DEFINITION 3.2. A parameter λ is said to be *stable* in Λ_{ad} for the observation operator $\mathscr{C}$, if $(\mathscr{C} \circ \Phi \circ A)^{-1}$ is continuous.

Remark. In case of nonidentifiable λ, stability is understood in the sense of continuity of multiple-valued mappings.

The identifiability of the parameter $\alpha(x)$ in

$$\frac{\partial u}{\partial t} - \sum_{j=1}^{n} \frac{\partial}{\partial x_j}\left(\alpha(x)\frac{\partial u}{\partial x_j}\right) = f \tag{3.2}$$

from a distributed observation of u, i.e. an observation of $u(x, t)$ in $\Omega \times]0, T[$, has been studied by [15] (one spatial dimension) and [4] (several spatial dimensions). It has been shown that in general $\alpha(x)$ need not be unique [15]. However, if the set $E(t) = \{x \in \bar{\Omega} | (\partial z_d/\partial x)(x, t) = 0\}$ is nonempty for every $t \in]0, T[$ and $\cap_{t\in]0,T[} E(t)$ is of measure zero, then $\alpha(x)$ is unique. The result in [4] is similar to that in [15], but involves quite restrictive assumptions concerning ∇z_d and $\nabla^2 z_d$. Thus, these results establish that $\alpha(x)$ in (3.2) is not identifiable in the sense of Definition 3.1; however, under certain additional conditions on z_d, there corresponds a unique $\alpha(x)$.

Due to the conditional nature of most distributed parameter identifiability results (see [7] for a review), we find it important to introduce a weaker concept of identifiability.

DEFINITION 3.3. A parameter λ is said to be *conditionally identifiable* in Λ_{ad} with respect to $\mathcal{H}_c \subset \mathcal{H}$, if the restriction of the mapping $\mathcal{C} \circ \Phi \circ A : \Lambda_{ad} \to \mathcal{H}$ on the set $(\mathcal{C} \circ \Phi \circ A)^{-1}\mathcal{H}_c$ has a unique inverse.

A degenerate case of conditional identifiability is obtained when $\mathcal{H}_c$ is a point set, i.e. $\mathcal{H}_c = \{\tilde{z}_d\} \subset \mathcal{C}(\Phi(A(\Lambda_{ad})))$.

DEFINITION 3.4. A parameter λ is said to be *pointwise identifiable* in Λ_{ad} for the observation $\tilde{z}_d \in \mathcal{C}(\Phi(A(\Lambda_{ad})))$, if $\tilde{z}_d$ has a unique preimage with respect to the mapping $\mathcal{C} \circ \Phi \circ A : \Lambda_{ad} \to \mathcal{H}$.

The concept of pointwise identifiability is the weakest possible concept of identifiability. It will be used in §§ 4 and 5 (Theorems 4.3 and 5.3).

As we have noted, the identification of distributed coefficients appearing in the differential operator of a partial differential equation is, as a rule, an unstable problem [18]. The homogenization theory [2] shows that operators with highly oscillatory coefficients can be "replaced" by very different operators and still yield practically the same response. Lions [18], has, in fact, cited the main difficulty in identifying distributed coefficients in partial differential equations as preventing excess of oscillations in the coefficients.

To illustrate the power of homogenization theory in proving instability of identification problems, let us consider the problem of identifying $\alpha(x)$ in (3.2).

Let $Y =]0, y_1^0[\times]0, y_2^0[\times \cdots \times]0, y_n^0[\subset \mathbb{R}^n$ and $\alpha : \mathbb{R}^n \to \mathbb{R}$ a function with properties

(i) $\alpha \in L^\infty(\mathbb{R}^n)$;

(ii) $\alpha(y) \geqq \alpha_0 > 0$ a.e. in y;

(iii) $\alpha(y)$ is Y-periodic i.e. it admits a period y_j^0 in the direction y_j, $j = 1, \cdots, n$.

Denote $\alpha^\varepsilon(x) = \alpha(x/\varepsilon)$, $\varepsilon > 0$. Now given Ω a bounded open subset of $\mathbb{R}^n$ and $T > 0$, consider

$$\begin{aligned} &\frac{\partial u_\varepsilon}{\partial t} - \sum_{j=1}^{n} \frac{\partial}{\partial x_j}\left(\alpha^\varepsilon(x)\frac{\partial u_\varepsilon}{\partial x_j}\right) = f \quad \text{in } \Omega \times]0, T[, \\ &u_\varepsilon(x, 0) = u_0(x) \quad \text{in } \Omega, \\ &\text{boundary condition.} \end{aligned} \tag{3.3}$$

Observe that as $\varepsilon \to 0$, the α^ε's are *highly oscillating functions.* They converge in a weak sense:

(3.4)

$$\alpha^\varepsilon \to \mathcal{M}(\alpha) \quad \text{in } L^\infty(\Omega) \text{ weak-*} \qquad \left(\text{i.e. } \int_\Omega \alpha^\varepsilon \phi \, dx \to \int_\Omega \mathcal{M}(\alpha)\phi \, dx \; \forall \phi \in L^1(\Omega)\right),$$

where $\mathcal{M}(\alpha) = (1/\text{meas } Y) \int_Y \alpha(y) \, dy$.

The question now concerns the behavior of the solution u_ε of (3.3) as $\varepsilon \to 0$. It is tempting to believe that u converges to the solution of

$$\frac{\partial u}{\partial t} - \mathcal{M}(\alpha)\Delta u = f \quad \text{in } \Omega \times]0, T[,$$

$$u(x, 0) = u_0(x) \quad \text{in } \Omega, \tag{3.5}$$

$$\text{boundary condition.}$$

But this is untrue [2, p. 242]. The correct result is given by the following proposition, which is an immediate consequence of a general result for second-order parabolic systems [2, pp. 241–243].

PROPOSITION 3.1. *The solution u_ε of* (3.3) *converges in $L^2(\Omega \times]0, T[)$ to the solution of the following homogenized problem*:

$$\frac{\partial u}{\partial t} - \frac{1}{\mathcal{M}(1/\alpha)}\Delta u = f \quad \textit{in } \Omega \times]0, T[,$$

$$u(x, 0) = u_0(x) \quad \textit{in } \Omega, \tag{3.6}$$

$$\textit{boundary condition.}$$

Thus, for sufficiently small ε, u_ε is approximately equal to the solution of (3.6); however α^ε and $1/\mathcal{M}(1/\alpha)$ can be very different.

The least-squares approach to distributed parameter system identification [4], [8] can be stated as follows:

Given $z_d \in \mathcal{H}$, find $\bar{\lambda} \in \Lambda_{ad}$ to minimize the functional

$$J_{LS}(\lambda) = \| \mathcal{C}(\Phi(A(\lambda))) - z_d \|_{\mathcal{H}}^2. \tag{3.7}$$

Conceptually, the least-squares approach consists of two steps:

(a) Project z_d in $\bar{z}_d$ on the set $\mathcal{C}(\Phi(A(\Lambda_{ad})))$.

(b) Find in Λ_{ad} a preimage $\bar{\lambda}$ of $\bar{z}_d$ for the mapping $\mathcal{C} \circ \Phi \circ A$.

It is therefore natural to inquire if a projection of an arbitrary $z_d \in \mathcal{H}$ on the set $\mathcal{C}(\Phi(A(\Lambda_{ad})))$ exists and is unique. Also, when z_d is perturbed slightly, does the perturbation correspond to a small change in $\bar{\lambda}$?

DEFINITION 3.3 [6]. A parameter λ is said to be *output least-square identifiable* (OLSI) in Λ_{ad} for the observation operator $\mathcal{C}$, if there exists a neighborhood $\mathcal{N} \supset \mathcal{C}(\Phi(A(\Lambda_{ad})))$ such that for every $z_d \in \mathcal{N}$ the least-squares estimate is unique and depends continuously on z_d.

It is easy to see that the following are necessary conditions for OLSI:

(i) Existence of a neighborhood $\mathcal{N} \supset \mathcal{C}(\Phi(A(\Lambda_{ad})))$ such that every $z_d \in \mathcal{N}$ has a unique projection on $\mathcal{C}(\Phi(A(\Lambda_{ad})))$.

(ii) Well-posedness of the identification problem for every $z_d \in \mathcal{C}(\Phi(A(\Lambda_{ad})))$, i.e. both identifiability and stability of λ in Λ_{ad} w.r.t. $\mathcal{C}$ (in the sense of Definitions 3.1 and 3.2.).

It has been shown in [6] that with Λ_{ad} convex and $\mathcal{C} \circ \Phi \circ A$ sufficiently regular, satisfaction of (i) can be guaranteed. Condition (ii) is the key one; unless a parameter

is both identifiable and stable, the least squares approach will not produce a reliable estimate.

4. Identification by regularization. Let us return to the general identification problem of § 2. In order to regularize the parameter λ, we introduce a more regular space $\mathcal{R}$, for which we assume:

(A8) $\mathcal{R}$ is a Hilbert space.

(A9) $\mathcal{R}$ is densely imbedded in Λ.

(A10) The imbedding operator from $\mathcal{R}$ into Λ is compact.

Define $\mathcal{R}_{ad} = \mathcal{R} \cap \Lambda_{ad}$. With (A5) and (A9) it readily follows that $\mathcal{R}_{ad}$ is a norm-closed convex subset of $\mathcal{R}$.

We now introduce the *stabilizing functional*

$$J_S(\lambda) = \|\lambda\|^2_{\mathcal{R}}, \qquad \lambda \in \mathcal{R}_{ad} \tag{4.1}$$

and the *smoothing functional*

$$J_\beta(\lambda) = J_{LS}(\lambda) + \beta J_S(\lambda), \qquad \lambda \in \mathcal{R}_{ad} \tag{4.2}$$

where $\beta > 0$ is the *regularization parameter.* Identification by regularization proceeds as follows. Given $z_d \in \mathcal{H}$ and $\beta > 0$, find $\lambda_\beta \in \mathcal{R}_{ad}$ so as to minimize $J_\beta(\lambda)$.

In this section we establish the basic results concerning the regularization method. The first result concerns differentiability of the functional $J_\beta(\lambda)$.

THEOREM 4.1. *Assume that* (A1)-(A4) *and* (A6)-(A9) *are valid. Then the functional*

$$J_\beta(\lambda) = \|\mathcal{C}(\Phi(A(\lambda))) - z_d\|^2_{\mathcal{H}} + \beta\|\lambda\|^2_{\mathcal{R}} \tag{4.3}$$

is of C^k-class. Its first derivative $J'_\beta(\lambda): \mathcal{R} \to \mathbb{R}$ is given by[2]

$$J'_\beta(\lambda) \cdot \delta\lambda = \left(\frac{\partial \Psi}{\partial A}(A(\lambda), u) \circ A'(\lambda) \cdot \delta\lambda, \rho\right)_{FF'} + 2\beta(\delta\lambda, \lambda)_{\mathcal{R}} \tag{4.4}$$

where u is the solution of $\Psi(A(\lambda), u) = f$ and ρ is the solution of

$$\left[\frac{\partial \Psi}{\partial u}(A(\lambda), u)\right]^* \rho = -2[\mathcal{C}'(u)]^* \Lambda_{\mathcal{H}}(\mathcal{C}(u) - z_d). \tag{4.5}$$

Proof. (A1)-(A4) and (A7) imply that $J_{LS}(\lambda) = \|\mathcal{C}(\Phi(A(\lambda))) - z_d\|^2_{\mathcal{H}}$ is of C^k-class with respect to the Λ-topology. Due to (A9), $J_{LS}(\lambda)$ will also be of C^k-class with respect to the $\mathcal{R}$-topology. Hence $J_\beta(\lambda)$ is also of C^k-class.

Existence and uniqueness of the solution of $\Psi(A(\lambda), u) = f$ is guaranteed by (A2) and (A6).

Existence and uniqueness of the solution of (4.5) follows from the following facts:

(a) $-2[\mathcal{C}'(u)]^* \Lambda_{\mathcal{H}}(\mathcal{C}(u) - z_d) \in U'$, since

$$(\mathcal{C}(u) - z_d) \in \mathcal{H}, \quad \Lambda_{\mathcal{H}}: \mathcal{H} \to \mathcal{H}', \quad [\mathcal{C}'(u)]^*: \mathcal{H}' \to U'.$$

(b) $[\partial\Psi/\partial u(A(\lambda), u)]^*$ is a linear homeomorphism of F' onto U', as a result of (A3).

Let us now calculate the first derivative of $J_\beta(\lambda)$. For every $\delta\lambda \in \Lambda$ we have

$$\begin{aligned} J'_\beta(\lambda) \cdot \delta\lambda &= 2(\mathcal{C}'(u) \cdot \delta u, \mathcal{C}(u) - z_d)_{\mathcal{H}} + 2\beta(\delta\lambda, \lambda)_{\mathcal{R}} \\ &= 2(\mathcal{C}'(u) \cdot \delta u, \Lambda_{\mathcal{H}}(\mathcal{C}(u) - z_d))_{\mathcal{H}\mathcal{H}'} + 2\beta(\delta\lambda, \lambda)_{\mathcal{R}} \\ &= 2(\delta u, [\mathcal{C}'(u)]^* \Lambda_{\mathcal{H}}(\mathcal{C}(u) - z_d))_{UU'} + 2\beta(\delta\lambda, \lambda)_{\mathcal{R}}. \end{aligned}$$

[2] Given a Banach space X and its dual X', we denote by $(\cdot, \cdot)_{XX'}$ the duality between X and X'. Given a Hilbert space $\mathcal{H}$, we denote by $(\cdot, \cdot)_{\mathcal{H}}$ the inner product in $\mathcal{H}$.

Taking into account (4.5) we have

$$J'_\beta(\lambda)\cdot\delta\lambda = -\left(\delta u, \left[\frac{\partial\Psi}{\partial u}(A,u)\right]^*\rho\right)_{UU'} + 2\beta(\delta\lambda,\lambda)_{\mathscr{R}}$$

$$= \left(-\frac{\partial\Psi}{\partial u}(A,u)\cdot\delta u, \rho\right)_{FF'} + 2\beta(\delta\lambda,\lambda)_{\mathscr{R}}.$$

Finally, from Proposition 2.2,

$$J'_\beta(\lambda)\cdot\delta\lambda = \left(\frac{\partial\Psi}{\partial A}(A,u)\circ A'(\lambda)\cdot\delta\lambda, \rho\right)_{FF'} + 2\beta(\delta\lambda,\lambda)_{\mathscr{R}}.$$

This completes the proof.

The next theorem establishes the existence of a global minimum of $J_\beta(\lambda)$ on $\mathscr{R}_{ad}$. We recall the following lemma.

LEMMA 4.1. *Let $\mathscr{R}$ and Λ be Banach spaces and assume that $\mathscr{R}$ is compactly imbedded in Λ. If $x_n \xrightarrow{\text{weak-top of } \mathscr{R}} x$, then $x_n \xrightarrow{\text{norm-top of } \Lambda} x$.*

THEOREM 4.2. *Under assumptions* (A1)-(A10), *the functional*

$$J_\beta(\lambda) = \|\mathscr{C}(\Phi(A(\lambda))) - z_d\|^2_{\mathscr{H}} + \beta\|\lambda\|^2_{\mathscr{R}}$$

admits a global minimum on $\mathscr{R}_{ad}$.

Proof. Let $m = \inf_{\lambda\in\mathscr{R}_{ad}} J_\beta(\lambda)$. Clearly, $m \geqq 0$. There is a minimizing sequence $\{\lambda_n\}$ in $\mathscr{R}_{ad}$ such that $\lim_{n\to\infty} J_\beta(\lambda_n) = m$. Clearly, we may assume that

$$\cdots \leqq J_\beta(\lambda_{n+1}) \leqq J_\beta(\lambda_n) \leqq \cdots \leqq J_\beta(\lambda_1).$$

Then, for every $n\in\mathbb{N}$

$$\|\lambda_n\|^2_{\mathscr{R}} \leqq \frac{1}{\beta}J_\beta(\lambda_n) \leqq \frac{1}{\beta}J_\beta(\lambda_1),$$

i.e., $\{\lambda_n\}$ is norm-bounded in $\mathscr{R}$. Hence, there is a subsequence $\{\lambda_{n_k}\}$ that converges in the weak topology of $\mathscr{R}$ to some $\lambda\in\mathscr{R}$. Since $\mathscr{R}_{ad}$ is norm-closed and convex, it is also weakly closed and hence $\lambda\in\mathscr{R}_{ad}$, so

$$\lambda_{n_k} \xrightarrow{\text{weak-top of } \mathscr{R}} \lambda \in \mathscr{R}_{ad}.$$

It follows from Lemma 4.1 that

$$\lambda_{n_k} \xrightarrow{\text{norm-top of } \Lambda} \lambda \in \mathscr{R}_{ad}.$$

Finally, using the continuity of the functional $J_{LS}(\lambda) = \|\mathscr{C}(\Phi(A(\lambda))) - z_d\|^2_{\mathscr{H}}$ in the norm-topology of Λ, as well as the weak lower semicontinuity of $J_S(\lambda) = \|\lambda\|^2_{\mathscr{R}}$ in $\mathscr{R}$, we conclude

$$m = \lim_{k\to\infty} J_\beta(\lambda_{n_k}) = \lim_{k\to\infty} \|\mathscr{C}(\Phi(A(\lambda_{n_k}))) - z_d\|^2_{\mathscr{H}} + \beta \lim_{k\to\infty} \|\lambda_{n_k}\|^2_{\mathscr{R}}$$

$$\geqq \left\|\mathscr{C}\left(\Phi\left(A\left(\lim_{k\to\infty}\lambda_{n_k}\right)\right)\right) - z_d\right\|^2_{\mathscr{H}} + \beta\left\|\lim_{k\to\infty}\lambda_{n_k}\right\|^2_{\mathscr{R}}$$

$$= J_\beta(\lambda).$$

Hence $J_\beta(\lambda) = m$. This completes the proof of Theorem 4.2.

Remark. We can say nothing about uniqueness of the minimum, since the functional $J_\beta(\lambda)$ will in general be nonconvex.

Now we can give a necessary condition for optimality:

PROPOSITION 4.1. *A necessary condition for* $\lambda \in \mathcal{R}_{ad}$ *to be a global minimum of* $J_\beta(\lambda)$ *on the set* $\mathcal{R}_{ad}$ *is*

$$J'_\beta(\lambda) \cdot (\nu - \lambda) \geqq 0 \quad \forall \nu \in \mathcal{R}_{ad}.$$

Proof. The proof is straightforward; see e.g. [20, p. 9] or [16].

So far we have established the existence of a minimum of the smoothing functional on $\mathcal{R}_{ad}$ and have given a necessary condition for optimality. Now we will show that minima of J_β depend continuously on the observation. This will be the key result of the regularization approach. Roughly speaking, what the next theorem says is the following:

Let $\tilde{\lambda}$ be the "true" value of the parameter and $\tilde{z}_d = \mathcal{C}(\Phi(A(\tilde{\lambda})))$, what we would have observed with a zero-error observation. Provided that

(i) $\tilde{\lambda}$ is the unique preimage of $\tilde{z}_d$;

(ii) β is an appropriately chosen function of the observation error;

any minimum of $J_\beta(\lambda)$ converges (in the norm of Λ) to $\tilde{\lambda}$, as the observation error tends (in the norm of $\mathcal{H}$) to zero.

Note that our theorem is a local version of the Tikhonov-Arsenin convergence theorem [29, p. 65] in the sense that:

(a) We relax their global identifiability assumption (i.e. in the sense of Definition 3.1).

(b) We refer to a specific pair $(\tilde{\lambda}, \tilde{z}_d)$ for which it is assumed that $\tilde{\lambda}$ is the unique preimage of $\tilde{z}_d$ in Λ_{ad} (pointwise identifiability assumption).

The need of such a generalization has been motivated by the fact that identifiability results are as a rule conditional identifiability results (see § 3). Note that the pointwise identifiability assumption (b) is the weakest possible assumption to ensure that the estimated parameter is "close enough" to the true parameter. (If $\tilde{z}_d$ has e.g. two preimages $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$, no mathematical method can "predict" which one is the true λ).

We first prove the following lemma:

LEMMA 4.2. *Let* $\langle X, d_X \rangle$, $\langle Y, d_Y \rangle$ *be metric spaces,* $f: X \to Y$ *a continuous mapping,* K *a precompact subset of* X. *Furthermore, we are given* $y^0 \in f(K)$ *to which there corresponds a unique* $x^0 \in X$ *with* $y^0 = f(x^0)$. *Then* $\forall \varepsilon > 0$, $\exists \gamma(\varepsilon) > 0$ *such that* $\forall x \in K\ d_Y(f(x), y^0) \leqq \gamma \Rightarrow d_X(x, x^0) \leqq \varepsilon$.

Proof. It suffices to prove that for every sequence $\{x_n\}$ in K such that $f(x_n) \to y^0$ we have $x_n \to x^0$., Since K is precompact, $\{x_n\}$ has a subsequence $\{x_{n_k}\}$ that converges to some $\tilde{x} \in X$. Since f is continuous, $f(x_{n_k}) \to f(\tilde{x})$. But $f(x_n) \to y^0$. Hence, $f(\tilde{x}) = y^0$. And since x^0 is the unique preimage of y^0, $\tilde{x} = x^0$. So, $x_{n_k} \to x^0$.

The same argument shows that $\{x_n\}$ cannot have any cluster point other than x^0. Thus x^0 is the unique cluster point of $\{x_n\}$, which is contained in the precompact set K. Hence [1, p. 68] $x_n \to x^0$. This completes the proof.

THEOREM 4.3. *For any* $\beta > 0$ *and* $z_d \in \mathcal{H}$, *denote by* $\lambda_\beta \in \mathcal{R}_{ad}$ *any minimum of* $J_\beta(\lambda)$ *on* $\mathcal{R}_{ad}$. *Also, denote by* T_{δ_1} *the class of functions that are nonnegative, nondecreasing and continuous on the interval* $[0, \delta_1]$. *Suppose*

$$\tilde{z}_d \in \mathcal{H},$$

$$\exists \text{ a unique } \tilde{\lambda} \in \mathcal{R}_{ad} \text{ with } \tilde{z}_d = \mathcal{C}(\Phi(A(\tilde{\lambda}))).$$

Then $\forall \varepsilon > 0\ \forall B_1, B_2 \in T_{\delta_1}$ *with*

$$B_2(0) = 0, \qquad \frac{\delta^2}{B_1(\delta)} \leqq B_2(\delta),$$

$\exists \delta_0(\varepsilon, B_1, B_2) \leqq \delta_1$ *such that* $\forall z_d \in \mathscr{H}\ \forall \delta \leqq \delta_0$

$$\|z_d - \tilde{z}_d\|_{\mathscr{H}} \leqq \delta \Rightarrow \|\lambda_\beta - \tilde{\lambda}\|_\Lambda \leqq \varepsilon,$$

for all β *satisfying* $\delta^2 / B_1(\delta) \leqq \beta \leqq B_2(\delta)$.

Proof. We have

$$\begin{aligned}
\beta \|\lambda_\beta\|_{\mathscr{R}}^2 &\leqq \|\mathscr{C}(\Phi(A(\lambda_\beta))) - z_d\|_{\mathscr{H}}^2 + \beta \|\lambda_\beta\|_{\mathscr{R}}^2 \\
&\leqq \|\mathscr{C}(\Phi(A(\tilde{\lambda}))) - z_d\|_{\mathscr{H}}^2 + \beta \|\tilde{\lambda}\|_{\mathscr{R}}^2 \\
&= \|\tilde{z}_d - z_d\|_{\mathscr{H}}^2 + \beta \|\tilde{\lambda}\|_{\mathscr{R}}^2 \leqq \delta^2 + \beta \|\tilde{\lambda}\|_{\mathscr{R}}^2 \\
&= \beta \left[\frac{\delta^2}{\beta} + \|\tilde{\lambda}\|_{\mathscr{R}}^2 \right] \leqq \beta [B_1(\delta) + \|\tilde{\lambda}\|_{\mathscr{R}}^2] \leqq \beta [B_1(\delta_1) + \|\tilde{\lambda}\|_{\mathscr{R}}^2].
\end{aligned}$$

Denote $H_0 = [B_1(\delta_1) + \|\tilde{\lambda}\|_{\mathscr{R}}^2]^{1/2}$. Clearly, $\|\lambda_\beta\|_{\mathscr{R}} \leqq H_0$ and $\|\tilde{\lambda}\|_{\mathscr{R}} \leqq H_0$. Thus we have shown that the elements $\tilde{\lambda}$ and λ_β belong to the set

$$\Lambda_{H_0} = \{\lambda \in \mathscr{R}_{ad} \mid \|\lambda\|_{\mathscr{R}} \leqq H_0\},$$

which is precompact in the norm-topology of Λ. It follows from Lemma 4.2 that

$\forall \varepsilon > 0\ \exists \gamma(\varepsilon) > 0$ such that $\forall \hat{\lambda} \in \Lambda_{H_0}$

$$\|\mathscr{C}(\Phi(A(\hat{\lambda}))) - \tilde{z}_d\|_{\mathscr{H}} \leqq \gamma \Rightarrow \|\hat{\lambda} - \tilde{\lambda}\|_\Lambda \leqq \varepsilon.$$

Now observe that

$$\begin{aligned}
\|\mathscr{C}(\Phi(A(\lambda_\beta))) - z_d\|_{\mathscr{H}}^2 &\leqq \|\mathscr{C}(\Phi(A(\lambda_\beta))) - z_d\|_{\mathscr{H}}^2 + \beta \|\lambda_\beta\|_{\mathscr{R}}^2 \\
&\leqq \|\mathscr{C}(\Phi(A(\tilde{\lambda}))) - z_d\|_{\mathscr{H}}^2 + \beta \|\tilde{\lambda}\|_{\mathscr{R}}^2 \\
&= \|\tilde{z}_d - z_d\|_{\mathscr{H}}^2 + \beta \|\tilde{\lambda}\|_{\mathscr{R}}^2 \\
&\leqq \delta^2 + B_2(\delta) \|\tilde{\lambda}\|_{\mathscr{R}}^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
\|\mathscr{C}(\Phi(A(\lambda_\beta))) - \tilde{z}_d\|_{\mathscr{H}} &\leqq \|\mathscr{C}(\Phi(A(\lambda_\beta))) - z_d\|_{\mathscr{H}} + \|z_d - \tilde{z}_d\|_{\mathscr{H}} \\
&\leqq (\delta^2 + B_2(\delta) \|\tilde{\lambda}\|_{\mathscr{R}}^2)^{1/2} + \delta.
\end{aligned}$$

The function $\psi(\delta) = (\delta^2 + B_2(\delta) \|\tilde{\lambda}\|_{\mathscr{R}}^2)^{1/2} + \delta$ is a continuous monotonically increasing function satisfying $\psi(0) = 0$. Hence, one can choose $\delta_0 = \psi^{-1}(\gamma(\varepsilon))$ and have $\|\mathscr{C}(\Phi(A(\lambda_\beta))) - \tilde{z}_d\|_{\mathscr{H}} \leqq \gamma(\varepsilon)\ \forall \delta \leqq \delta_0$. Thus we see that for all β satisfying $\delta^2 / B_1(\delta) \leqq \beta \leqq B_2(\delta)$, the inequality $\|z_d - \tilde{z}_d\|_{\mathscr{H}} \leqq \delta$ implies the inequality $\|\lambda_\beta - \tilde{\lambda}\|_\Lambda \leqq \varepsilon$. This completes the proof.

5. Selection of the regularization parameter. In § 4 we established that the regularization approach provides a stable method for distributed system identification. One question was not addressed, the selection of the regularization parameter β. In this section we will discuss two methods for selection of β.

Let $\tilde{\lambda} \in \mathscr{R}_{ad}$ be the "true" value of the parameter λ and $\tilde{z}_d$ be the error-free observation, $\tilde{z}_d = \mathscr{C}(\Phi(A(\tilde{\lambda})))$. We assume that (i) $\tilde{\lambda}$ is the unique preimage of $\tilde{z}_d$; (ii) An upper bound in the observation error is known, i.e. $\|z_d - \tilde{z}_d\|_{\mathscr{H}} \leqq \delta$.

Method 1. When an a priori upper bound on $\|\tilde{\lambda}\|_{\mathscr{R}}$ is known, i.e. $\|\tilde{\lambda}\|_{\mathscr{R}} \leqq \Delta$, Miller [22] suggests $\beta(\delta) = (\delta/\Delta)^2$. (When $\mathscr{R}$ is a Sobolev space, $\|\cdot\|_{\mathscr{R}}$ is a measure of smoothness.) We note first that this choice of β satisfies the assumptions of Theorem

4.3. Furthermore, if $\lambda_{\beta(\delta)}$ is a minimizer of

$$J_\beta(\lambda) = \| \mathscr{C}(\Phi(A(\lambda))) - z_d \|_{\mathscr{H}}^2 + \left(\frac{\delta}{\Delta}\right)^2 \| \lambda \|_{\mathscr{R}}^2 \tag{5.1}$$

on $\mathscr{R}_{ad}$, then

$$\| \mathscr{C}(\Phi(A(\lambda_{\beta(\delta)}))) - z_d \|_{\mathscr{H}}^2 + \left(\frac{\delta}{\Delta}\right)^2 \| \lambda_{\beta(\delta)} \|_{\mathscr{R}}^2$$

$$= J_\beta(\lambda_{\beta(\delta)}) \leqq J_\beta(\tilde{\lambda}) = \| \mathscr{C}(\Phi(A(\tilde{\lambda}))) - z_d \|_{\mathscr{H}}^2 + \left(\frac{\delta}{\Delta}\right)^2 \| \tilde{\lambda} \|_{\mathscr{R}}^2 \leqq 2\delta^2.$$

Hence,

$$\| \mathscr{C}(\Phi(A(\lambda_{\beta(\delta)}))) - z_d \|_{\mathscr{H}} \leqq \sqrt{2}\delta, \qquad \| \lambda_{\beta(\delta)} \|_{\mathscr{R}} \leqq \sqrt{2}\Delta \tag{5.2}$$

i.e. regularized solutions satisfy the constraints up to a factor of $\sqrt{2}$.

Method 2. This method has been suggested by Tikhonov and Arsenin [29]. Their suggestion is to choose $\beta(\delta)$ so that

$$\| \mathscr{C}(\Phi(A(\lambda_{\beta(\delta)}))) - z_d \|_{\mathscr{H}} = \delta$$

where $\lambda_{\beta(\delta)}$ minimizes

$$J_\beta(\lambda) = \| \mathscr{C}(\Phi(A(\lambda))) - z_d \|_{\mathscr{H}}^2 + \beta(\delta) \| \lambda \|_{\mathscr{R}}^2.$$

Before one discusses the stability of the method, one has to examine the existence of such a β.

To this end, we follow a different approach than Tikhonov and Arsenin, who give a simple sufficient condition for existence. We give here a much weaker condition which is both necessary and sufficient. In the development of this condition we have used concepts and results from the theory of minimization of vector-valued functionals.

DEFINITION 5.1 (ordering relations in $\mathbb{R}^n$). Let $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ be two arbitrary elements of $\mathbb{R}^n$. We will write

(i) $x \leqq y$ iff $x_i \leqq y_i$ for all i;

(ii) $x \lneqq y$ iff $x \leqq y$ and $x_i < y_i$ for at least one i;

(iii) $x < y$ iff $x_i < y_i$ for all i.

DEFINITION 5.2. Let $\Upsilon : \Lambda \to \mathbb{R}^n$. We will say that $\hat{\lambda}$ is a *Pareto-minimal point* of the vector-valued functional Υ if $\nexists \lambda \in \Lambda$ with $\Upsilon(\lambda) \lneqq \Upsilon(\hat{\lambda})$. The set

$$\{\Upsilon(\hat{\lambda}) | \hat{\lambda} \text{ is a Pareto-minimal point of } \Upsilon\}$$

is called the *Pareto-minimal set* of Υ.

PROPOSITION 5.1 [30, p. 94]. *Let* $\Upsilon(\lambda) \equiv (\Upsilon_1(\lambda), \cdots, \Upsilon_n(\lambda))$ *be a vector-valued functional on* Λ. *An element* $\hat{\lambda} \in \Lambda$ *is a Pareto-minimal point iff for every* $j \in \{1, \cdots, n\}$ $\hat{\lambda}$ *minimizes* $\Upsilon_j(\lambda)$ *on the set*

$$\Lambda_j = \{\lambda \in \Lambda | \Upsilon_i(\lambda) \leqq \Upsilon_i(\tilde{\lambda}) \ \forall i \in \{1, \cdots, n\} \text{ with } i \neq j\}.$$

LEMMA 5.1. *Let*

$$\lambda_{\min} = \text{the minimum-norm element of } \mathscr{R}_{ad},$$

$$\delta_{\max} = J_{LS}(\lambda_{\min}) = \| \mathscr{C}(\Phi(A(\lambda_{\min}))) - z_d \|_{\mathscr{H}}^2,$$

$$\delta_{\min} = \inf_{\lambda \in \mathscr{R}_{ad}} J_{LS}(\lambda) = \inf_{\lambda \in \mathscr{R}_{ad}} \| \mathscr{C}(\Phi(A(\lambda))) - z_d \|_{\mathscr{H}}^2.$$

Given $\delta > \delta_{\min}$, *there exists an element* λ_δ *minimizing the functional* $J_S(\lambda)$ *on the set*

$\{\lambda \in \mathcal{R}_{ad} \,|\, \|\mathcal{C}(\Phi(A(\lambda))) - z_d\|_{\mathcal{H}}^2 \leqq \delta\}$. *Furthermore, if* $\delta \leqq \delta_{\max}$, *then*

$$\|\mathcal{C}(\Phi(A(\lambda_\delta))) - z_d\|_{\mathcal{H}}^2 = \delta.$$

Remark. Lemma 5.1 holds for $\delta = \delta_{\min}$ if $J_{LS}(\lambda)$ admits a minimum on $\mathcal{R}_{ad}$.

THEOREM 5.1. *Let* λ_δ, $\delta_{\max}$ *as in Lemma* 5.1. *The function*

$$\Theta(\delta) = J_S(\lambda_\delta), \qquad \delta \leqq \delta_{\max}$$

is monotonically decreasing. Its graph coincides with the Pareto-minimal set of the vector-valued functional $\Upsilon : \mathcal{R}_{ad} \to \mathbb{R}^2$ *defined by* $\Upsilon(\lambda) = (J_{LS}(\lambda), J_S(\lambda))$.

Proof of Lemma 5.1. The proof of existence of a minimum is almost the same as that of Theorem 4.2.

Consider a minimizing sequence $\{\lambda_n\}$. This will have a subsequence $\{\lambda_{n_k}\}$ that converges in the weak topology of $\mathcal{R}$ to some $\lambda \in \mathcal{R}$. We conclude that

$$\lambda \in \mathcal{R}_{ad}, \qquad \lambda_{n_k} \xrightarrow{\text{strong topology of } \Lambda} \lambda.$$

Also, due to the continuity of $\mathcal{C} \circ \Phi \circ A$, it is easy to see that the limit has to satisfy

$$\|\mathcal{C}(\Phi(A(\lambda))) - z_d\|_{\mathcal{H}}^2 \leqq \delta.$$

Finally, using the weak lower semicontinuity of $J_S(\lambda)$ in $\mathcal{R}$, we conclude that λ minimizes $J_S(\lambda)$ on the set $\{\lambda \in \mathcal{R}_{ad} \,|\, \|\mathcal{C}(\Phi(A(\lambda))) - z_d\|_{\mathcal{H}}^2 \leqq \delta\}$.

To prove the second part of the lemma, suppose

$$\|\mathcal{C}(\Phi(A(\lambda_\delta))) - z_d\|_{\mathcal{H}}^2 = \delta^* < \delta.$$

Since $\mathcal{C} \circ \Phi \circ A$ is continuous, there is a ball $B(\lambda_\delta)$, centered at λ_δ, such that

$$\|\mathcal{C}(\Phi(A(\lambda))) - \mathcal{C}(\Phi(A(\lambda_\delta)))\|_{\mathcal{H}} < \frac{\sqrt{\delta} - \sqrt{\delta^*}}{2} \quad \forall \lambda \in B(\lambda_\delta) \cap \mathcal{R}_{ad}.$$

Now observe that

(i) We can always have $\lambda_{\min} \notin B(\lambda_\delta)$, since $\delta^* < \delta_{\max}$ implies $\lambda_\delta \neq \lambda_{\min}$.

(ii) $B(\lambda_\delta) \cap \mathcal{R}_{ad} \subset \{\lambda \in \mathcal{R}_{ad} \,|\, \|\mathcal{C}(\Phi(A(\lambda))) - z_d\|_{\mathcal{H}}^2 \leqq \delta\}$ since

$$\|\mathcal{C}(\Phi(A(\lambda))) - z_d\|_{\mathcal{H}} \leqq \|\mathcal{C}(\Phi(A(\lambda))) - \mathcal{C}(\Phi(A(\lambda_\delta)))\|_{\mathcal{H}} + \|\mathcal{C}(\Phi(A(\lambda_\delta))) - z_d\|_{\mathcal{H}}.$$

But from (i) and convexity of $\mathcal{R}_{ad}$ it follows that $\exists \lambda^* \in B(\lambda_\delta) \cap \mathcal{R}_{ad}$ so that $\|\lambda^*\|_{\mathcal{R}} < \|\lambda_\delta\|_{\mathcal{R}}$.

This contradicts the definition of λ_δ and (ii).

Proof of Theorem 5.1. If λ_δ minimizes $J_S(\lambda)$ subject to the constraint $J_{LS}(\lambda) \leqq \delta < \delta_{\max}$, then from Lemma 5.1

$$\Upsilon(\lambda_\delta) = (\delta, \Theta(\delta)).$$

It is clear that $\delta_1 \leqq \delta_2 \leqq \delta_{\max}$ implies $\Theta(\delta_2) \leqq \Theta(\delta_1)$ i.e. Θ is monotonically decreasing. Furthermore, Proposition 5.1 implies that the Pareto-minimal set of Υ is a subset of the graph of Θ.

Finally, if $\Upsilon(\hat{\lambda}) \gneqq \Upsilon(\lambda_\delta)$ for some $\hat{\lambda} \in \mathcal{R}_{ad}$ and $\delta \leqq \delta_{\max}$, this would mean

either

$$J_{LS}(\hat{\lambda}) < \delta, \qquad J_S(\hat{\lambda}) \leqq J_S(\lambda_\delta)$$

or

$$J_{LS}(\hat{\lambda}) \leqq \delta, \qquad J_S(\hat{\lambda}) < J_S(\lambda_\delta).$$

Both cases are impossible since they contradict the definition of λ_δ and/or Lemma 5.1. Hence, $\Upsilon(\lambda_\delta) \equiv (\delta, \Theta(\delta))$ is an element of the Pareto-minimal set.

This completes the proof of the theorem.

Before we proceed to the main result of this section, we will state an important proposition by Yu [33], which will be needed in the proof. We first give the definition of cone convexity, introduced in the same paper.

DEFINITION 5.3. Let $S \subset \mathbb{R}^n$ and C a convex cone in $\mathbb{R}^n$. S will be called *C-convex* if $S + C$ is convex.

PROPOSITION 5.2. [33, p. 28]. *Let* $\Upsilon : \Lambda \to \mathbb{R}^n$ *and suppose that* Ran Υ *is* $\mathbb{R}^n_+$*-convex, where* $\mathbb{R}^n_+ = \{x \in \mathbb{R}^n \mid x \geqq 0\}$. *If* $\hat{\lambda}$ *is a Pareto-minimal point, there exists* $\eta \leqq 0$ *such that*

$$\eta^T \Upsilon(\hat{\lambda}) \leqq \eta^T \Upsilon(\lambda) \quad \forall \lambda \in \Lambda.$$

THEOREM 5.2. *Let* Θ *and* Υ *be as in Theorem* 5.1. *Then the following assertions are equivalent*:

(i) *For every* $\delta \in]\delta_{\min}, \delta_{\max}[$ *there is* $\beta > 0$ *and a minimizer* λ_β *of the functional* $J_\beta(\lambda) = J_{LS}(\lambda) + \beta J_S(\lambda)$ *on* $\mathcal{R}_{ad}$, *such that* $J_{LS}(\lambda_\beta) = \delta$.

(ii) Ran Υ *is* $\mathbb{R}^2_+$*-convex.*

(iii) Θ *is convex.*

We first prove the following lemma:

LEMMA 5.2. *Let*

$$\tilde{\Theta}(\delta) = \begin{cases} \Theta(\delta), & \text{for } \delta \leqq \delta_{\max}, \\ J_S(\lambda_{\min}), & \text{for } \delta \geqq \delta_{\max}. \end{cases}$$

Then[3] Ran $\Upsilon + \mathbb{R}^2_+ = \text{Epi } \tilde{\Theta}$.

Proof of Lemma 5.2. Take $(x, y) \in \text{Epi } \tilde{\Theta}$. If $x \geqq \delta_{\max}$, then $(x, y) \in \{\Upsilon(\lambda_{\min})\} + \mathbb{R}^2_+$. If $x \leqq \delta_{\max}$, then $(x, y) \in \{\Upsilon(\lambda_x)\} + \mathbb{R}^2_+$, where λ_x minimizes $J_S(\lambda)$ on $\mathcal{R}_{ad}$ subject to $J_{LS}(\lambda) \leqq x$. So: Ran $\Upsilon + \mathbb{R}^2_+ \supseteq \text{Epi } \tilde{\Theta}$.

Now take $z \in \text{Ran } \Upsilon + \mathbb{R}^2_+$. This means $\exists \lambda \in \mathcal{R}_{ad}\ \exists \nu \in \mathbb{R}^2_+$ with $z = \Upsilon(\lambda) + \nu$. Assume that $z \notin \text{Epi } \tilde{\Theta}$, hence $\exists \delta \geqq 0$ with $z \geqq (\delta, \Theta(\delta))$.

Case 1. $\delta \leqq \delta_{\max}$. Then $\Upsilon(\lambda) + \nu \geqq \Upsilon(\lambda_\delta)$, where λ_δ minimizes $J_S(\lambda)$ on $\mathcal{R}_{ad}$ subject to $J_{LS}(\lambda) \leqq \delta$. This implies that λ_δ is not Pareto-minimal.

Case 2. $\delta \geqq \delta_{\max}$. Then $\Upsilon(\lambda) + \nu \geqq \Upsilon(\lambda_{\min}) \Rightarrow \lambda_{\min}$ is not Pareto-minimal.

Thus, we see that in all cases the assumption $z \notin \text{Epi } \tilde{\Theta}$ leads to contradiction. Hence, Ran $\Upsilon + \mathbb{R}^2_+ \subseteq \text{Epi } \tilde{\Theta}$. This completes the proof.

Proof of Theorem 5.2.

(i) $\Rightarrow$ (ii). Given $\beta > 0$ denote by λ_β a minimizer of $J_\beta(\lambda) = J_{LS}(\lambda) + \beta J_S(\lambda)$ on $\mathcal{R}_{ad}$ and define

$$P_\beta = \{(x, y) \in \mathbb{R}^2 \mid x + \beta y \geqq J_{LS}(\lambda_\beta) + \beta J_S(\lambda_\beta)\},$$

$$P_\infty = \{(x, y) \in \mathbb{R}^2 \mid y \geqq J_S(\lambda_{\min})\}.$$

Furthermore, define

$$P_0 = \begin{cases} \{(x, y) \in \mathbb{R}^2 \mid x > \delta_{\min}\} \cup \{(x, y) \in \mathbb{R}^2 \mid x = \delta_{\min}, y \geqq J_S(\bar{\lambda})\} \\ \qquad\qquad \text{if } \bar{\lambda} \text{ minimizes } J_{LS}(\lambda) \text{ on } \mathcal{R}_{ad}, \\ \{(x, y) \in \mathbb{R}^2 \mid x > \delta_{\min}\} \quad \text{if } J_{LS}(\lambda) \text{ does not admit a minimum of } \mathcal{R}_{ad}. \end{cases}$$

We will show that $\bigcap_{0 \leqq \beta \leqq \infty} P_\beta = \text{Ran } \Upsilon + \mathbb{R}^2_+$. Clearly $\forall \lambda \in \mathcal{R}_{\text{ad}}\ \forall \beta > 0\ \Upsilon(\lambda) \in P_\beta$. Hence $\forall \lambda \in \mathcal{R}_{ad}\ \forall z \in \mathbb{R}^2_+\ \forall \beta > 0\ \Upsilon(\lambda) + z \in P_\beta$. Thus Ran $\Upsilon + \mathbb{R}^2_+ \subseteq P_\beta\ \forall \beta > 0$. It is also trivial to see that the above relation holds for $\beta = 0$ and $\beta = \infty$. Hence Ran $\Upsilon + \mathbb{R}^2_+ \subseteq \bigcap_{0 \leqq \beta \leqq \infty} P_\beta$.

[3] By the symbol Epi F we mean the epigraph of a function F, i.e. the set $\{(x, y) \in \mathbb{R}^2 \mid y \geqq F(x)\}$.

To show that $\bigcap_{0\leqq\beta\leqq\infty} P_\beta \subseteq \operatorname{Ran}\Upsilon+\mathbb{R}^2_+$ we will take $(\hat{x},\hat{y})\notin \operatorname{Ran}\Upsilon+\mathbb{R}^2_+$ and show that $\exists\beta$ such that $(\hat{x},\hat{y})\notin P_\beta$. We only need to consider the case $\hat{x}\in]\delta_{\min},\delta_{\max}[$, since $\hat{x}=\delta_{\min}$ clearly implies $(\hat{x},\hat{y})\notin P_0$ and $\hat{x}\geqq\delta_{\max}$ implies $(\hat{x},\hat{y})\notin P_\infty$. By (i), $\exists\hat{\beta}>0$ $\exists$ minimizer $\lambda_{\hat{\beta}}$ of the functional $J_{\hat{\beta}}(\lambda)=J_{LS}(\lambda)+\hat{\beta}J_S(\lambda)$ on $\mathscr{R}_{\text{ad}}$ satisfying $J_{LS}(\lambda_{\hat{\beta}})=\hat{x}$. Observe that $J_S(\lambda_{\hat{\beta}})>\hat{y}$, since otherwise $\Upsilon(\lambda_{\hat{\beta}})\leqq(\hat{x},\hat{y})$ which would imply that $(\hat{x},\hat{y})\in \operatorname{Ran}\Upsilon+\mathbb{R}^2_+$. But

$$\left.\begin{aligned} J_{LS}(\lambda_{\hat{\beta}})&=\hat{x}\\ J_S(\lambda_{\hat{\beta}})&>\hat{y}\end{aligned}\right\}\Rightarrow J_{LS}(\lambda_{\hat{\beta}})+\hat{\beta}J_S(\lambda_{\hat{\beta}})>\hat{x}+\hat{\beta}\hat{y}\Rightarrow(\hat{x},\hat{y})\notin P_{\hat{\beta}}.$$

So

$$\bigcap_{0\leqq\beta\leqq\infty} P_\beta=\operatorname{Ran}\Upsilon+\mathbb{R}^2_+.$$

Taking into account the convexity of the sets P_β we conclude that $\operatorname{Ran}\Upsilon+\mathbb{R}^2_+$ is convex.

(ii)$\Rightarrow$(i). Consider an arbitrary $\delta\in]\delta_{\min},\delta_{\max}[$ and denote by λ_δ a minimizer of $J_S(\lambda)$ on $\mathscr{R}_{ad}$ subject to the constraint $J_{LS}(\lambda)\leqq\delta$.

Since $\operatorname{Ran}\Upsilon$ is $\mathbb{R}^2_+$-convex, by Proposition 5.2, there is $\eta\equiv(\eta_1,\eta_2)\gneqq 0$ such that:

$$\eta_1 J_{LS}(\lambda_\delta)+\eta_2 J_S(\lambda_\delta)\leqq\eta_1 J_{LS}(\lambda)+\eta_2 J_S(\lambda)\ \forall\lambda\in\mathscr{R}_{ad}.$$

Since $\delta>\delta_{\min}$, λ_δ does not minimize J_{LS} on $\mathscr{R}_{ad}$. Hence $\eta_2\neq 0$.

Since $\delta<\delta_{\max}$, we have $\lambda_\delta\neq\lambda_{\min}$. Hence $\eta_1\neq 0$.

So we may choose $\beta=\eta_2/\eta_1>0$ and have

$$J_{LS}(\lambda_\delta)+\beta J_S(\lambda_\delta)\leqq J_{LS}(\lambda)+\beta J_S(\lambda)\quad\forall\lambda\in\mathscr{R}_{ad}.$$

But by construction of λ_δ and Lemma 5.1 it follows that $J_{LS}(\lambda_\delta)=\delta$.

(ii)$\Leftrightarrow$(iii).

$$\operatorname{Ran}\Upsilon \text{ is } \mathbb{R}^n_+\text{-convex}\overset{\text{Lemma 5.2}}{\Longleftrightarrow}\operatorname{Epi}\tilde{\Theta}\text{ is convex}\Leftrightarrow\tilde{\Theta}\text{ is convex.}$$

Since Θ is decreasing, the latter is equivalent to Θ convex. This completes the proof of the theorem.

What remains to show is that regularized solutions obtained by this method converge (in the norm of Λ) to $\tilde{\lambda}$ as the observation error tends (in the norm of $\mathscr{H}$) to zero. This will be done independently of the theory of § 4. Note, however, that the argument is almost the same as in Theorem 4.3.

THEOREM 5.3. *Suppose*

$\tilde{z}_d\in\mathscr{H}$,

$\exists$ *a unique* $\tilde{\lambda}\in\mathscr{R}_{ad}$ *with* $\tilde{z}_d=\mathscr{C}(\Phi(A(\tilde{\lambda})))$,

the function Θ *defined in Theorem* 5.1 *is convex.*

Then $\forall\varepsilon>0\ \exists\delta_0(\varepsilon)>0$ *such that* $\forall z_d\in\mathscr{H}\ \forall\delta\leqq\delta_0$

$$\|z_d-\tilde{z}_d\|_{\mathscr{H}}\leqq\delta\Rightarrow\|\lambda_{\beta(\delta)}-\tilde{\lambda}\|_\Lambda\leqq\varepsilon,$$

where

$\beta(\delta)$ *denotes a regularization parameter,*

$\lambda_{\beta(\delta)}$ *denotes a minimizer of* $J_{\beta(\delta)}(\lambda)$ *on* $\mathscr{R}_{ad}$,

satisfying $\|\mathscr{C}(\Phi(A(\lambda_{\beta(\delta)})))-z_d\|_{\mathscr{H}}=\delta$.

Proof. Denote $\hat{\Lambda} = \{\lambda \in \mathcal{R}_{ad} \mid \|\lambda\|_{\mathcal{R}} \leqq \|\tilde{\lambda}\|_{\mathcal{R}}\}$ which is precompact in the norm-topology of Λ. It follows from Lemma 4.2 that

$\forall \varepsilon > 0\ \exists \gamma(\varepsilon) > 0$ such that $\forall \hat{\lambda} \in \hat{\Lambda}$

$$\|\mathcal{C}(\Phi(A(\hat{\lambda}))) - \tilde{z}_d\|_{\mathcal{H}} \leqq \gamma \Rightarrow \|\hat{\lambda} - \tilde{\lambda}\|_{\Lambda} \leqq \varepsilon.$$

Clearly, $\lambda_{\beta(\delta)}$ minimizes $J_S(\lambda)$ on $\mathcal{R}_{ad}$ subject to the constraint $\|\mathcal{C}(\Phi(A(\lambda))) - z_d\|_{\mathcal{H}} \leqq \delta$. Since

$$\|\mathcal{C}(\Phi(A(\tilde{\lambda}))) - z_d\|_{\mathcal{H}} = \|\tilde{z}_d - z_d\|_{\mathcal{H}} \leqq \delta$$

it is obvious that $J_S(\lambda_{\beta(\delta)}) \leqq J_S(\tilde{\lambda})$ i.e. $\lambda_{\beta(\delta)} \in \hat{\Lambda}$. Also, observe that

$$\|\mathcal{C}(\Phi(A(\lambda_{\beta(\delta)}))) - \tilde{z}_d\|_{\mathcal{H}} \leqq \|\mathcal{C}(\Phi(A(\lambda_{\beta(\delta)}))) - z_d\|_{\mathcal{H}} + \|z_d - \tilde{z}_d\|_{\mathcal{H}} \leqq 2\delta.$$

Thus we can choose $\delta_0 = \gamma(\varepsilon)/2$ and have

$$\|\mathcal{C}(\Phi(A(\lambda_{\beta(\delta)}))) - \tilde{z}_d\|_{\mathcal{H}} \leqq \gamma(\varepsilon) \quad \forall \delta \leqq \delta_0.$$

Thus we see that for all $\delta \leqq \delta_0 = \gamma(\varepsilon)/2$, the inequality $\|z_d - \tilde{z}_d\|_{\mathcal{H}} \leqq \delta$ implies the inequality $\|\lambda_{\beta(\delta)} - \tilde{\lambda}\|_{\Lambda} \leqq \varepsilon$.

This completes the proof.

Before we close this section, we should say a few words concerning the practical implementation of the two methods of selection of the regularization parameter presented previously. Method 1 provides a concrete formula for $\beta(\delta)$ given an upper bound Δ on the smoothness of the unknown parameter. Thus, one can obtain an approximate solution of the identification problem by numerical minimization of the corresponding smoothing functional (see § 9). In Method 2, $\beta(\delta)$ is determined implicitly as a solution of the equation

$$\|\mathcal{C}(\Phi(A(\lambda_\beta))) - z_d\|_{\mathcal{H}} = \delta. \tag{5.3}$$

Therefore, in order to find β (in terms of δ) in practice, one will have to numerically solve the above equation; this will involve a *sequence* of numerical minimizations of $J_\beta(\lambda)$ (each corresponding to the value of β at each iteration).

For instance, one can use the interval halving algorithm. Convergence to a root of (5.3) is an immediate consequence of the following monotonicity and lower- and upper- semicontinuity properties:

PROPOSITION 5.3. *If $\beta_1 < \beta_2$, then $J_{LS}(\lambda_{\beta_1}) < J_{LS}(\lambda_{\beta_2})$ for all minimizers λ_{β_1} of $J_{\beta_1}(\lambda) = J_{LS}(\lambda) + \beta_1 J_S(\lambda)$ and all minimizers λ_{β_2} of $J_{\beta_2}(\lambda) = J_{LS}(\lambda) + \beta_2 J_S(\lambda)$.*

PROPOSITION 5.4. *Let $\beta^* > 0$ and M_{β^*} be the set of minimizers λ_{β^*} of $J_{\beta^*}(\lambda) = J_{LS}(\lambda) + \beta^* J_S(\lambda)$.*

(a) *If $\{\beta_n\}$ is an increasing sequence of positive numbers converging to β^* and $\{\lambda_{\beta_n}\}$ is a sequence of minimizers of $J_{\beta_n}(\lambda) = J_{LS}(\lambda) + \beta_n J_S(\lambda)$, then $J_{LS}(\lambda_{\beta_n}) \to \sup_{\lambda_{\beta^*} \in M_{\beta^*}} J_{LS}(\lambda_{\beta^*})$.*

(b) *If $\{\beta_n\}$ is a decreasing sequence converging to β^* and $\{\lambda_{\beta_n}\}$ is a sequence of minimizers of $J_{\beta_n}(\lambda) = J_{LS}(\lambda) + \beta_n J_S(\lambda)$, then $J_{LS}(\lambda_{\beta_n}) \to \inf_{\lambda_{\beta^*} \in M_{\beta^*}} J_{LS}(\lambda_{\beta^*})$.*

The proof of these propositions is omitted for brevity.

6. Identification of a second order linear parabolic system from distributed observation. Let

Ω a bounded open subset of $\mathbb{R}^n$,

Γ the boundary of Ω, a C^1-manifold with Ω locally on one side of Γ,

T a real number with $0 < T < \infty$,

$Q = \Omega \times]0, T[$,

$\Sigma = \Gamma \times]0, T[$,

and consider

$$\begin{aligned} &\frac{\partial u}{\partial t} - \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{jk}(x)\frac{\partial u}{\partial x_k}\right) + a_0(x)u = f(x,t) \quad \text{in } Q, \\ &u(x,0) = u_0(x) \quad \text{in } \Omega \\ &\frac{\partial u}{\partial \nu} = 0 \quad \text{on } \Sigma. \end{aligned} \tag{6.1}$$

The variational formulation of the above Neumann problem is as follows

$$\begin{aligned} &\int_\Omega \frac{\partial u}{\partial t} v + \int_\Omega \sum_{j,k=1}^{n} a_{jk}(x)\frac{\partial u}{\partial x_k}\frac{\partial v}{\partial x_j} + \int_\Omega a_0(x)uv = \int_\Omega fv \quad \forall v \in V, \\ &u(x,0) = u_0(x) \end{aligned} \tag{6.2}$$

where $V = H^1(\Omega)$.

We assume that the parameters a_{jk}, $a_0 \in C^0(\bar{\Omega})$, i.e. the parameter space

$$\Lambda = \left(\bigtimes_{j,k=1}^{n} C^0(\bar{\Omega})\right) \times C^0(\bar{\Omega}) \tag{6.3}$$

which is a Banach space with norm

$$\|\lambda\|_\Lambda = \max\{\|a_{jk}\|_{C^0(\bar{\Omega})}, \|a_0\|_{C^0(\bar{\Omega})}\}.$$

The set of admissible parameters is taken to be

$$\Lambda_{ad} = \left\{\lambda \in \Lambda \,\middle|\, \sum_{j,k=1}^{n} a_{jk}(x)\xi_j\xi_k \geqq \kappa(\xi_1^2 + \cdots + \xi_n^2)\ \forall \xi \in \mathbb{R}^n\ \forall x \in \Omega \text{ and } a_0(x) \geqq \kappa_0\ \forall x \in \Omega\right\}, \tag{6.4}$$

where κ and κ_0 are given positive numbers; so (A5) is satisfied.

Now given $\lambda \in \Lambda$, define $A \in \mathscr{A} = \mathscr{L}(V, V')$[4] by

$$(Au, v)_{V'V} = \int_\Omega \sum_{j,k=1}^{n} a_{jk}(x)\frac{\partial u}{\partial x_k}\frac{\partial v}{\partial x_j} + \int_\Omega a_0(x)uv \quad \forall u, v \in V. \tag{6.5}$$

The open subset of $\mathscr{A}$

$$\mathscr{A}_c = \{A \in \mathscr{A} \,|\, \exists \zeta > 0 : (Av, v)_{V'V} \geqq \zeta \|v\|_V^2\ \forall v \in V\} \tag{6.6}$$

is the set of coercive operators. It is straightforward to verify assumptions (A4) (with $k = \infty$) and (A6).

Equation (6.2) can be rewritten as follows

$$\begin{aligned} &\frac{du}{dt} + Au = f, \\ &u(0) = u_0, \end{aligned} \tag{6.7}$$

where we have used the notation $(du/dt)(v)$ in place of $\int_\Omega \partial u/\partial t\, v$ and $f(v)$ in place of $\int_\Omega fv$. It is known [20, p. 102] that for every $A \in \mathscr{A}_c$, $f \in L^2(0, T; V')$ and $u_0 \in L^2(\Omega)$,

[4] Given X and Y Banach spaces, we denote by $\mathscr{L}(X, Y)$ the space of bounded linear operators from X into Y.

(6.7) admits a unique solution $u \in U = W(0, T) = \{u | u \in L^2(0, T; V), \ du/dt \in L^2(0, T; V')\}$ which depends continuously on f and u_0.

Taking $F = L^2(0, T; V') \times L^2(\Omega)$, $U_c = U = W(0, T)$ and defining the mapping

$$\Psi; (A, u) \in \mathcal{A} \times U \to \left(\frac{du}{dt} + Au, u(0)\right) \in F,$$

it is not difficult to verify assumptions (A1) (with $k = \infty$), (A2) and (A3).

Finally, suppose one wants to identify a_{jk} and a_0 from distributed observation i.e. by observing $u(x, t)$ in Q. Take

$$\mathcal{H} = L^2(Q), \tag{6.8}$$

$$\Lambda_{\mathcal{H}} = \text{identity}, \tag{6.9}$$

$$\mathcal{C} = \text{injection of } W(0, T) \text{ into } L^2(Q), \tag{6.10}$$

$$\mathcal{R} = \left(\sum_{j,k=1}^{n} H^l(\Omega)\right) \times H^l(\Omega) \quad \text{with } l > \frac{n}{2}. \tag{6.11}$$

Thus (A7)-(A10) are automatically satisfied.

As a consequence of Theorem 4.1 we have the following:

THEOREM 6.1. *Given $z_d \in L^2(Q)$ and $\beta > 0$, the smoothing functional*

$$J_\beta(\lambda) = \int_Q [u(x, t; \lambda) - z_d(x, t)]^2 \, dx \, dt + \beta \|\lambda\|^2_{\mathcal{R}} \tag{6.12}$$

where $u(x, t; \lambda) \in W(0, T)$ is the weak solution of (6.1), *is of C^∞-class. Its first derivative is given by*

$$J'_\beta(\lambda) \cdot \delta\lambda = \int_Q \left[\sum_{j,k=1}^{n} \delta a_{jk} \frac{\partial u}{\partial x_k} \frac{\partial \phi}{\partial x_j} + \delta a_0 u \phi\right] dx \, dt + 2\beta(\delta\lambda, \lambda)_{\mathcal{R}} \tag{6.13}$$

where $\phi \in W(0, T)$ is the weak solution of the adjoint equation

$$\frac{\partial \phi}{\partial t} + \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{kj}(x) \frac{\partial \phi}{\partial x_k}\right) - a_0(x)\phi = 2(u(x, t) - z_d(x, t)) \quad \text{in } Q,$$

$$\frac{\partial \phi}{\partial \nu} = 0 \quad \text{on } \Sigma, \tag{6.14}$$

$$\phi(x, T) = 0 \quad \text{in } \Omega.$$

Remark 6.1. The analysis of this section can be generalized for the system described by (1.1)-(1.5). See [16] for details.

Remark 6.2. A similar analysis has been performed by Chavent concerning the least-squares identification of linear parabolic systems [4, pp. 69-71].

7. Identification of a second order linear parabolic system from point observation. Given a set of discrete points $x_1, \cdots, x_\mu \in \Omega$, we now consider the identification of the system (6.1) by observing $u(x_i, t)$, $i = 1, \cdots, \mu$. We have seen in § 6 that the weak solution of (6.1) lies in $L^2(0, T; H^1(\Omega))$. Thus, for a weak solution u, the point value $u(x_i, t)$ has meaning if $H^1(\Omega) \subset C^0(\Omega) \Leftrightarrow n < 2$. Since such an assumption is overly restrictive, we will consider here strong solutions, which lie in $H^{2,1}(Q)$ (hence they lie in $C^0(\bar{Q})$ for $n \leqq 3$).[5]

[5] For a definition and properties of the spaces $H^{r,s}$ see [21, pp. 6-10].

We will make stronger regularity assumptions, such as

(7.1) Γ is an $(n-1)$-dimensional C^2-manifold, with Ω locally on one side of Γ and $a_{jk} \in C^1(\bar{\Omega})$, $a_0 \in C^0(\bar{\Omega})$, i.e. the parameter space

$$\Lambda = \left(\bigtimes_{j,k=1}^{n} C^1(\bar{\Omega}) \right) \times C^0(\bar{\Omega}) \tag{7.2}$$

which is a Banach space with norm

$$\|\lambda\|_\Lambda = \max \{ \|a_{jk}\|_{C^1(\bar{\Omega})}, \|a_0\|_{C^0(\bar{\Omega})} \}.$$

The set of admissible parameters is taken to be

(7.3)

$$\Lambda_{ad} = \left\{ \lambda \in \Lambda \,\middle|\, \sum_{j,k=1}^{n} a_{jk}(x)\xi_j\xi_k \geqq \kappa(\xi_1^2 + \cdots + \xi_n^2) \ \forall \xi \in \mathbb{R}^n \ \forall x \in \Omega \text{ and } a_0(x) \geqq \kappa_0 \ \forall x \in \Omega \right\}$$

where κ and κ_0 are given positive constants; so (A5) is satisfied.

As operator space we take

$$\mathscr{A} = \left\{ A \in \mathscr{L}(H^{2,1}(Q), L^2(Q)) \,\middle|\, A = -\sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{jk}(x)\frac{\partial}{\partial x_k} \right) + a_0(x) \right. \tag{7.4}$$
$$\left. \text{with } a_{jk} \in C^1(\bar{\Omega}) \text{ and } a_0 \in C^0(\bar{\Omega}) \right\}$$

which is a Banach space with norm

$$\|A\|_{\mathscr{A}} = \max \{ \|a_{jk}\|_{C^1(\bar{\Omega})}, \|a_0\|_{C^0(\bar{\Omega})} \}$$

and denote by $\mathscr{A}_c$ its open subset

(7.5)

$$\mathscr{A}_c = \left\{ A \in \mathscr{A} \,\middle|\, \left\{ A - e^{i\theta}\frac{\partial^2}{\partial y^2}, \frac{\partial}{\partial \nu} \right\} \text{ is a regular elliptic system on } \bar{\Omega} \times \mathbb{R}_y \ \forall \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2} \right] \right\}.$$

It is straightforward to verify assumptions (A4) (with $k = \infty$) and (A6).

It is known[6] [21, p. 33] that for every $A \in \mathscr{A}_c$, $f \in L^2(Q)$, $u_0 \in H^1(\Omega)$ and $g \in H^{1/2,1/4}(\Sigma)$, the boundary-value problem

$$\begin{aligned} &\frac{\partial u}{\partial t} + Au = f \quad \text{in } Q, \\ &u(x,0) = u_0 \quad \text{in } \Omega, \\ &\frac{\partial u}{\partial \nu} = g \quad \text{on } \Sigma \end{aligned} \tag{7.6}$$

admits a unique solution $u \in U = H^{2,1}(Q)$ that depends continuously on f, u_0 and g.

[6] Lions and Magenes use sharper regularity conditions for Γ and the coefficients of A than (7.1) and (7.2). However, the result remains unaltered. See [21, Remark 6.1, p. 35] and [4, Thm. 3.3, p. 32].

Taking $F = L^2(Q) \times H^1(\Omega) \times H^{1/2,1/4}(\Sigma)$, $U_c = U = H^{2,1}(Q)$ and defining the mapping

$$\Psi\colon (A, u) \in \mathscr{A} \times U \to \left(\frac{\partial u}{\partial t} + Au,\ u(x, 0),\ \frac{\partial u}{\partial \nu}\right) \in F$$

it is not difficult to verify assumptions (A1) (with $k = \infty$), (A2) and (A3).

Now to identify (a_{jk}) and a_0 in (6.1) from an observation of u at the points x_i, $i = 1, \cdots, \mu$, take

$$\mathscr{H} = (L^2(0, T))^{\mu}, \tag{7.7}$$

$$\Lambda_{\mathscr{H}} = \text{identity}, \tag{7.8}$$

$$\mathscr{C}\colon u(x, t) \in H^{2,1}(Q) \to (u(x_i, t), i = 1, \cdots, \mu) \in (L^2(0, T))^{\mu}, \tag{7.9}$$

$$\mathscr{R} = \left(\underset{j,k=1}{\overset{n}{\times}} H^l(\Omega)\right) \times H^{l_0}(\Omega) \quad \text{with } l > 1 + \frac{n}{2},\quad l_0 > \frac{n}{2}. \tag{7.10}$$

Thus (A7)-(A10) are automatically satisfied.

As a consequence of Theorem 4.1 we have the following:

THEOREM 7.1. *Given $z_d = (z_{d_1}(t), \cdots, z_{d_\mu}(t)) \in (L^2(0, T))^{\mu}$ and $\beta > 0$, the smoothing functional*

$$J_\beta(\lambda) = \sum_{i=1}^{\mu} \int_0^T [u(x_i, t; \lambda) - z_{d_i}(t)]^2\, dt + \beta \|\lambda\|_{\mathscr{R}}^2 \tag{7.11}$$

where $u(x, t; \lambda) \in H^{2,1}(Q)$ is the strong solution of (6.1), *is of C^∞-class. Its first derivative is given by*

$$J'_\beta(\lambda) \cdot \delta\lambda = \int_Q \left[-\sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(\delta a_{jk} \frac{\partial u}{\partial x_k}\right) + \delta a_0 u\right] \not{p}\, dx\, dt + 2\beta(\delta\lambda, \lambda)_{\mathscr{R}} \tag{7.12}$$

where $\not{p} \in L^2(Q)$ is the unique solution of

(7.13)

$$\int_Q \left[\frac{\partial v}{\partial t} - \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{jk}(x) \frac{\partial v}{\partial x_k}\right) + a_0(x) v\right] \not{p}\, dx\, dt = -2 \sum_{i=1}^{\mu} \int_0^T [u(x_i, t) - z_{d_i}(t)] v(x_i, t)\, dt$$

$\forall v \in H^{2,1}(Q)$ satisfying $\partial v/\partial \nu = 0$ on Σ, $v(x, 0) = 0$ in Ω.

In other words, $\not{p}$ is a distributional solution of

$$\begin{aligned} &\frac{\partial \not{p}}{\partial t} + \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{kj}(x) \frac{\partial \not{p}}{\partial x_k}\right) - a_0(x) \not{p} = 2 \sum_{i=1}^{\mu} (u(x_i, t) - z_{d_i}(t)) \otimes \delta(x - x_i) \quad \text{in } Q, \\ &\frac{\partial \not{p}}{\partial \nu} = 0 \quad \text{on } \Sigma, \\ &\not{p}(x, T) = 0 \quad \text{in } \Omega. \end{aligned} \tag{7.14}$$

Remark 7.1. The first term in (7.12) can be *formally* rewritten as

$$\int_Q \left[\sum_{j,k=1}^{n} \delta a_{jk} \frac{\partial u}{\partial x_k} \frac{\partial \not{p}}{\partial x_j} + \delta a_0 u \not{p}\right] dx\, dt$$

by using Green's formula.

Remark 7.2. The analysis of this section can be generalized for the system described by (1.1)-(1.5). See [16] for details.

Remark 7.3. A similar analysis has been performed by Chavent concerning the least-squares identification of linear parabolic systems [4, pp. 88-92].

8. Identification of a nonlinear parabolic system. Let Ω, Γ, T, Q, Σ as in § 6 and consider the identification of $a_{jk}(x)$ in

$$\begin{aligned} &\frac{\partial u}{\partial t} - \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{jk}(x)\frac{\partial u}{\partial x_k}\right) + b(x)u + |u|^{\gamma}u = f \quad \text{in } Q, \\ &u(x,0) = u_0(x) \quad \text{in } \Omega, \\ &u = 0 \quad \text{on } \Sigma \end{aligned} \tag{8.1}$$

from distributed observation of u, where $\gamma > 0$ and $b(x)$ is bounded below by a positive number.

The variational formulation of (8.1) is as follows:

$$\begin{aligned} &\int_{\Omega} \frac{\partial u}{\partial t} v + \int_{\Omega} \sum_{j,k=1}^{n} a_{jk}(x)\frac{\partial u}{\partial x_k}\frac{\partial v}{\partial x_j} + \int_{\Omega} b(x)uv + \int_{\Omega} |u|^{\gamma}uv = \int_{\Omega} fv \quad \forall v \in V, \\ &u(x,0) = u_0(x), \end{aligned} \tag{8.2}$$

where $V = H_0^1(\Omega)$.

The parameter space is taken to be

$$\Lambda = \bigtimes_{j,k=1}^{n} C^0(\bar{\Omega}), \tag{8.3}$$

and the set of admissible parameters

$$\Lambda_{\text{ad}} = \left\{ \lambda = (a_{jk}) \in \Lambda \;\middle|\; \sum_{j,k=1}^{n} a_{jk}(x)\xi_j\xi_k \geqq \kappa(\xi_1^2 + \cdots + \xi_n^2)\ \forall \xi \in \mathbb{R}^n\ \forall x \in \Omega \right\}, \tag{8.4}$$

where κ is a given positive number; so (A5) is satisfied.

Now given $\lambda \in \Lambda$, define $A \in \mathscr{A} = \mathscr{L}(V, V')$ by

$$(Au, v)_{V'V} = \int_{\Omega} \sum_{j,k=1}^{n} a_{jk}(x)\frac{\partial u}{\partial x_k}\frac{\partial v}{\partial x_j} \quad \forall u, v \in V. \tag{8.5}$$

The open subset of $\mathscr{A}$

$$\mathscr{A}_c = \{A \in \mathscr{A} \mid \exists \zeta > 0 \colon (Av, v)_{V'V} \geqq \zeta \|v\|_V^2\ \forall v \in V\} \tag{8.6}$$

is the set of coercive operators. It is straightforward to verify assumptions (A4) (with $k = \infty$) and (A6).

Note that (8.2) can be studied using methods of monotone operators ([13, Ch. VI § 1], [19, Ch. 2 § 1]); this leads to an existence-uniqueness result for (8.2) which is not enough to verify (A3) unless additional assumptions are made. Using the maximum principle and assuming in addition that $\exists c_1, c_2$ with $0 < c_1 < c_2$ such that

$$\begin{aligned} &0 < c_1 \leqq \frac{f(x,t)}{b(x)} \leqq c_2 \quad \text{a.e. in } Q, \\ &0 < c_1 \leqq u_0(x) \leqq c_2 \quad \text{a.e. in } \Omega, \end{aligned} \tag{8.7}$$

it can be shown [4, p. 37] that (8.2) admits a unique solution $u \in U = \{u \mid u \in L^2(0, T\colon V) \cap L^{\infty}(Q),\ du/dt + Au \in L^{\infty}(Q),\ u(0) \in L^{\infty}(\Omega)\}$ satisfying $u(x,t) \geqq c_1$ a.e. in Q.

Taking $F = L^\infty(Q) \times L^\infty(\Omega)$, $U_c = \{u \in U \mid \exists c_1 > 0: u(x, t) \geqq c_1 \text{ a.e. in } Q\}$ and defining the mapping

$$\Psi: (A, u) \in \mathscr{A}_c \times U_c \to \left(\frac{du}{dt} + Au + bu + |u|^\gamma u, u(0)\right) \in F,$$

it is not difficult to verify assumptions (A1) (with $k = 1$), (A2) and (A3).

For the identification of $\lambda = (a_{jk})$ from distributed observation of u, we take

$$\mathscr{H} = L^2(Q), \tag{8.8}$$

$$\Lambda_{\mathscr{H}} = \text{identity}, \tag{8.9}$$

$$\mathscr{C} = \text{injection of } L^\infty(Q) \text{ into } L^2(Q), \tag{8.10}$$

$$\mathscr{R} = \left(\bigtimes_{j,k=1}^{n} H^l(\Omega)\right) \quad \text{with } l > \frac{n}{2}. \tag{8.11}$$

Thus (A7)-(A10) are automatically satisfied.

As a consequence of Theorem 4.1 we have the following:

THEOREM 8.1. *The smoothing functional*

$$J_\beta(\lambda) = \int_Q [u(x, t; \lambda) - z_d(x, t)]^2 \, dx \, dt + \beta \|\lambda\|_{\mathscr{R}}^2 \tag{8.12}$$

where $u(x, t; \lambda) \in U$ is the weak solution of (8.1), *is of the C^1-class. Its derivative is given by*

$$J'_\beta(\lambda) \cdot \delta\lambda = \int_Q \sum_{j,k=1}^{n} \delta a_{jk} \frac{\partial u}{\partial x_k} \frac{\partial \not{p}}{\partial x_j} \, dx \, dt + 2\beta(\delta\lambda, \lambda)_{\mathscr{R}} \tag{8.13}$$

where $\not{p} \in L^2(0, T; H_0^1(\Omega))$ is the weak solution of the adjoint equation

$$\begin{aligned} &\frac{\partial \not{p}}{\partial t} + \sum_{j,k=1}^{n} \frac{\partial}{\partial x_j}\left(a_{kj}(x) \frac{\partial \not{p}}{\partial x_k}\right) - [b(x) + (\gamma+1)|u|^\gamma] \not{p} = 2(u(x, t) - z_d(x, t)) \quad \text{in } Q, \\ &\not{p} = 0 \quad \text{on } \Sigma \\ &\not{p}(x, T) = 0 \quad \text{in } \Omega. \end{aligned} \tag{8.14}$$

Remark. If the data are more regular, e.g. $a_{jk} \in C^1(\bar{\Omega})$, $u_0 \in H_0^1(\Omega)$, the solution of (8.1) will be in $H^{2,1}(Q)$. Hence, one will be able to consider point observation as well.

9. Numerical implementation of the regularization method. The minimization of $J_\beta(\lambda)$ can be conveniently carried out by a gradient method [3], [11], in which J_β is iteratively minimized along the gradient direction $\partial J_\beta / \partial \lambda$, which is defined as the unique element $\phi \in \mathscr{R}$ satisfying $J'_\beta(\lambda) \cdot h = (\phi, h)_{\mathscr{R}}$ $\forall h \in \mathscr{R}$. To illustrate the theory, we will consider the identification of $\alpha(x)$ in the one-dimensional diffusion equation from point observations $z_{d_i}(t)$ of $u(x_i, t)$, $i = 1, \cdots, \mu$.

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\alpha(x) \frac{\partial u}{\partial x}\right) + f \quad \text{in } \Omega \times]0, T[, \\ &u(x, 0) = u_0(x) \quad \text{in } \Omega, \\ &\frac{\partial u}{\partial x} = 0 \quad \text{on } \Gamma \times]0, T[. \end{aligned} \tag{9.1}$$

Following the analysis of § 7, we take:

$$\Lambda = C^1(\bar{\Omega}),$$
$$\Lambda_{ad} = \{\alpha \in \Lambda | \alpha(x) \geqq \alpha_0 > 0 \ \forall x \in \Omega\},$$
$$\mathscr{A} = \{A \in \mathscr{L}(H^{2,1}(Q), L^2(Q)) | A = (\partial/\partial x)(\alpha(x)\, \partial/\partial x), \text{ where } \alpha \in C^1(\bar{\Omega})\},$$
$$U = H^{2,1}(Q),$$
$$\mathscr{H} = (L^2(0, T))^{\mu},$$
$$\mathscr{R} = H^2(\Omega),$$
$$\mathscr{R}_{ad} = \{\alpha \in H^2(\Omega) | \alpha(x) \geqq \alpha_0 > 0 \ \forall x \in \Omega\}.$$

The smoothing functional is

$$J_\beta(\alpha) = \sum_{i=1}^{\mu} \int_0^T [u(x_i, t) - z_{d_i}(t)]^2 \, dt + \beta \|\alpha\|^2_{H^2(\Omega)}. \tag{9.2}$$

Applying Theorem 7.1 we find that its first derivative is given by

$$J'_\beta(\alpha) \cdot \delta\alpha = \int_\Omega \delta\alpha \int_0^T \frac{\partial u}{\partial x} \frac{\partial \not{p}}{\partial x} \, dt \, dx + 2\beta(\delta\alpha, \alpha)_{H^2(\Omega)}, \tag{9.3}$$

where $\not{p}$ is the solution of

$$\begin{aligned} &\frac{\partial \not{p}}{\partial t} + \frac{\partial}{\partial x}\left(\alpha(x) \frac{\partial \not{p}}{\partial x}\right) = 2 \sum_{i=1}^{\mu} (u(x_i, t) - z_{d_i}(t)) \otimes \delta(x - x_i) \quad \text{in } \Omega \times]0, T[, \\ &\frac{\partial \not{p}}{\partial x} = 0 \quad \text{on } \Gamma \times]0, T[, \\ &\not{p}(x, T) = 0 \quad \text{in } \Omega. \end{aligned} \tag{9.4}$$

If $H^2(\Omega)$ is equipped with the norm

$$\|f\|_{H^2(\Omega)} = \left(\int_\Omega (f^2 + f''^2) \, dx\right)^{1/2}, \tag{9.5}$$

the gradient $\partial J_\beta / \partial \alpha$ is given by

$$\frac{\partial J_\beta}{\partial \alpha}(\alpha) = \psi + 2\beta\alpha, \tag{9.6}$$

where ψ is the weak solution of

$$\begin{aligned} &\frac{d^4\psi}{dx^4} + \psi = \int_0^T \frac{\partial u}{\partial x} \frac{\partial \not{p}}{\partial x} \, dt \quad \text{in } \Omega, \\ &\psi'' = 0 \quad \text{on } \Gamma, \\ &\psi''' = 0 \quad \text{on } \Gamma, \end{aligned} \tag{9.7}$$

with u and $\not{p}$ being the solutions of the state and adjoint equations respectively. The gradient algorithm in this case proceeds as follows:

(1) Initialize $\alpha \in H^2(\Omega)$.
(2) Solve the state and adjoint equations.
(3) Calculate $J_\beta(\alpha)$ and $\int_0^T (\partial u/\partial x)(\partial \not{p}/\partial x) \, dt$.

(4) If $|J_\beta(\alpha^{\text{old}}) - J_\beta(\alpha^{\text{new}})| < \text{Tolerance}$, stop.
(5) Solve (9.7) for ψ and calculate $(\partial J_\beta/\partial\alpha)(\alpha)$.
(6) Set $\alpha^{\text{new}} = \alpha^{\text{old}} + \varepsilon(\partial J_\beta/\partial\alpha)(\alpha^{\text{old}})$ where ε is a step length parameter to be determined by one-dimensional line-search.
(7) Go to (2).

We have considered the three cases given in Table 1. Data were generated by first numerically solving (9.1) using the Crank–Nicolson scheme with 50 grid points and then adding to $u(x_i, t)$ random numbers with zero mean and standard deviation $\sigma = 0.2$.

The smoothing functional

$$J_\beta(\alpha) = \frac{1}{5}\int_0^{0.5} \sum_{i=1}^{10} (u(x_i, t) - z_{d_i}(t))^2\, dt + \beta \int_0^1 [(\alpha(x))^2 + (\alpha''(x))^2]\, dx \tag{9.8}$$

was minimized by applying the gradient algorithm described above. The state and adjoint equations were solved by the Crank–Nicolson method. The fourth order O.D.E. giving ψ was solved by a finite-difference scheme. The one-dimensional line search for the step length was performed by the golden section search method. Finally, the test for stopping the iterations was

$$|J_\beta(\alpha^{\text{new}}) - J_\beta(\alpha^{\text{old}})| < 10^{-3}.$$

The initial guess for $\alpha(x)$, the true $\alpha(x)$ and the result after six iterations of the gradient method are shown for Case 1 in Fig. 2. Similarly, the estimated $\alpha(x)$ after six iterations is shown for Case 2 in Fig. 3. In each of Cases 1 and 2 the value of the regularization parameter β was selected based on the suggestion of Miller (see § 5). In Case 1, with an assumed upper bound of 0.05 for the squared error and an assumed upper bound for smoothness of $\|\alpha^{\text{true}}\|_{H^2}^2 \leqq 1$, we obtain $\beta = 5 \times 10^{-2}$. In Case 2, with the same assumed upper bound of 0.05 for the squared error and that for smoothness of $\|\alpha^{\text{true}}\|_{H^2}^2 \leqq 10$, we have $\beta = 5 \times 10^{-3}$.

TABLE 1

Numerical values for identification of $\alpha(x)$ *in*

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\alpha(x)\frac{\partial u}{\partial x}\right),\ 0 < x < 1,\ 0 < t < 0.5,$$

$$u(x, 0) = 10 + 270x^2 - 180x^3,$$

$$\frac{\partial u}{\partial x}(0, t) = \frac{\partial u}{\partial x}(1, t) = 0,$$

based on noisy observations of u at the points $x_i = (5i-3)/49$, $i = 1, \cdots, 10$.

Case	True $\alpha(x)$	$\|\alpha\|_{H^2}^2$	Standard deviation of errors in the data	Regularization parameter
1	$0.5 + 0.5x$	0.583	0.2	5×10^{-2}
2	$0.5 + x - 5x^4 + 6x^5 - 2x^6$	6.236	0.2	5×10^{-3}
3	$0.5 + 0.5x - 0.5x^2$	1.342	0.2	10^{-4}
				10^{-2}
				1

The effect of the choice of β is examined in Case 3. Figure 4 shows three estimated $\alpha(x)$ profiles corresponding to $\beta = 10^{-4}$, 10^{-2}, and 1. The value $\beta = 10^{-2}$ is consistent with the suggestion of Miller. We note that when $\beta = 10^{-4}$ the oscillations in $\alpha(x)$, characteristic of numerical instability, are setting in. In the absence of a good estimate

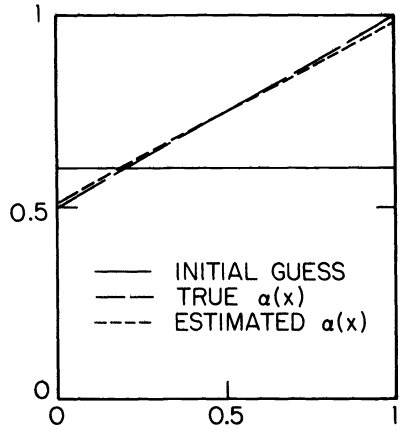


FIG. 2. *True and estimated profiles of α for Case* 1.

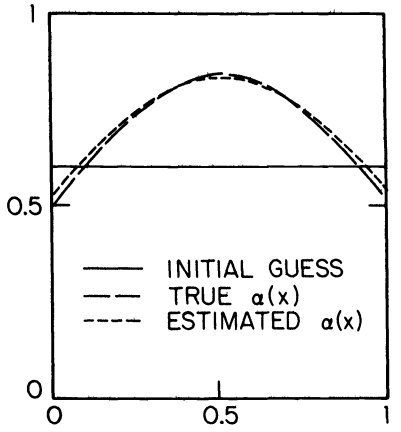


FIG. 3. *True and estimated profiles of α for Case* 2.

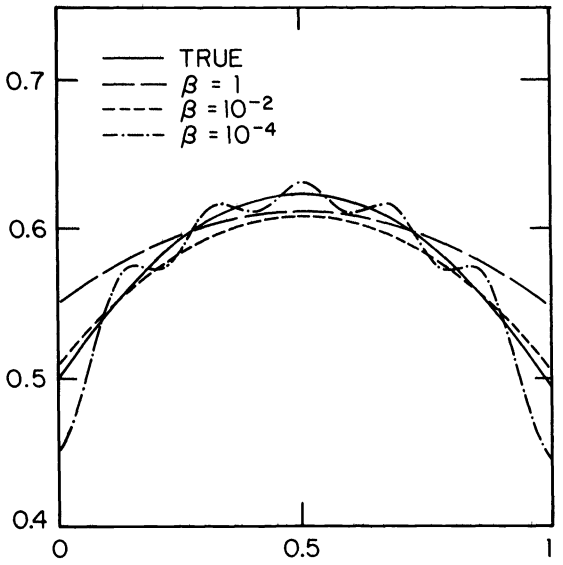


FIG. 4. *True and estimated profiles of α for Case* 3.

for the errors and/or smoothness, it is a good idea to examine the solution as a function of β.

REFERENCES

[1] J. P. AUBIN, *Applied Abstract Analysis*, John Wiley, New York, 1977.

[2] A. BENSOUSSAN, J. L. LIONS AND G. PAPANICOLAOU, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, 1978.

[3] J. CÉA, *Optimization: théorie et algorithmes*, Dunod, Paris, 1971.

[4] G. CHAVENT, *Analyse fonctionelle et identification de coefficients répartis dans les équations aux dérivées partielles*, Thèse d'État, Faculté des Sciences de Paris, 1971.

[5] G. CHAVENT, M. DUPUY AND P. LEMONNIER, *History matching by use of optimal control theory*, Soc. Pet. Eng. J., 15 (1975), pp. 74-86.

[6] G. CHAVENT, *About the stability of the optimal control solution of inverse problems*, in Inverse and Improperly Posed Problems in Differential Equations, G. Anger, ed., Akademie-Verlag, Berlin, 1979, pp. 45-78.

[7] G. CHAVENT, *Identification of distributed parameter systems: About the output least square method, its implementation and identifiability*, Proc. 5th IFAC Symposium on Identification and System Parameter Estimation, Vol. I, R. Isermann, ed., Pergamon Press, New York, 1980, pp. 85-97.

[8] W. H. CHEN AND J. H. SEINFELD, *Estimation of spatially varying parameters in partial differential equations*, Int. J. Control, 15 (1972), pp. 487-495.

[9] W. H. CHEN, G. R. GAVALAS, J. H. SEINFELD AND M. L. WASSERMAN, *A new algorithm for automatic history matching*, Soc. Pet. Eng. J., 14 (1974), pp. 593-608.

[10] K. H. COATS, J. R. DEMPSEY AND J. H. HENDERSON, *A new technique for determining reservoir descriptions from field performance data*, Soc. Pet. Eng. J., 10 (1970), pp. 66-74.

[11] J. W. DANIEL, *The Approximate Minimization of Functionals*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[12] N. DISTEFANO AND A. RATH, *An identification approach to subsurface hydrological systems*, Water Resour. Res., 11 (1975), pp. 1005-1012.

[13] H. GAJEWSKI, K. GRÖGER AND K. ZACHARIAS, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.

[14] G. R. GAVALAS, P. C. SHAH AND J. H. SEINFELD, *Reservoir history matching by Bayesian estimation*, Soc. Pet. Eng. J., 16 (1976), pp. 337-350.

[15] 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.

[16] C. KRAVARIS, *Identification of spatially-varying parameters in distributed parameter systems*, Ph.D. Thesis, California Institute of Technology, Pasadena, 1984.

[17] M. M. LAVRENTIEV, V. G. ROMANOV AND V. G. VASILIEV, *Multidimensional Inverse Problems for Differential Equations*, Springer-Verlag, Berlin, 1970.

[18] J. L. LIONS, *Some aspects of modeling problems in distributed parameter systems*, in Proc. IFIP Working Conference, Rome, 1976, A. Ruberti, ed., Lecture Notes in Control and Information Sciences Vol. 1, Springer-Verlag, Berlin, 1978, pp. 11-41.

[19] ——, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris, 1969.

[20] ——, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[21] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. II, Springer-Verlag, Berlin, 1972.

[22] K. MILLER, *Least-squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52-74.

[23] M. Z. NASHED, *Approximate regularized solutions to improperly posed linear integral and operator equations*, in Constructive and Computational Methods for Differential and Integral Equations, D. L. Colton and R. P. Gilbert, eds., Springer-Verlag, Berlin, 1974, pp. 289-332.

[24] S. P. NEUMAN AND S. YAKOWITZ, *A statistical approach to the inverse problem of aquifer hydrology. 1. Theory*, Water Resour. Res., 15 (1979), pp. 845-860.

[25] P. C. SHAH, G. R. GAVALAS AND J. H. SEINFELD, *Error analysis in history matching: The optimum level of parameterization*, Soc. Pet. Eng. J., 18 (1978), pp. 219-228.

[26] L. SCHWARTZ, *Cours d'analyse*, Hermann, Paris, 1967.

[27] A. N. TIKHONOV, *Solution of ill-posed problems and the regularization method*, Dokl. Akad. Nauk SSSR, 151 (1963), pp. 501-504; Soviet Math. Dokl., 4 (1963), pp. 1035-1038.

[28] ——, *Regularization of ill-posed problems*, Dokl. Akad. Nauk SSSR, 153 (1963), pp. 49-52; Soviet Math. Dokl., 4 (1963), pp. 1624-1627.

[29] A. N. TIKHONOV AND V. YA. ARSENIN, *Solutions of Ill-Posed Problems*, Winston-Wiley, New York, 1977.

[30] T. L. VINCENT AND W. J. GRANTHAM, *Optimality in Parametric Systems*, Wiley, New York, 1981.

[31] G. WAHBA, *Practical approximate solutions to linear operator equations when data are noisy*, SIAM J. Numer. Anal., 14 (1977), pp. 651-667.

[32] S. YAKOWITZ AND L. DUCKSTEIN, *Instability in aquifer identification: Theory and case studies*, Water Resour. Res., 16 (1980), pp. 1054-1064.

[33] P. L. YU, *Cone convexity, cone extreme points, and nondominated solutions in decision problems with multiobjectives*, in Multicriteria Decision Making and Differential Games, G. Leitmann, ed., Plenum, New York, 1976, pp. 1-60.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

ON DETERMINISTIC CONTROL PROBLEMS: AN APPROXIMATION PROCEDURE FOR THE OPTIMAL COST I THE STATIONARY PROBLEM*

R. GONZALEZ† AND E. ROFMAN‡

Abstract. We study deterministic optimal control problems having stopping time, continuous and impulse controls in each strategy.

We obtain the optimal cost, considered as the maximum element of a suitable set of subsolutions of the associated Hamilton-Jacobi equation, using an approximation method. A particular derivative discretization scheme is employed.

Convergence of approximate solutions is shown taking advantage of a discrete maximum principle which is also proved.

For the numerical solutions of approximate problems we use a method of relaxation type. The algorithm is very simple; it can be run on computers with small central memory.

In Part I we study the stationary case, in Part II [SIAM J. Control Optim., 23 (1985), pp. 267-285] we study the nonstationary case.

Key words. deterministic control, Hamilton-Jacobi-Bellman equations, finite elements

Introduction. Previously [9], we have dealt with the numerical solution of some optimal deterministic problems using as a basic tool of analysis the characterization (introduced in [7], [8]) of the optimal cost function as the maximum element of a suitable set of subsolutions of the associated Hamilton-Jacobi equation. In this paper, to compute this maximum element, we present a new algorithm that makes possible solutions of nontrivial problems on computers with small central memories.

In part I we study the stationary case. In Part II (this issue, pp. 267-285) we study the nonstationary case.

In § 1 we introduce a control problem with a cost function J to be minimized. We consider in each strategy stopping times, continuous and impulse control; so $V(x) = \inf_{\tau,u(\cdot),z(\cdot)} J(x; \tau, u(\cdot), z(\cdot))$. After the definition of a suitable set W of subsolutions and following the same techniques used in [7], [8], [9], we characterize $V(x)$ as the unique solution of the equivalent problem (P): *Find the maximum element of the set W.*

In § 2 we consider the discretized problem (P_h), its solution $(\bar{w}_h)$, the algorithm to compute it and its properties. Using a particular scheme to discretize the partial derivatives of the functions under consideration we are enabled to define an algorithm that, with successive iterations, increases the values of these functions in the vertices of the triangulation employed, until the approximate solution $\bar{w}_h$ is found.

In § 3 the convergence is proved. An estimation of the rate of convergence is given.

1. The optimal control problem and an equivalent formulation (the problem (P)). To control a system with trajectories in an open bounded set $\Omega \in \mathbb{R}^n$ we use stopping time control, impulse control and continuous control.

* Received by the editors October 7, 1982, and in final revised form April 15, 1984. This work was supported in part by the U.S. Department of Energy, Office of Electric Energy Systems, under contract 01-80RA-50154.

† Electronics Department, University of Rosario, Argentina. The work of this author was supported in part by CONICET, Argentina, under grant 9977/81.

‡ INRIA, Domaine de Voluceau, BP105—Rocquencourt, 78153 Le Chesnay Cedex, France.

In the intervals of time free of the action of impulse controls, the trajectory of the system satisfies the differential equation

$$\begin{aligned} &\frac{dy}{dt} = f(y, u), \\ &y(0) = x, \qquad x \in \Omega \subset \mathbb{R}^n, \end{aligned} \tag{1.1}$$

where $u(\cdot)$ is a measurable function of time with values in a compact set $U \subset \mathbb{R}^m$.

At times $\theta_\nu (0 \leqq \theta_1 < \theta_2 < \cdots)$ impulses $z(\theta_\nu)$ are applied; they produce jumps of amplitude g_ν in the trajectory of the system:

$$y(\theta_\nu^+) = y(\theta_\nu^-) + g(y(\theta_\nu^-), z(\theta_\nu)); \tag{1.2}$$

$y(\theta_\nu^+)$, $(y(\theta_\nu^-))$ is the right (left) limit of the trajectory $y(\cdot)$. The set Z of admissible impulse controls is a compact set of $\mathbb{R}^p$.

The control strategy is determined by the stopping time $\tau \geqq 0$, the function $u(\cdot)$, the times $\{\theta_\nu\}$ and the impulses $\{z(\theta_\nu)\}$; it will be noted by $(\tau, u(\cdot), z(\cdot))$.

In the following we shall suppose that $\forall t$, $y(t) \in \Omega$.

We assign to each control strategy the cost value J:

$$\begin{aligned} J(x; \tau, u(\cdot), z(\cdot)) = &\int_0^\tau e^{-\alpha s} l(y(s), u(s))\, ds \\ &+ \phi(y(\tau))\, e^{-\alpha\tau} + \sum_\nu e^{-\alpha\theta_\nu} q(y(\theta_\nu^-), z(\theta_\nu)) \end{aligned} \tag{1.3}$$

in which l is the instantaneous cost, ϕ is the final cost, $\alpha > 0$ is the discount factor and $q > 0$ is the cost of application of an impulse.

Our aim is to find the optimal cost function $V(x)$ defined by

$$V(x) = \inf_{\tau, u(\cdot), z(\cdot)} J(x; \tau, u(\cdot), z(\cdot)) \quad \forall x \in \Omega. \tag{1.4}$$

In what follows we will always suppose that

i) f, l, ϕ, g, q are continuous and bounded functions ($M_f, \cdots, M_q$ being the bounds);

ii) f, l, ϕ, g, q are Lipschitzian functions of y ($L_f, \cdots, L_q$ being the Lipschitz constants);

iii)
$$\alpha > \begin{cases} L_f + \mu_\delta \ln \lambda_g & (\text{if } \lambda_g > 1), \\ L_f & (\text{if } \lambda_g \leqq 1), \end{cases}$$

with

$$\lambda_g = \sup \left\{ \frac{\|x + g(x, z) - x' - g(x', z)\|}{\|x - x'\|} \,\middle|\, x \neq x';\ x, x' \in \Omega,\ z \in Z \right\},$$

$$\mu_\delta = \frac{2\, e(M_l + \alpha M_\phi)}{q_0}, \quad \mu_0 = \frac{2\, e(M_l + \alpha M_\phi)}{\alpha q_0}, \quad q_0 = \inf_{\substack{x \in \Omega \\ z \in Z}} q(x, z) > 0.$$

The characterization of $V(x)$ given in Theorem 1.2 concerns Lipschitzian functions. So it is useful to recall the following result, which is easily obtained as a combination of those shown in [9].

THEOREM 1.1 (Lipschitz continuity of $V(x)$). *Under assumptions* i), ii) *and iii)* $V(x)$ *is a Lipschitzian function, i.e.*

$$|V(x) - V(x')| \leqq L_v \|x - x'\| \quad \forall x, x' \in \Omega,$$

in which

$$L_v = \begin{cases} L_l \lambda_g^{\mu_0} \dfrac{1}{\alpha - L_f - \mu_\delta \ln \lambda_g} + L_q \dfrac{e^{(\alpha - L_f)/\alpha}}{1 - \lambda_g e^{(L_f - \alpha)/\mu_\delta}} + L_\phi \lambda_g^{\mu_0} & \text{if } \lambda_g > 1, \\ L_l \dfrac{1}{\alpha - L_f} + L_q \dfrac{e^{(\alpha - L_f)/\alpha}}{1 - \lambda_g e^{(L_f - \alpha)/\mu_\delta}} + L_\phi & \text{if } \lambda_g \leqq 1. \end{cases}$$

As a consequence $V(x)$ is a.e. differentiable in Ω. Using the techniques of dynamic programming it is possible to show (cf. [9]) that $V(x)$ is a solution of the Hamilton-Jacobi inequality associated with the optimal control problem,

$$\min_{u \in U} \left(\frac{\partial V(x)}{\partial x} \cdot f(x, u) + l(x, u) - \alpha V(x) \right) \geqq 0, \quad x \text{ a.e. in } \Omega, \tag{1.5}$$

$$V(x) - \min_{z \in Z} (V(x + g(x, z)) + q(x, z)) \leqq 0 \qquad \forall x \in \Omega, \tag{1.6}$$

$$V(x) - \phi(x) \leqq 0 \qquad \forall x \in \Omega. \tag{1.7}$$

(1.8) For all x at which $V(\cdot)$ is differentiable, $V(\cdot)$ satisfies one at least of (1.5), (1.6), (1.7) with equality.

Following the technique used in [7], [8] we have proved in [9]:

THEOREM 1.2 (characterization of $V(x)$). *Let*

$$W = \{w: \Omega \to \mathbb{R} \mid (1.10), (1.11), (1.12), (1.13)\}, \tag{1.9}$$

where

(1.10) *w is a Lipschitzian function,*

$$\min_{u \in U} \left(\frac{\partial w(x)}{\partial x} \cdot f(x, u) + l(x, u) - \alpha w(x) \right) \geqq 0 \quad \text{a.e. } x \in \Omega, \tag{1.11}$$

$$w(x) - \min_{z \in Z} (w(x + g(x, z)) + q(x, z)) \leqq 0 \qquad \forall x \in \Omega, \tag{1.12}$$

$$w(x) \leqq \phi(x) \qquad \forall x \in \Omega. \tag{1.13}$$

Then $V(x)$ is the maximum element of the set (1.9) *i.e. $V(x) \in W$ and*

$$V(x) \geqq w(x) \quad \forall x \in \Omega, \quad \forall w \in W. \tag{1.14}$$

Clearly Theorem 1.2 makes possible the determination of the optimal cost function defined in (1.4) by solving the equivalent problem:

(P): Find the maximum element in the set W defined by (1.9).

2. The discretized problem (P_h).

2.1. Preliminary comments. In this section we shall introduce sets W^h, finite-dimensional approximations of W, looking for a numerical device to compute $V(x)$. Following this idea, after a discretization Ω^h of the set Ω, we shall define W^h by functions w^h verifying properties related to (1.10)-(1.13). The main difficulty of this approach is the choice of W^h having maximum element $\bar{w}^h$.

In fact, after introducing in W^h the natural partial order

$$w_1 \leqq \hat{w}_2 \Leftrightarrow w_1(x_i^h) \leqq w_2(x_i^h) \quad \forall x_i^h \text{ vertex of } \Omega^h, \tag{2.1}$$

it is not possible, in general, to ensure the existence of $\bar{w}^h$. We show in what follows, that thanks to a criterion used in the discretization of the derivatives that appear in (1.11) (see (2.3)) we obtain

a) the existence of an unique maximum element $\bar{w}^h$ in W^h;

b) a characterization of $\bar{w}^h$ that enables us to compute it with an iterative algorithm of relaxation type.

2.2. The discretization procedure.

a) The set Ω is approximated with a triangulation Ω^h (union of simplices) (see Fig. 1).

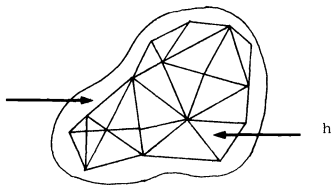


FIG. 1

b) We consider, in place of W, the set W^h of functions $w^h: \bar{\Omega}^h \to \mathbb{R}$, w^h continuous in $\bar{\Omega}^h$, $\partial w^h/\partial x$ constant in the interior of each simplex of Ω^h (i.e. w^h are linear finite elements) satisfying at every x_i^h of Ω^h the restrictions (2.2), (2.4), (2.5), (discretization of (1.11), (1.12), (1.13)):

$$\frac{\partial w^h}{\partial x_f}(x_i^h; u) \cdot \|f(x_i^h, u)\| + l(x_i^h, u) - \alpha w^h(x_i^h) \geqq 0 \tag{2.2}$$

for all $u \in U^h$. Here U^h is a finite set which approximates the set of admissible continuous controls U.

We denote by $(\partial w^h/\partial x_f)(x_i^h; u)$ the derivative of w^h in the direction of f, more precisely (see Fig. 2):

$$\frac{\partial w^h}{\partial x_f}(x_i^h; u) \cdot \|f(x_i^h, u\| = \begin{cases} \dfrac{w^h(a_i(u)) - w^h(x_i^h)}{\|a_i(u) - x_i^h\|} \cdot \|f(x_i^h, u)\| & \text{if } f(x_i^h, u) \neq 0, \\ 0 \quad \text{if } f(x_i^h, u) = 0; \end{cases} \tag{2.3}$$

$$w^h(x_i^h) \leqq w^h(x_i^h + g(x_i^h, z)) + q(x_i^h, z) \quad \forall z \in Z^h \subset Z, \tag{2.4}$$

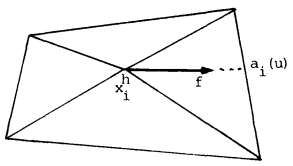


FIG. 2

with Z^h a finite set that approximates the set of admissible impulse controls Z, and

$$w^h(x_i^h) \leqq \phi(x_i^h). \tag{2.5}$$

c) We introduce problem (P_h), a discretized version of problem (P):

(P_h): Find the maximum element $\bar{w}^h$ of the set W^h with respect to the partial order (2.1) in W^h.

d) *Remarks relevant to comment* b).

d_1) We suppose always that $x_i^h + g(x_i^h, z) \in \Omega^h$ in order to ensure that (2.4) makes sense.

d_2) If D is the diameter of a simplex, there exists $\gamma_1 > 0$ such that for each simplex in Ω^h there exists a sphere of radius $r \geqq \gamma_1 D$ in the interior of the simplex (Fig. 3).

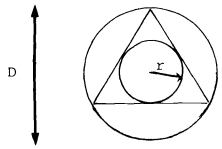


FIG. 3

d_3) Denoting by $\|h\|$ the maximum of the diameters of the simplices of Ω^h, the sets U^h and Z^h approximate the sets U and Z in the following sense:

$$U^h \subset U, \qquad Z^h \subset Z,$$

$$\|h\| \leqq \|h'\| \Rightarrow U^{h'} \subset U^h, \qquad Z^{h'} \subset Z^h,$$

$$\bigcup_h U^h = U, \qquad \bigcup_h Z^h = Z.$$

2.3. Existence of a solution of problem (P_h). In the set W^h a partial order is defined. As a consequence we can speak of maximal elements but, as was said before, it is not obvious, a priori, that a maximum element exists. To prove property a) of § 2.1 we will transform the constraints (2.2) and (2.4) into more useful equivalent relations (2.2)′ and (2.4)′.

Considering that $a_i(u)$ and $x_i^h + g(x_i^h, z)$ are linear convex combinations of the simplices to which they belong, we have (Fig. 4):

$$\text{(2.6)} \qquad \begin{aligned} a_i(u) &= \sum_{j=1}^{n_h} \lambda_j(x_i^h, u) \cdot x_j^h, \qquad \lambda_j \geqq 0, \quad \sum_{j=1}^{n_h} \lambda_j = 1, \\ x_i^h + g(x_i^h, z) &= \sum_{j=1}^{n_h} \lambda_j'(x_i^h, z) \cdot x_j^h, \quad \lambda_j' \geqq 0, \quad \sum_{j=1}^{n_h} \lambda_j' = 1. \end{aligned}$$

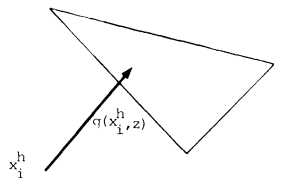


FIG. 4

Since w^h is an affine function of each simplex, (2.2) and (2.4) are equivalent to:

$$\text{(2.2)}' \qquad w^h(x_i^h) \leqq \min_{u \in U^h} \left[\beta_1(x_i^h, u) \sum_{j=1}^{n_h} \lambda_j(x_i^h, u) \cdot w^h(x_j^h) + \beta_2(x_i^h, u) l(x_i^h, u) \right],$$

$$\text{(2.4)}' \qquad w^h(x_i^h) \leqq \min_{z \in Z^h} \left[\sum_{j=1}^{n_h} \lambda_j'(x_i^h, z) \cdot w^h(x_j^h) + q(x_i^h, z) \right],$$

where

$$\beta_1(x_i^h, u) = \begin{cases} 0 & \text{if } f(x_i^h, u) = 0, \\ \dfrac{\|f(x_i^h, u)\|}{\|f(x_i^h, u)\| + \alpha \|a_i(u) - x_i^h\|} & \text{if } f(x_i^h, u) \neq 0; \end{cases}$$

$$\beta_2(x_i^h, u) = \begin{cases} \dfrac{1}{\alpha} & \text{if } f(x_i^h, u) = 0, \\ \dfrac{\|a_i(u) - x_i^h\|}{\|f(x_i^h, u)\| + \alpha \|a_i(u) - x_i^h\|} & \text{if } f(x_i^h, u) \neq 0. \end{cases}$$

So, we can now consider W^h as the set of linear finite elements on Ω^h satisfying (2.2)′, (2.4)′, (2.5), and we can pass to:

THEOREM 2.1. *There exists $\bar{w}^h$, maximum element of W^h.*

Proof. Let be

$$\hat{w}^h(x_i^h) = \sup \{w^h(x_i^h) | w^h \in W^h\}. \tag{2.7}$$

$\hat{w}^h$ is well-defined by virtue of (2.5). From (2.5), (2.7) it follows that $\hat{w}^h$ verifies (2.5).

In (2.2)′, (2.4)′ the factors that multiply $w^h(x_j^h)$ are nonnegative; then by virtue of (2.2)′ and (2.7) we have

$$w^h(x_i^h) \leqq \min_{u \in U^h} \left[\beta_1(x_i^h, u) \cdot \sum_{j=1}^{n_h} \lambda_j(x_i^h, u) \hat{w}^h(x_j^h) + \beta_2(x_i^h, u) l(x_i^h, u) \right];$$

and taking into account (2.7) we have

$$\hat{w}^h(x_i^h) \leqq \min_{u \in U^h} \left[\beta_1(x_i^h, u) \cdot \sum_{j=1}^{n_h} \lambda_j(x_i^h, u) \cdot \hat{w}^h(x_j^h) + \beta_2(x_i^h, u) l(x_i^h, u) \right],$$

i.e. $\hat{w}^h$ verifies (2.2)′. In a similar way it is proved that $\hat{w}^h$ verifies (2.4)′ and in consequence $\hat{w}^h \in W^h$. Now, by virtue of (2.7), $\hat{w}_h$ is the maximum element of W^h, i.e. $\hat{w}^h = \bar{w}^h$.

2.4. Characterization of the maximum element $\bar{w}^h$. We define the operator $M: \mathbb{R}^{n_h} \to \mathbb{R}^{n_h}$ in the following manner

$$\begin{aligned} (Mw^h)(x_i^h) = \min \Bigg\{ & \min_{u \in U^h} \left[\beta_1(x_i^h, u) \cdot \sum_{j=1}^{n_h} \lambda_j(x_i^h, u) \cdot w^h(x_j^h) + \beta_2(x_i^h, u) l(x_i^h, u) \right], \\ & \min_{z \in Z^h} \left[\sum_{j=1}^{n_h} \lambda_j'(x_i^h, z) w^h(x_j^h) + q(x_i^h, z) \right], \phi(x_i^h) \Bigg\}, \end{aligned} \tag{2.8}$$

and we obtain the following characterization of $\bar{w}^h$.

THEOREM 2.2. *$\bar{w}^h$ is the maximum element of W^h if and only if $\bar{w}^h \equiv M\bar{w}^h$ (i.e. if and only if for all $x_i^h \in \Omega^h$ one at least of (2.2)′ (2.4)′, (2.5) is an equality).*

a) *Proof of the necessary condition.* Let $\bar{w}^h$ be the maximum element of W^h, and suppose that i_0 and $\varepsilon > 0$ exist such that

$$\bar{w}^h(x_{i_0}^h) + \varepsilon \leqq (M\bar{w}^h)(x_{i_0}^h). \tag{2.9}$$

We define

$$\begin{aligned} &\check{w}^h(x_i^h) = \bar{w}^h(x_i^h) \quad \forall i \neq i_0, \\ &\check{w}^h(x_{i_0}^h) = \bar{w}^h(x_{i_0}^h) + \varepsilon. \end{aligned} \tag{2.10}$$

Then, by virtue of (2.9) and the monotonicity of M, we obtain

$$\begin{aligned} &\check{w}^h(x_i^h) = \bar{w}^h(x_i^h) \leqq (M\bar{w}^h)(x_i^h) = (M\check{w}^h)(x_i^h) \quad \forall i \neq i_0, \\ &\check{w}(x_{i_0}^h) \leqq (M\bar{w}^h)(x_{i_0}^h) \leqq (M\check{w}^h)(x_{i_0}^h). \end{aligned}$$

In consequence $\check{w}^h \in W^h$ and by (2.10), $\check{w}^h > \bar{w}^h$; this contradiction has the origin in (2.9). So $\bar{w}^h(x_i^h) = M\bar{w}^h(x_i^h) \forall i$, i.e. $\bar{w}^h \equiv M\bar{w}^h$.

For the proof of the sufficient condition we will introduce the following lemma.

LEMMA 2.1 (discrete maximum principle). *Let C^h be a subset of vertices of Ω^h and S^h its complement. If for all $x_i \in C^h$*

$$\min_{u \in U_i^h} \left(\frac{\partial w^h}{\partial x_f}(x_i^h; u) \cdot \|f(x_i^h, u)\| + l(x_i^h, u) - \alpha w^h(x_i^h) \right) \geqq 0, \tag{2.11}$$

then there exists γ, $0 < \gamma < 1$ such that

$$\max_{x_i^h \in C^h} w^h(x_i^h) \leqq \gamma \left[\max_{x_j^h \in S^h} w^h(x_j^h) \vee 0 \right] + \frac{1}{\alpha} \left[\max_{\substack{x_i^h \in C^h \\ u \in U_i^h}} l(x_i^h, u) \vee 0 \right]. \tag{2.12}$$

Proof. We rewrite (2.11) in its equivalent form (2.2)′:

$$w^h(x_i^h) \leqq \min_{u \in U_i^h} \left(\beta_j(x_i^h, u) \sum_{j=1}^{n_h} \lambda_j(x_i^h, u) w^h(x_j^h) + \beta_2(x_i^h, u) l(x_i^h, u) \right). \tag{2.13}$$

Let $x_{i_0}^h \in C^h$ be such that

$$w^h(x_{i_0}^h) = \max_{x_i^h \in C^h} (w^h(x_i^h)) = M_{C^h};$$

we denote

$$M_{S^h}^+ = \max_{x_j^h \in S^h} (w^h(x_j^h) \vee 0), \qquad M_l^+ = \max_{\substack{x_j^h \in C^h \\ u \in U_i^h}} (l(x_j^h, u) \vee 0), \tag{2.14}$$

and we have $\forall u \in U_{i_0}^h$ from (2.13)

$$\begin{aligned} M_{C^h} = w^h(x_{i_0}^h) \leqq \beta_1(x_{i_0}^h, u) \Bigg(&\sum_{j/x_j^h \in C^h} \lambda_j(x_{i_0}^h, u) M_{C^h} \\ &+ \sum_{j/x_j^h \in S^h} \lambda_j(x_{i_0}^h, u) M_{S^h}^+ \Bigg) + \beta_2(x_{i_0}^h, u) M_l^+ \Bigg). \end{aligned} \tag{2.15}$$

After putting

$$\lambda^C(x_{i_0}^h, u) = \sum_{j/x_j^h \in C^h} \lambda_j(x_{i_0}^h, u)$$

we obtain

$$\sum_{j/x_j^h \in S^h} \lambda_j(x_{i_0}^h, u) = 1 - \lambda^C(x_{i_0}^h, u);$$

so we can rewrite (2.15) as follows:

$$M_{C^h} \leqq \frac{\beta_1(x^h_{i_0}, u)(1-\lambda^C(x^h_{i_0}, u))}{1-\beta_1(x^h_{i_0}, u)\lambda^C(x^h_{i_0}, u)} M^+_{S_h} + \frac{\beta_2(x^h_{i_0}, u) M^+_l}{1-\beta_1(x^h_{i_0}, u)\lambda^C(x^h_{i_0}, u)} \tag{2.16}$$

$\forall u \in U^h_{i_0}$. As $0 \leqq \beta < 1$, $0 \leqq \lambda^C \leqq 1$ we have

$$\frac{\beta_1(x^h_{i_0}, u)(1-\lambda^C(x^h_{i_0}, u))}{1-\beta_1(x^h_{i_0}, u)\lambda^C(x^h_{i_0}, u)} \leqq \beta_1(x^h_{i_0}, u),$$

$$\frac{1}{1-\beta_1(x^h_{i_0}, u)\lambda^C(x^h_{i_0}, u)} \leqq \frac{1}{1-\beta_1(x^h_{i_0}, u)};$$

so (2.16) tells us that

$$M_{C^h} \leqq \beta_1(x^h_{i_0}, u) M^+_{S^h} + \frac{\beta_2(x^h_{i_0}, u)}{1-\beta_1(x^h_{i_0}, u)} M^+_l \quad \forall u \in U^h_{i_0}. \tag{2.17}$$

But, taking in account the definition of β_1, β_2, we have

$$\frac{\beta_2(x^h_{i_0}, u)}{1-\beta_1(x^h_{i_0}, u)} = \frac{1}{\alpha}.$$

Furthermore, by definition of β_1, for all x^h_i, we have

$$0 \leqq \beta_1(x^h_i, u) < 1; \text{ so } \exists 0 < \gamma < 1 / \forall x^h_i, \forall u \in U^h_i, \beta_1(x^h_{i_0}, u) \leqq \gamma.$$

Using this relation in (2.17) we obtain (2.12). □

Remarks on Lemma 2.1.

- As $0 < \gamma < 1$ we also have:

$$M_{C^h} \leqq M^+_{S^h} + \frac{1}{\alpha} M^+_l. \tag{2.18}$$

- If $\forall x^h_i \in C^h$, $\forall u \in U^h_i$, $l(x^h_i, u) = 0$, then

$$M_{C^h} \leqq \gamma M^+_{S^h}. \tag{2.19}$$

- Even if γ is independent of x^h_i and u, it depends on the triangulation Ω_h. We could emphasize this point by writing $\gamma = \gamma(h) < 1$.
- In the proof we have supposed that the set of controls U^h depend on x^h_i; this is why this set is denoted by U^h_i.

b) *Proof of the sufficient condition.* Let w^h be an arbitrary element of W^h and $\bar{w}^h$ such that $\bar{w}^h \equiv M\bar{w}^h$. We define

$$\tilde{w}(x^h_i) = w^h(x^h_i) - \bar{w}(x^h_i) \quad \forall x^h_i \in \Omega^h \tag{2.20}$$

and a partition of the vertices of Ω^h in three disjoint sets:

$$S^h = \{x^h_j / \bar{w}^h(x^h_j) = \phi(x^h_j)\}, \tag{2.21}$$

$$I^h = \left\{ x^h_i \notin S^h / \bar{w}^h(x^h_i) = \min_{z \in Z^h} \left\{ q(x^h_i, z) + \sum_j \lambda'_j(x^h_i, z) \bar{w}^h(x^h_j) \right\} \right\}, \tag{2.22}$$

$$C^h = \left\{ x^h_i \notin I^h \cup S^h / \bar{w}^h(x^h_i) = \min_{u \in U^h} \left\{ \beta_1(x^h_i, u) \sum_j \lambda_j(x^h_i, u) \bar{w}^h(x^h_j) + \beta_2(x^h_i, u) \cdot l(x^h_i, u) \right\} \right\}. \tag{2.23}$$

Remark. From the definition of M and the hypothesis $\bar{w}^h = M\bar{w}^h$ we have

$$\Omega^h = S^h \cup I^h \cup C^h.$$

From (2.5) and (2.21) we obtain

$$w^h(x_i^h) \leqq \bar{w}^h(x_i^h) = \phi(x_i^h) \quad \forall x_i^h \in S^h. \tag{2.24}$$

To achieve our proof we need similar inequalities in C^h and I^h.

Let $\bar{u}_i^h$ be a control for which equality (2.23) holds. So (using the equivalent form (2.2))

$$\frac{\partial \bar{w}^h}{\partial x_f}(x_i^h, \bar{u}_i^h) \cdot \|f(x_i^h, \bar{u}_i^h)\| + l(x_i^h, \bar{u}_i^h) - \alpha \bar{w}^h(x_i^h) = 0 \quad \forall x_i^h \in C^h.$$

But, as $w^h \in W^h$,

$$\frac{\partial w^h}{\partial x_f}(x_i^h, \bar{u}_i^h) \cdot \|f(x_i^h, \bar{u}_i^h)\| + l(x_i^h, \bar{u}_i^h) - \alpha w^h(x_i^{(h)}) \geqq 0 \quad \forall x_i^h \in \Omega^h,$$

we have, $\forall x_i^h \in C^h$,

$$\frac{\partial \tilde{w}^h}{\partial x_f}(x_i^h, \bar{u}_i^h) \cdot \|f(x_i^h, \bar{u}_i^h)\| - \alpha \tilde{w}^h(x_i^h) \geqq 0. \tag{2.25}$$

We can apply Lemma 2.1 with C^h given by (2.23), $S^h \cup I^h$ as its complement and $U_i^h = \{\bar{u}_i^h\}$. (2.12), (2.19), (2.24) tell us that there exists γ, $0 < \gamma < 1$ such that

$$\min_{x_i^h \in C^h} \tilde{w}(x_i^h) \leqq \gamma[\max_{x_i^h \in S^h} \tilde{w}(x_i^h) \vee \max_{x_i^h \in I^h} \tilde{w}(x_i^h)]^+ \leqq \gamma \max_{x_i^h \in I^h} \tilde{w}(x_i^h) \vee 0. \tag{2.26}$$

On the other hand, let $\bar{z}_i^h$ be an impulse control z for which (2.22) holds:

$$\bar{w}^h(x_i^h) = q(x_i^h, \bar{z}_i^h) + \sum_j \lambda_j'(x_i^h, \bar{z}_i^h) \bar{w}^h(x_j^h) \quad \forall x_i^h \in I^h. \tag{2.27}$$

Since $w^h \in W^h$ we have for $\bar{z}_i^h$ as above

$$w^h(x_i^h) \leqq q(x_i^h, \bar{z}_i^h) + \sum_j \lambda_j'(x_i^h, z_i^h) w^h(x_j^h) \quad \forall x_i^h \in \Omega^h;$$

in consequence

$$\tilde{w}^h(x_i^h) \leqq \sum_j \lambda_j'(x_i^h, \bar{z}_j^h) \tilde{w}^h(x_j^h) \quad \forall x_i^h \in I^h. \tag{2.28}$$

With a view to finding an upper bound of $\max_{x_i^h \in I^h} \tilde{w}^h(x_i^h)$ we shall introduce the set of indices

$$I_M^h = \{x_i^h \in I^h / \tilde{w}^h(x_i^h) = \max_{x_j^h \in I^h} \tilde{w}^h(x_j^h)\},$$

and the vertex

$$x_{i_0}^h / \bar{w}^h(x_{i_0}^h) = \min_{x_i^h \in I_M^h} \bar{w}^h(x_i^h).$$

As $x_{i_0}^h \in I_h$ we have, from (2.27),

$$\bar{w}^h(x_{i_0}^h) = q(x_{i_0}^h, \bar{z}_{i_0}^h) + \sum_{j/x_j^h \in I^h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h)\bar{w}^h(x_j^h) + \sum_{j/x_j^h \notin I^h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h)\bar{w}^h(x_j^h). \tag{2.29}$$

We will suppose that

$$\lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) = 0 \quad \forall j / x_j^h \notin I_h. \tag{2.30}$$

In this case (2.28) shows that

$$\tilde{w}^h(x_{i_0}^h) \leqq \sum_{j/x_j^h \in I^h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h)\tilde{w}^h(x_j^h). \tag{2.31}$$

As $x_{i_0}^h \in I_M^h$, (2.31) gives

$$\tilde{w}^h(x_{i_0}^h) \leqq \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h)\tilde{w}^h(x_j^h) + \sum_{\substack{k \neq j \\ k/x_k^h \in I^h}} \lambda_k'(x_{i_0}^h, \bar{z}_{i_0}^h)\tilde{w}^h(x_{i_0}^h).$$

As

$$\sum_{\substack{k \neq j \\ k/x_k^h \in I^h}} \lambda_k'(x_{i_0}^h, \bar{z}_{i_0}^h) = 1 - \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h),$$

we obtain

$$\tilde{w}(x_{i_0}^h)\lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) \leqq \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h)\tilde{w}^h(x_j^h),$$

which implies $\forall j$ such that $\lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) \neq 0$, $\tilde{w}^h(x_j^h) = \tilde{w}^h(x_{i_0}^h)$ i.e. $x_j^h \in I_M^h$.

If we use this result in (2.29), (2.30) taking with account the definition of $x_{i_0}^h$, we have

$$\bar{w}^h(x_{i_0}^h) = q(x_{i_0}^h, \bar{z}_{i_0}^h) + \sum_{j/x_j^h \in I_m^h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h)\bar{w}^h(x_j^h) \geqq q(x_{i_0}^h, \bar{z}_{i_0}^h) + \bar{w}^h(x_{i_0}^h),$$

i.e. $q(x_{i_0}^h, \bar{z}_{i_0}^h) \leqq 0$, contradicting our initial hypothesis $q > 0$.

This contradiction comes from supposition (2.30). Then there exists at least a vertex $x_j^h \notin I_h$ such that $\lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) > 0$.

We return to (2.28) and we have

$$\begin{aligned}\tilde{w}(x_{i_0}^h) &\leqq \sum_{j/x_j \in I_h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) \cdot \tilde{w}^h(x_j^h) + \sum_{j/x_j \notin I_h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) \cdot \tilde{w}^h(x_j^h) \\ &\leqq \left(\sum_{j/x_j \in I_h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) \right) \tilde{w}^h(x_{i_0}^h) + \left(\sum_{j/x_j^h \notin I^h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) \right) \max_{x_j^h \in C^h \cup S^h} \tilde{w}^h(x_j^h),\end{aligned}$$

and, as

$$1 - \sum_{j/x_j^h \in I^h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) = \sum_{j/x_j \notin I^h} \lambda_j'(x_{i_0}^h, \bar{z}_{i_0}^h) > 0,$$

we obtain

$$\max_{x_j^h \in I^h} \tilde{w}(x_j^h) = \tilde{w}^h(x_{i_0}^h) \leqq \max_{x_j^h \in C^h \cup S^h} \tilde{w}(x_j^h) \leqq (\max_{x_j^h \in C^h} \tilde{w}^h(x_j^h)) \vee 0. \tag{2.32}$$

So, after (2.24), (2.32), if $\tilde{w}^h$ has a positive maximum it will follow that $\max_{x_j^h \in C^h} \tilde{w}^h(x_j^h) > 0$. But from (2.26), (2.32),

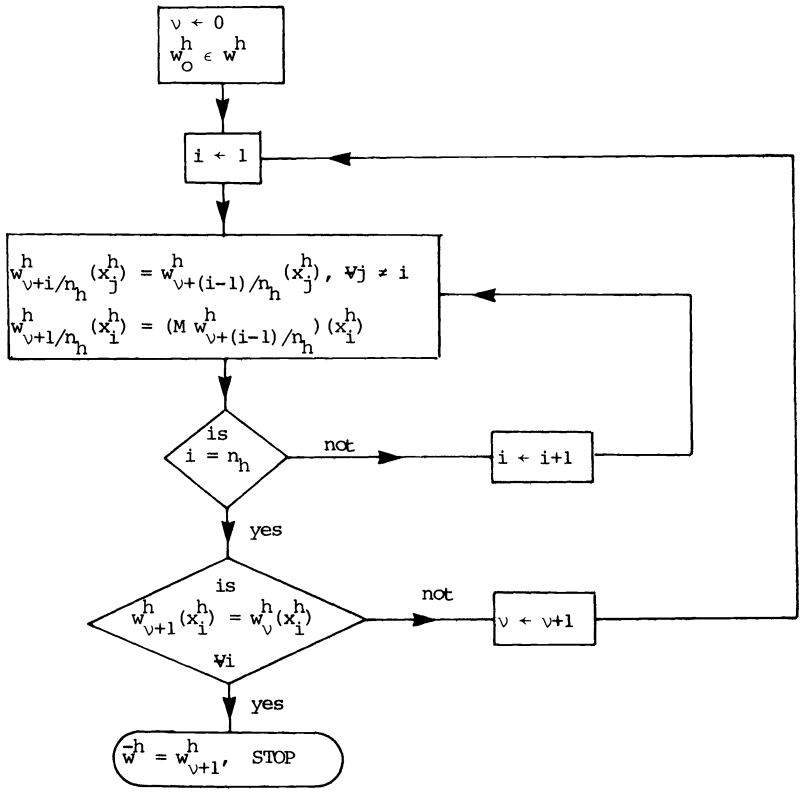
$$\max_{x_i^h \in C^h} \tilde{w}^h(x_i^h) \leqq \gamma \max_{x_j^h \in I^h} \tilde{w}^h(x_j^h) \vee 0 \leqq \gamma \max_{x_j^h \in C^h} \tilde{w}^h(x_j^h) \vee 0. \tag{2.33}$$

Then, if $\max_{x_j^h \in C^h} \tilde{w}^h(x_j^h) > 0$ we have in (2.33) $(1-\gamma)\max_{x_j^h \in C^h} \tilde{w}^h(x_j^h) \leqq 0$, i.e. $\max_{x_j^h \in C^h} \tilde{w}^h(x_j^h) \leqq 0$, contradicting our previous supposition.

We conclude $\tilde{w}^h(x_i^h) \leqq 0 \ \forall x_i^h \in \Omega^h$, i.e. $\bar{w}^h(x_i^h) \geqq w^h(x_i^h) \ \forall x_i^h \in \Omega^n$. As w^h is an arbitrary element of W^h, the sufficiency condition is proved.

2.5. Algorithm to compute $\bar{w}^h$. We will take advantage of the characterization of $\bar{w}^h$ given in Theorem 2.2 to define an algorithm generating an increasing sequence of functions w_ν^h, convergent to a function satisfying $Mw^h = w^h$, i.e., convergent to $\bar{w}^h$.

ALGORITHM 2.1.



This algorithm gives the solution of problem (P_h) in the following sense.

THEOREM 2.3. *Algorithm* 2.1 *terminates at $\bar{w}^h$ in a finite number of steps or generates a sequence $\{w_\nu^h\}$ convergent to $\bar{w}^h$.*

Proof. If the algorithm terminates in a finite number of steps ($\bar{\nu}$), that means $w_{\bar{\nu}}^h$ is not modified in the last loop, i.e. $w_{\bar{\nu}}^h(x_i^h) = Mw_{\bar{\nu}}^h(x_i^h)$, $\forall i$, then in virtue of Theorem 2.2, $w_{\bar{\nu}}^h = \bar{w}^h$.

If this is not the case, since $w_0^h \in W^h$, it results by induction that

$$w^h_{\nu+1/n_h} \leqq w^h \quad \text{and} \quad w_\nu^h \leqq \cdots \leqq w^h_{\nu+(i-1)/n_h} \leqq w^h_{\nu+i/n_h} \cdots \leqq w^h_{\nu+1} \leqq \cdots \leqq \phi;$$

then there exists $\check{w}^h$ such that

$$\lim_{\nu\to\infty} w^h_{\nu+i/n^h} = \check{w}^h \quad \forall i = 1, \cdots, n_h.$$

We remark that by definition of operator M we have $w^h \in W^h \Leftrightarrow w^h \leqq Mw^h$; $w^h \leqq \tilde{w}^h \Leftrightarrow Mw^h \leqq M\tilde{w}^h$; hence, as Algorithm 2.1 imposes $w^h_{\nu+i/n_h}(x^h_i) = Mw^h_{\nu+(i-1)/n_h}(x^h_i)$, from (2.34) we obtain $\check{w}^h(x^h_i) = M\check{w}^h(x^h_i)$, i.e. by Theorem 2.2, $\lim_{\nu\to\infty} w^h_\nu = \bar{w}^h$. □

Remarks on Algorithm 2.1.

a) The algorithm only needs the values of functions $w^h_{\nu+(i-1)/n_h}$ at points $a_i(x)$ and $x^h_i + g(x^h_i, z)$ to compute $w^h_{\nu+i/n_h}(x^h_i)$.

This property allows the application of this algorithm on computers with small central memories (minicomputers).

b) As Theorem 2.3 shows, the convergence of the algorithm does not depend on the order of vertices x^h_j in triangulation Ω^h; however a careful choice of that order may allow:

1. An easy retrieval of the information needed for the computation, from the mass memory to the central memory of the computer.
2. An acceleration of the convergence of the algorithm.

c) The algorithm implies the fulfillment with equality of at least one of the constraints (2.2)′, (2.4)′, (2.5) in each iteration. In practice, the convergence will not be lost if that saturation is omitted in some steps.

3. Convergence of discrete solutions $\bar{w}^h(x)$ to the optimal cost functions $V(x)$.

3.1. Preliminary comments. The result will be achieved in two steps. In the first step we will restrict ourselves to consider only stopping time problems (P_S). $V_S(x)$ will be, in this case, our optimal cost function and to show $V_S(x) \leqq \lim_{\|h\|\to 0} \bar{w}^h(x)$ we introduce the same techniques used in [9]. But the discrete maximum principle will be essential to show $V_S(x) \geqq \overline{\lim}_{\|h\|\to 0} \bar{w}^h(x)$. Furthermore, this DMP implicitly gives the stability of the method.

In a second step we consider the original problem (P), with continuous and impulse controls. We introduce a suitable sequence of stopping-time problems which are able to define a sequence of solutions convergent to $V(x)$.

3.2. Convergence in stopping time problems (P_S). In the dynamics (1.1) of the system

$$\frac{dy}{dt} = f(y, u), \qquad y(0) = x, \tag{3.1}$$

we consider a constant value of $u \in U$, and we look for

$$V_S(x) = \inf_{0\leqq\theta} \int_0^\theta e^{-\alpha s} l(y(s), u)\, ds + e^{-\alpha\theta} \psi(y(\theta)). \tag{3.2}$$

If we suppose ψ Lipschitzian (L_ψ its Lipschitz constant) we obtain following what was done in Theorem 1.1 and Theorem 1.2 that $V_S(x)$ is Lipschitzian (having L_S as Lipschitz constant: $L_S = L_l/(\alpha - L_f) + L\psi$) and it is the maximum element of the set:

$$W_S = \{w: \Omega \to R / (3.4), (3.5), (3.6)\}, \tag{3.3}$$

$$w(x) \text{ is Lipschitzian,} \tag{3.4}$$

$$\frac{\partial w(x)}{\partial x} \cdot f(x, u) + l(x, u) - \alpha w(x) \geqq 0, \tag{3.5}$$

$$w(x) \leqq \psi(x) \quad \forall x \in \Omega. \tag{3.6}$$

As in § 2.2 we introduce the discretization procedure and we pose the approximate problem (over the same triangulations of problem (P^h)):

(P_s^h): Find the maximum element $\bar{w}_S^h$ of the set

$$W_S^h = \{w^h : \Omega^h \to R / (3.8), (3.9), (3.10)\}, \tag{3.7}$$

$$w^h \text{ is a linear finite element characterized by the values } w^h(x_i^h), \tag{3.8}$$

$$\frac{\partial w^h}{\partial x_f}(x_i^h; u) \cdot \|f(x_i^h, u)\| + l(x_i^h, u) - \alpha w^h(x_i^h) \geqq 0 \quad \forall x_i^h \text{ vertex of } \Omega^h, \tag{3.9}$$

$$w^h(x_i^h) \leqq \psi(x_i^h) \quad \forall x_i^h \text{ vertex of } \Omega^h. \tag{3.10}$$

Similarly to what was done in § 2.3 we show that (P_S^h) has an unique solution $\bar{w}_S^h$ given by

$$\bar{w}_S^h(x_i^h) = \min\left(\psi(x_i^h), \beta_1(x_i^h, u) \sum_j \lambda_j(x_i^h, u) \cdot \bar{w}_S^h(x_j^h) + \beta_2(x_i^h, u) : l(x_i^h, u)\right), \tag{3.11}$$

and relation (3.9) can be equivalently written

$$w^h(x_i^h) \leqq \beta_1(x_i^h, u) : \sum_j \lambda_j(x_i^h, u) \cdot w^h(x_j^h) + \beta_2(x_i^h, u) l(x_i^h, u) \quad \forall x_i^h \in \Omega^h. \tag{3.12}$$

To show the uniform convergence of $\bar{w}_S^h$ to V_S we introduce three hypotheses:

H_1) The functions $f(x, u)$ and $l(x, u)$ can be extended to some open set which contains $\bar{\Omega}$ in such a manner that the Lipschitz continuity is preserved (and the Lipschitz constants L_f and L_l are unaltered).

H_2) There exist $\eta > 0$ and an injective continuous differentiable mapping $A_\eta : \Omega \to R^n$ such that

$$\exists c_1 > 0 \quad \text{such that} \quad \left\| \frac{\partial A_\eta(x)}{\partial x} - I \right\| \leqq \frac{c_1}{2} \eta \quad \forall x \in \Omega, \tag{3.13}$$

with I the unit matrix of $\mathbb{R}^{n \times n}$;

$$\|A_\eta(x) - x\| \leqq x \quad \forall x \in \Omega, \tag{3.14}$$

$$\Omega + B_{\eta/2} \subset A_\eta(\Omega), \quad \text{with } B_{\eta/2} = \left\{ y \in \mathbb{R}^n \Big/ \|y\| \leqq \frac{\eta}{2} \right\}. \tag{3.15}$$

Remark. (3.13) implies the existence and continuity of $\partial A_\eta^{-1}/\partial x$; furthermore

$$\left\| \frac{\partial A_\eta^{-1}(x)}{\partial x} - I \right\| \leqq c_1 \eta \quad \forall \eta \leqq \frac{1}{c_1}. \tag{3.13'}$$

H_3) For each vertex x_i^h on the boundary of Ω^h there exists $\varepsilon_i^h > 0$ such that for $0 \leqq \varepsilon \leqq \varepsilon_1^h$ the segment $x_i^h + \varepsilon f(x_i^h, u)$ belongs to Ω^h; so $\partial w^h / \partial x_f(x_i^h, u)$ is well defined (Fig. 5).

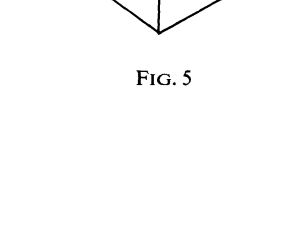


FIG. 5

The convergence property asserted in Theorem 3.1 is an immediate consequence of the following two lemmas (whose proofs will be given after the theorem).

LEMMA 3.1. *If* H_1), H_2), H_3) *hold there exists a real single valued function* $G_1(\eta, \rho, \|h\|)$ *such that*

$$V_S(x_i^h) - G_1(\eta, \rho, \|h\|) \leqq \bar{w}_S^h(x_i^h) \quad \forall x_i^h \in \Omega^h, \tag{3.16}$$

with $0 < \rho \leqq \eta/4$, ρ *a regularization parameter and*

$$\begin{aligned} G_1(\eta, \rho, \|h\|) &= \eta L_S + \rho L_S(1 + c_1\eta) + L_\psi(\eta + \rho) \\ &+ \frac{1}{\alpha}\left\{ \eta[L_S(c_1 M_f + L_f) + L_l] \right. \\ &\left. + \rho[L_S(1 + c_1\eta)L_f + L_l] + [c_3 L_S(1 + c_1\eta)M_f]\frac{\|h\|}{\rho}\right\}, \end{aligned} \tag{3.17}$$

with c_3 *constant.*

LEMMA 3.2. *With the same hypothesis as Lemma* 3.1, *there exists a real single valued function* $G_2(\eta, \rho, \|h\|)$ *such that*

$$\bar{w}_S^h(x_i^h) \leqq V_S(x_i^h) + G_2(\eta, \rho, \|h\|) \quad \forall x_i^h \in \Omega^h, \tag{3.18}$$

with $0 < \rho < \eta/4$ *and*

$$\begin{aligned} G_2(\eta, \rho, \|h\|) &= \eta(3L_S + L_\psi) + \rho[3L_S(1 + c_1\eta) + L_\psi] \\ &+ \frac{1}{\alpha}\left\{ \eta[L_S(c_1 M_f + L_f) + L_l] \right. \\ &\left. + \rho[L_S(1 + c_1\eta)L_f + L_l] + c_3 L_S M_f \frac{\|h\|}{\rho}(1 + c_1\eta)\right\}. \end{aligned} \tag{3.19}$$

THEOREM 3.1. *If* H_1), H_2), H_3) *hold the functions* $\bar{w}_S^h(x)$ *converge uniformly to* $V_S(x)$.

Proof. In Lemmas 3.1 and 3.2 if we put $\rho = \|h\|^{1/2}$, $\eta = 4\rho$ we obtain, from (3.16), (3.17), (3.18) and (3.19) a positive constant C such that $\forall x_i^h$(vertex) of Ω^h:

$$V_S(x_i^h) - C\|h\|^{1/2} \leqq \bar{w}_S^h(x_i^h) \leqq V_S(x_i^h) + C\|h\|^{1/2}. \tag{3.20}$$

We now consider an arbitrary point $x \in \Omega^h$. As we know we can express x as

$$x = \sum_j \hat{\lambda}_j x_j^h, \quad \hat{\lambda}_j \geqq 0, \quad \sum_j \hat{\lambda}_j = 1;$$

so we can obtain

$$\begin{aligned} V_S(x) - \bar{w}_S^h(x) &= \left(V_S(x) - \sum_j \hat{\lambda}_j V_S(x_j^h)\right) + \left(\sum_j \hat{\lambda}_j V_S(x_j^h) - \sum_j \hat{\lambda}_j \bar{w}^h(x_j^h)\right) \\ &= \sum_j \hat{\lambda}_j (V_S(x) - V_S(x_j^h)) + \sum \hat{\lambda}_j (V_S(x_j^h) - \bar{w}_S^h(x_l^h)) \\ &\leqq \sum_j \hat{\lambda}_j \cdot L_S\|x - x_j^h\| + \sum_j \hat{\lambda}_j C\|h\|^{1/2} \\ &\leqq L_S\|h\| + C\|h\|^{1/2}. \end{aligned} \tag{3.21}$$

In the same way

$$V_S - \bar{w}_S^h(x) \geqq -(L_S\|h\| + C\|h\|^{1/2}),$$

which gives together with (3.21)

$$\max_{x\in\Omega^h} |\bar{w}_S^h(x) - V_S(x)| \leqq L_S\|h\| + C\|h\|^{1/2}$$

and the proof is achieved. □

Proof of Lemma 3.1. By a constructive device we shall obtain $V_S^h \in W_S^h$ such that

$$V_S(x_i^h) - G_1(\eta, \rho, \|h\|) \leqq V_S^h(x_i^h) \quad \forall x_i^h \in \Omega^h; \tag{3.22}$$

then, as $V_S^h(x_i^h) \leqq \bar{w}_S^h(x_i^h)$, (3.16) will hold.

The construction of V_S^h needs four steps:

a) Using H_2) we introduce the Lipschitzian function

$$V_\eta(x) = V_S(A_\eta^{-1}(x)) \quad \forall x \in A_\eta(\Omega). \tag{3.23}$$

$\partial V_\eta/\partial x$ exists a.e. in $A_\eta(\Omega)$. So, from $\partial V_\eta/\partial x = (\partial/\partial A_\eta^{-1}(x)) V_S(A_\eta^{-1}(x)) \cdot \partial A_\eta^{-1}(x)/\partial x$, using (3.13) we have

$$\left\|\frac{\partial V_\eta}{\partial x}(x)\right\| \leqq L_S(1 + c_1\eta) \quad \text{a.e. } x \in A_\eta(x). \tag{3.24}$$

We remark that the domain of V_η contains the set $\Omega + B_{\eta/2}$ (from definition of V_η and (3.15)).

We compute some bounds concerning $V_\eta(x)$

$$|V_S(x) - V_\eta(x)| = |V_S(x) - V_S(A_\eta^{-1}(x))| \leqq L_S\|x - A_\eta^{-1}(x)\|; \tag{3.25}$$

but, after (3.15), $x \in \Omega \Rightarrow \exists y \in \Omega$ such that $x = A_\eta(y)$; so $\|x - A_\eta^{-1}(x)\| = \|A_\eta(y) - y\|$ which, by (3.14) is bounded by η. Then, in (3.25) we have

$$|V_S(x) - V_\eta(x)| \leqq L_S \cdot \eta \quad \forall x \in \Omega. \tag{3.26}$$

On the other hand, thanks to (3.6), (3.14),

$$\begin{aligned} V_\eta(x) - \psi(x) &= V_S(A_\eta^{-1}(x)) - \psi(A_\eta^{-1}(x)) + (\psi(A_\eta^{-1}(x)) - \psi(x)) \\ &\leqq L_\psi\|A_\eta^{-1}(x) - x\| \leqq L_\psi \cdot \eta \quad \forall x \in A_\eta(x). \end{aligned} \tag{3.27}$$

If we compute the first term of (3.5) with $V_\eta(x)$ in place of $w(x)$, we have

$$\begin{aligned} \frac{\partial V_\eta(x)}{\partial x} \cdot f(x, u) + l(x, u) - \alpha V_\eta(x) &= \frac{\partial V_S}{\partial x}(A_\eta^{-1}(x)) f(A_\eta^{-1}(x), u) \\ &\quad + l(A_\eta^{-1}(x), u) - \alpha V_S(A_\eta^{-1}(x)) + \gamma_\eta(x), \end{aligned} \tag{3.28}$$

with

$$\gamma_\eta(x) = \frac{\partial V_S}{\partial x}(A_\eta^{-1}(x))\left(\frac{\partial A_\eta^{-1}(x)}{\partial x} - I\right) f(x, u)$$

$$+ \frac{\partial V_S}{\partial x}(A_\eta^{-1}(x))(f(x, u) - f(A_\eta^{-1}(x), u)) + l(x, u) - l(A_\eta^{-1}(x), u).$$

Using H_2), (3.13), (3.14), we have

$$|\gamma_\eta(x)| \leqq L_S c_1 \eta M_f + L_S L_f \eta + L_l \eta.$$

Also, in (3.28) we obtain (recalling (3.5))

$$\frac{\partial V_\eta(x)}{\partial x} f(x,u) + l(x,u) - \alpha V_\eta(x) \geqq -\eta[L_S(c_1 M_f + L_f) + L_l] \quad \forall x \in A_\eta(x). \tag{3.29}$$

b) *Regularization of* $V_\eta(x)$. Let $\beta_1(\cdot)$ be a function such that

$$\beta_1(\cdot) \in C^\infty(\mathbb{R}^n), \quad \beta_1(x) \geqq 0 \ \forall x, \quad \operatorname{supp} \beta_1 \subset B_1, \quad \int_{\mathbb{R}^n} \beta_1(x)\, ds = 1.$$

We define $\beta_\rho(x) = (1/\rho^n)\beta_1(x/\rho)$, $\rho \in \mathbb{R}^+$. As $\forall \rho < \eta/4$, dom $V_\eta \supset \Omega + B_{\eta/2}$, we can define

$$V_{\eta,\rho}(x) = (V_\eta * \beta_\rho)(x) \quad \forall x \in \Omega + B_{\eta/4}.$$

We remark that $V_{\eta,\rho}$ is infinitely differentiable; furthermore

$$\left\| \frac{\partial}{\partial x} V_{\eta,\rho}(x) \right\| \leqq \left\| \frac{\partial V_\eta}{\partial x} \right\|; \quad \text{so}$$

$$\left\| \frac{\partial}{\partial x} V_{\eta,\rho}(x) \right\| \leqq L_S(1 + c_1\eta) \quad \forall x \in \Omega + B_{\eta/4}, \tag{3.30}$$

$$\left\| \frac{\partial^2}{\partial x_i\, \partial x_j} V_{\eta,\rho}(x) \right\| \leqq c_2 \cdot \frac{1}{\rho} \left\| \frac{\partial V_\eta}{\partial x} \right\| \leqq c_2 \frac{1}{\rho} L_S(1 + c_1\eta), \tag{3.31}$$

with c_2 a constant depending on $\beta_1(\cdot)$.

As before we compute some bounds concerning $V_{\eta,\rho}$. Using (3.24), (3.26) and the properties of the convolution operator, we have

$$\begin{aligned} |V_S(x) - V_{\eta,\rho}(x)| &\leqq |V_S(x) - V_\eta(x)| + |V_\eta(x) - V_{\eta,\rho}| \\ &\leqq L_S\eta + \rho L_S(1 + c_1\eta) \quad \forall x \in \Omega. \end{aligned} \tag{3.32}$$

Using (3.27) gives

$$V_{\eta,\rho}(x) - \psi(x) = ((V_\eta - \psi) * \beta_\rho)(x) + (\psi * \beta_\rho - \psi)(x) \leqq L_\psi \cdot \eta + L_\psi \cdot \rho. \tag{3.33}$$

Now in (3.5)

$$\begin{aligned} &\frac{\partial V_{\eta,\rho}(x)}{\partial x} \cdot f(x,u) + l(x,u) - \alpha V_{\eta,\rho}(x) \\ &\quad = \left(\frac{\partial V_\eta}{\partial x}(\cdot) \cdot f(\cdot, u) + l(\cdot, u) - \alpha V_\eta(\cdot) * \beta_\rho(\cdot) \right)(x) + \gamma_\rho(x) \end{aligned} \tag{3.34}$$

with

$$\gamma_\rho(x) = \int_{B_\rho} \left\{ \frac{\partial V_\eta}{\partial x}(x-y)[(f(x) - f(x-y))] + (l(x) - l(x-y)) \right\} B_\rho(y)\, dy,$$

and from H_2) and (3.24),

$$|\gamma_\rho(x)| \leqq (L_S(1 + c_1\eta)L_f + L_l)\rho.$$

It follows from (3.29) and (3.34) that

$$\begin{aligned} &\frac{\partial}{\partial x} V_{\eta,\rho}(x) \cdot f(x,u) + l(x,u) - \alpha V_{\eta,\rho}(x) \\ &\quad \geqq -\eta(L_S(c_1 M_f + L_f) + L_l) - \rho(L_S(1 + c_1\eta)L_f + L_l) \quad \forall x \in \Omega + B_{\eta/4}. \end{aligned} \tag{3.35}$$

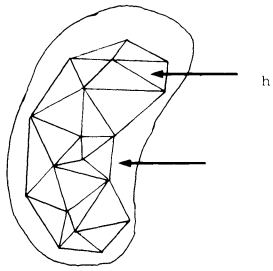


FIG. 6

c) *Discretization.* In Ω^h we define $V^h_{\eta,\rho}$ to be the linear interpolation of $V_{\eta,\rho}$, which coincides with $V_{\eta,\rho}$ at the vertices x^h_i of Ω^h, i.e.

$$V^h_{\eta,\rho}(x^h_i) = V_{\eta,\rho}(x^h_i) \quad \forall x^h_i \in \Omega^h. \tag{3.36}$$

Some properties of $V^h_{\eta,\rho}(x^h_i)$ are, by (3.32),

$$|V_S(x^h_i) - V^h_{\eta,\rho}(x^h_i)| \leqq L_S\eta + \rho L_S(1 + c_1\eta) \quad \forall x^h_i \in \Omega^h; \tag{3.37}$$

and by (3.33),

$$V^h_{\eta,\rho}(x^h_i) - \psi(x^h_i) = V_{\eta,\rho}(x^h_i) - \psi(x^h_i) \leqq L_\psi(\eta + \rho) \quad \forall x^h_i \in \Omega^h. \tag{3.38}$$

On the other hand, from (3.31),

$$\begin{aligned} &\left| \frac{\partial V^h_{\eta,\rho}}{\partial x_f}(x^h_i, u) \cdot \|f(x^h_i, u)\| - \frac{\partial V_{\eta,\rho}}{\partial x}(x^h_i) \cdot f(x^h_i, u) \right| \\ &\quad \leqq c_4 \left\| \frac{\partial^2}{\partial x_i\, \partial x_j} V_{\eta,\rho} \right\| \|h'\| M_f \leqq c_3 \frac{1}{\rho} L_S(1 + c_1\eta)\|h\| M_f; \end{aligned} \tag{3.39}$$

in consequence (3.35) and (3.39) allow us to write

$$\begin{aligned} &\frac{\partial V^h_{\eta,\rho}}{\partial x_f}(x^h_i, u) \cdot \|f(x^h_i, u)\| + l(x^h_i, u) - \alpha V^h_{\eta,\rho}(x^h_i) \\ &\quad = \frac{\partial V_{\eta,\rho}}{\partial x}(x^h_i) \cdot f(x^h_i, u) + l(x^h_i, u) - \alpha V_{\eta,\rho}(x^h_i) \\ &\quad\quad + \left(\frac{\partial V^h_{\eta,\rho}}{\partial x_f}(x^h_i, u) \cdot \|f(x^h_i, u)\| - \frac{\partial V_{\eta,\rho}}{\partial x}(x^h_i) f(x^h_i, u) \right) \\ &\quad \geqq -\eta(L_S(L_f + c_1 M_f) + L_l) - \rho(L_S(1 + c_1\eta)L_f + L_l) \\ &\quad\quad - c_3 L_S(1 + c_1\eta) M_f \frac{\|h\|}{\rho} \quad \forall x^h_i \in \Omega^h. \end{aligned} \tag{3.40}$$

d) *Definition of V^h_S.* We define $\forall x \in \Omega^h$ the function V^h_S as the linear interpolation of the values of the vertex given by

$$\begin{aligned} V^h_S(x^h_i) &= V^h_{\eta,\rho}(x^h_i) - L_\psi(\eta + \rho) \\ &\quad - \frac{1}{\alpha}\left\{ (L_S(c_1 M_f + L_f) + L_l)\eta + (L_S(1 + c_1\eta)L_f + L_l)\rho + c_3 L_S(1 + c_1\eta) M_f \right\} \frac{\|h\|}{\rho}. \end{aligned} \tag{3.41}$$

We will show that

$$V_S^h \leqq \bar{w}_S^h. \tag{3.42}$$

In fact, after (3.38), (3.40) and (3.41) we can easily obtain,

$$\frac{\partial V_S^h}{\partial x_f}(x_i^h, u) \cdot \|f(x_i^h, u)\| + l(x_i^h, u) - \alpha V_S^h(x_i^h) \geqq 0, \tag{3.43}$$

$$V_S^h(x_i^h) \leqq \psi(x_i^h) \quad \forall x_i^h \text{ vertex of } \Omega^h; \tag{3.44}$$

so, $V_S^h \in W_S^h$, and by definition of $\bar{w}_S^h$ we have $V_S^h(x_i^h) \leqq \bar{w}_S^h(x_i^h)\ \forall x_i^h \in \Omega^h$. □

Furthermore, from (3.17), (3.37) and (3.41),

$$\begin{aligned} |V_S(x_i^h) - V_S^h(x_i^h)| &\leqq |V_S^h(x_i^h) - V_{\eta,\rho}^h(x_i^h)| + |V_{\eta,\rho}^h(x_i^h) - V_S(x_i^h)| \\ &\leqq G_1(\eta, \rho, \|h\|); \end{aligned}$$

so (3.22) is proved, as is Lemma 3.1. □

Proof of Lemma 3.2. We recall that function V_S verifies (see [9, p. 3])

$$\frac{\partial V_S}{\partial x}(x) f(x, u) + l(x, u) - \alpha V_S(x) = 0 \quad \text{a.e. } x \in C \tag{3.45}$$

with

$$C = \{x \in \Omega \,|\, V_S(x) < \psi(x)\}, \tag{3.46}$$

$$S = \{x \in \Omega \,|\, V_S(x) = \psi(x)\}. \tag{3.47}$$

In the following we will consider the functions V_η, $V_{\eta,\rho}$ and $V_{\eta,\rho}^h$ as they were defined in the proof of Lemma 3.1. We put

$$S_\eta = \{x \in A_\eta(\Omega) \,|\, d(x, S) \leqq \eta\}, \tag{3.48}$$

$$C_\eta = \{x \in A_\eta(\Omega) \,|\, x \notin S_\eta\}. \tag{3.49}$$

Let us consider V_η and its behaviour in S_η, C_η. If $x \in C_\eta$, it follows that $d(x, S) > \eta$; so recalling that $\|x - A_\eta^{-1}(x)\| < \eta$ we have that $d(A_\eta^{-1}(x), S) > 0$, implying $A_\eta^{-1}(x) \in C$. In consequence, using (3.28) and (3.45) we have

$$\begin{aligned} &\left| \frac{\partial V_\eta(x)}{\partial x} \cdot f(x, u) + l(x, u) - \alpha V(x) \right| \\ &\quad \leqq |\gamma_\eta(x)| \leqq \eta[L_S(c_1 M_f + L_f) + L_l] \quad \text{a.e. } x \in C_\eta. \end{aligned} \tag{3.50}$$

If $x \in S_\eta$, $\exists\{x_\xi\}$, $x_\xi \in S$, $\xi = 1, 2, \cdots$ such that $\|x_\xi - x\| \to d(x, S)$ if $\xi \to \infty$. Using (3.14), (3.47), we obtain

$$\begin{aligned} |V_\eta(x) - \psi(x)| &= |V_S(A_\eta^{-1}(x)) - \psi(x)| \\ &= |V_S(A_\eta^{-1}(x)) - V_S(x)| + |V_S(x) + V_S(x_\xi)| \\ &\quad + |V_S(x_\xi) - \psi(x_\xi)| + |\psi(x_\xi) - \psi(x)| \\ &\leqq L_S \eta + L_S \|x - x_\xi\| + L_\psi \|x - x_\xi\|. \end{aligned}$$

Finally for $\xi \to \infty$

$$|V_\eta(x) - \psi(x)| \leqq \eta(2L_S + L_\psi) \quad \forall x \in S_\eta. \tag{3.51}$$

Now we consider $V_{\eta,\rho}(x)$. We define

$$S_\rho = \{x \in \Omega + B_{\eta/4} | d(x, S_\eta) \leqq \rho\}, \tag{3.52}$$

$$C_\rho = \{x \in \Omega + B_{\eta/4} | x \notin S_\rho\}. \tag{3.53}$$

If $x \in C_\rho$ it follows that $(x-y) \in C_\eta$, for all y such that $\|x-y\| \leqq \rho$. So, using (3.34), (3.50), we have

$$\begin{aligned} &\left| \frac{\partial V_{\eta,\rho}(x)}{\partial x} \cdot f(x,u) + l(x,u) - \alpha V_{\eta,\rho}(x) \right| \\ &\leqq (\eta(L_S(c_1 M_f + L_f) + L_l) * \beta_\rho)(x) + |\gamma_\rho(x)| \\ &\leqq \eta(L_S(c_1 M_f + L_f) + L_l) + (L_S(1 + c_1\rho) + L_l) \quad \forall x \in C_\rho. \end{aligned} \tag{3.54}$$

If $x \in S_\rho$ we use a sequence $\{x_\xi\} \subset S_\eta / \lim_{\xi \to \infty} \|x_\xi - x\| = d(x, S_\eta) \leqq \rho$; we obtain (3.24), (3.51). Now

$$\begin{aligned} |V_{\eta,\rho}(x) - \psi(x)| &\leqq |V_{\eta,\rho}(x) - V_\eta(x)| + |V_\eta(x) - V_\eta(x_\xi)| + |V_\eta(x_\xi) - \psi(x_\xi)| + |\psi(x_\xi) - \psi(x)| \\ &\leqq L_S(1 + c_1\eta)\rho + L_S(1 + c_1\eta)\|x - x_\xi\| + \eta(2L_S + L_\psi) + L_\psi\|x - x_\xi\|, \end{aligned}$$

and for $\xi \to \infty$

$$|V_{\eta,\rho}(x) - \psi(x)| \leqq \eta(2L_S + L_\psi) + \rho(2L_S(1 + c_1\eta) + L_\psi) \quad \forall x \in S_\rho. \tag{3.55}$$

Finally concerning $V^h_{\eta,\rho}$ we define S_h, C_h, sets of vertices of Ω^h, $(S_h \cap C_h = \varnothing)$,

$$S_h = \{x_i^h / x_i^h \in S_\rho\}, \tag{3.56}$$

$$C_h = \{x_i^h / x_i^h \in C_\rho\}. \tag{3.57}$$

If $x_i^h \in S_h$, as $V^h_{\eta,\rho}(x_i^h) = V_{\eta,\rho}(x_i^h)$ we obtain, from (3.55),

$$|V^h_{\eta,\rho}(x_i^h) - \psi(x_i^h)| \leqq \eta(2L_S + L_\psi) + \rho(2L_S(1 + c_1\eta) + L_\psi). \tag{3.58}$$

If $x_i^h \in C_h$ we have from (3.39), (3.54),

$$\begin{aligned} &\left| \frac{\partial V^h_{\eta,\rho}(x_i^h, u)}{\partial x_f} \|f(x_i^h, u)\| + l(x_i^h, u) - \alpha V^h_{\eta,\rho}(x_i^h) \right| \\ &\leqq \left| \frac{\partial V_{\eta,\rho}(x_i^h)}{\partial x} \cdot f(x_i^h, u) + l(x_i^h, u) - \alpha V_{\eta,\rho}(x_i^h) \right| \\ &\quad + c_3 \frac{L_S}{\rho}(1 + c_1\eta) M_f \|h\| \\ &\leqq \eta(L_S(c_1 M_f + L_f) + L_l) \\ &\quad + \rho(L_S(1 + c_1\eta) L_f + L_l) + c_3 \frac{L_S}{\rho}(1 + c_1\eta) M_f \|h\|. \end{aligned} \tag{3.59}$$

Let us define $\tilde{w}^h = \bar{w}_S^h - V_{\eta,\rho}^h$. As $\bar{w}^h \in W_S^h$, we obtain by (3.9), (3.10), (3.58), (3.59),

$$\tilde{w}^h(x_i^h) \leqq \psi(x_i^h) - V_{\eta,\rho}^h(x_i^h) \leqq \eta(2L_S + L_\psi) + \rho(2L_S(1+c_1\eta) + L_\psi) \quad \forall x_i^h \in S_h, \tag{3.60}$$

$$\frac{\partial \tilde{w}^h(x_i^h, u)}{\partial x_f} \cdot \|f(x_i^h, u)\| - \alpha \tilde{w}^h(x_i^h)$$

$$+\left\{\eta[L_S(c_1 M_f + L_f) + L_l] + \rho(L_S(1+c_1\eta)L_f + L_l) + c_3 \frac{L_S}{\rho}(1+c_1\eta)M_f\|h\|\right\} \tag{3.61}$$

$$\geqq 0 \quad \forall x_i^h \in C^h.$$

Using (3.60), (3.61) and Lemma 2.1 (Discrete maximum principle), we can ensure that

$$\tilde{w}^h(x_i^h) \leqq \eta(2L_S + L_\psi) + \rho(2L_S(1+c_1\eta) + L_\psi)$$

$$+\frac{1}{\alpha}\left\{\eta(L_S(c_1 M_f + L_f) + L_l) + \rho(L_S(1+c_1\eta)L_f + L_l) + c_3 \frac{1}{\rho} L_S(1+c_1\eta)M_f\|h\|\right\} \tag{3.62}$$

$$\forall x_i^h \in C_h.$$

Finally (3.37), (3.60), (3.61) give us

$$\bar{w}_S^h(x_i^h) = V_{\eta,\rho}^h(x_i^h) + \tilde{w}^h(x_i^h) \leqq \tilde{w}^h(x_i^h) + V_S(x_i^h) + \eta L_S + \rho L_S(1+c_1\eta)$$

$$\leqq V_S(x_i^h) + G_2(\eta, \rho, \|h\|) \quad \forall x_i^h \in \Omega^h. \qquad \square$$

3.3. Convergence in the framework of problem (P). We begin by a lemma similar to Lemma 3.1 (for the proof see [12]).

LEMMA 3.3. *If* i), ii) *and* iii) *of* § 1.1 *and* H_1), H_2), H_3) *hold, there exists a single real function* $G_3(\eta, \rho, \|h\|)$ *defined for* $\rho \leqq \eta/4$ *such that*:

$$\text{a)} \quad V(x_i^h) \leqq \bar{w}^h(x_i^h) + G_3(\eta, \rho, \|h\|) \quad \forall x_i^h \in \Omega^h, \tag{3.63}$$

$$\text{b)} \quad \text{if } \eta = 4\rho, \qquad \rho = \|h\|^{1/2},$$

$$\lim_{\|h\| \to 0} G_3(4\|h\|^{1/2}, \|h\|^{1/2}, \|h\|) = 0. \tag{3.64}$$

Remark. With the same technique used in Theorem 3.1, we obtain

$$V(x) \leqq \tilde{G}_3(\eta, \rho, \|h\|) + \bar{w}^h(x) \quad \forall x \in \Omega^h \tag{3.65}$$

with $\tilde{G}_3(\eta, \rho, \|h\|) = G_3(\eta, \rho, \|h\|) + L_V\|h\|$.

Furthermore, (3.65) implies

$$\lim_{\|h\| \to \infty} \bar{w}^h(x) \geqq V(x). \tag{3.66}$$

In view of (3.66) convergence will be assured if we can prove the inequality

$$\overline{\lim_{\|h\| \to \infty}} \bar{w}^h(x) \leqq V(x). \tag{3.67}$$

To obtain (3.67), let us consider a suitable sequence of stopping time problems whose solution is the sequence of functions $V_{\mu,\nu,i}$.

DEFINITION OF $V_{\mu,\nu,i}$. To simplify notation we will suppose that the sets of controls U and Z are finite sets

$$U = \{u_i; i = 1, \cdots, n_u\}, \qquad Z = \{z_k; k = 1, \cdots, n_z\}.$$

We introduce the set of control policies

$$U_{\mu,\nu,i} = \{(u(\cdot), z(\cdot)) | \text{a}), \text{b})\}. \tag{3.68}$$

a) $u(\cdot)$ is a left continuous piecewise constant function with a maximum of ν switching points, with values in U and $u(0) = u_i$.

b) $z(\cdot)$ has a maximum of μ impulses and $z(\theta_s) \in Z$.

Corresponding to (3.68) we introduce the functions

$$V_{\mu,\nu,i}(x) = \inf \{J(x, u(\cdot), z(\cdot), \theta) | (u(\cdot), z(\cdot)) \in U_{\mu,\nu,i},\ \theta \geqq 0\}, \tag{3.69}$$

having the following properties as it is easy to verify from the definition (3.69): $\forall x \in \Omega$

$$V_{\mu,\nu,i}(x) \leqq V_{\mu',\nu,i}(x) \quad \text{if } \mu > \mu', \tag{3.70}$$

$$V_{\mu,\nu,i}(x) \leqq V_{\mu,\nu',j}(x) \quad \text{if } \nu > \nu' \quad \forall j = 1, \cdots, n_u, \tag{3.71}$$

$$V(x) \leqq V_{\mu,\nu,i}(x) \quad \forall \mu \geqq 0, \quad \nu \geqq 0, \quad i = 1, \cdots, n_u. \tag{3.72}$$

Furthermore, V is the limit of $V_{\mu,\nu,i}$ in the sense specified in the following:

PROPOSITION 3.1. *There exists* $\delta_{\mu,\nu}(x)$ *such that*

$$V_{\mu,\nu,i}(x) - V(x) \leqq \delta_{\mu,\nu}(x) \quad \forall x \in \Omega \tag{3.73}$$

with $\lim_{(\nu,\mu)\to\infty} \delta_{\mu,\nu}(x) = 0$.

We can also characterize $V_{\mu,\nu,i}$ as solutions of stopping time problems. In fact we have as follows using dynamic programming techniques.

PROPOSITION 3.2. $V_{\mu,\nu,i}$ *is the optimal cost function of a stopping time problem defined by the recursion device*

$$V_{\mu,\nu,i}(x) = \min_{\theta \geqq 0} \left(\int_0^\theta e^{-\alpha s} l(y(s), u_i)\, ds + e^{-\alpha\theta} \psi_{\mu,\nu,i}(y(\theta)) \right) \tag{3.74}$$

with $dy/ds = f(y, u_i)$, $y(0) = x$, *and* $\psi_{\mu,\nu,i}$ *defined by*

$$\begin{aligned}
\psi_{\mu,\nu,i}(x) = \min \Bigg\{ & \min_{\substack{j=1,n_u \\ j\neq i}} V_{\mu,\nu-1,j}(x), \\
& \min_{k=1,n_z} (q(x, z_k) + V_{\mu-1,\nu,i}(x + g(x, z_k)), \phi(x)) \Bigg\} \quad \text{if } \nu \geqq 1, \quad \mu \geqq 1, \\
\psi_{0,\nu,i}(x) = \min & \left\{ \min_{\substack{j=1,n_u \\ j\neq i}} V_{0,\nu-1,j}(x), \phi(x) \right\} \quad \text{if } \mu = 0, \quad \nu \geqq 1, \\
\psi_{\mu,0,i}(x) = \min & \left\{ \min_{k=1,n_z} (q(x, z_k) + V_{\mu-1,0,i}(x + g(x, z_k))), \phi(x) \right\} \quad \text{if } \nu = 0, \quad \mu \geqq 1, \\
\psi_{0,0,i}(x) = \phi(x) & \quad \text{if } \nu = 0, \quad \mu = 0.
\end{aligned} \tag{3.75}$$

In the same way that we established Lipschitz continuity of $V(x)$, it is possible to ensure:

PROPOSITION 3.3. $V_{\mu,\nu,i}$ *are equi-Lipschitzian; more precisely*

$$|V_{\mu,\nu,i}(x) - V_{\mu,\nu,i}(x')| \leqq L_V \|x - x'\| \quad \forall x, x' \in \Omega. \tag{3.76}$$

As a consequence of (3.76) and of arguments similar to those leading to (1.9)-(1.14) we have

(3.77) $V_{\mu,\nu,i}$ is the maximum element of the set $W_{\mu,\nu,i}$ with $W_{\mu,\nu,i}=\{w:\Omega\to\mathbb{R}|\text{a) b) c)}\}$,

a) w Lipschitzian,

b) $w(x)\leqq\psi_{\mu,\nu,i}(x)$, $\forall x\in\Omega$,

c) $(\partial w(x)/\partial x)\, f(x,u_i)+l(x,u_i)-\alpha w(x)\geqq 0$, a.e. $x\in\Omega$.

After introducing a triangulation Ω^h in Ω we set, as in (3.75)

$$\psi^h_{\mu,\nu,i}(x^h_s)=\min\left\{\phi(x^h_s),\ \min_{\substack{j=1,n_u\\ j\neq i}}\bar w^h_{\mu,\nu-1,j}(x^h_s),\ \min_{k=1,n_z}\left(q(x^h_s,z_k)+\bar w^h_{\mu-1,\nu,i}(x^h_s+g(x^h_s,z_k))\right)\right\}$$

$$\text{if } \mu\geqq 1,\quad \nu\geqq 1,$$

$$\psi^h_{0,\nu,i}(x^h_s)=\min\left\{\phi(x^h_s),\ \min_{\substack{j=1,n_u\\ j\neq i}}(\bar w^h_{0,\nu-1,j}(x^h_s))\right\}\quad\text{if } \mu=0,\quad \nu\geqq 1, \tag{3.78}$$

$$\psi^h_{\mu,0,i}(x^h_s)=\min\left\{\phi(x^h_s),\ \min_{k=1,n_z}\left(q(x^h_s,z_k)+\bar w^h_{\mu-1,0,i}(x^h_s+g(x^h_s,z_k))\right)\right\}\quad\text{if } \mu\geqq 1,\quad \nu=0,$$

$$\psi^h_{0,0,i}(x^h_s)=\phi(x^h_s)\quad\text{if } \mu=0,\quad \nu=0;$$

where $\bar w^h_{\mu,\nu,i}$ is iteratively obtained from (3.78) using (3.79):

(3.79)
$\bar w^h_{\mu,\nu,i}$ is the maximum element of $W^h_{\mu,\nu,i}$ with

$W^h_{\mu,\nu,i}=\{w^h:\Omega^h\to R|\text{b}^h),\text{c}^h)\}$,

b^h) $w^h(x^h_s)\leqq\psi^h_{\mu,\nu,i}(x^h_s)$, $\forall x^h_s$ vertex of Ω^h,

c^h) $\dfrac{\partial w^h}{\partial x_f}(x^h_s,u_i)\|f(x^h_s,u_i)\|+l(x^k_s,u_i)-\alpha w^h(x^h_s)\geqq 0$.

Simultaneously we introduce the functions $\hat w^h_{\mu,\nu,i}$

$\hat w^h_{\mu,\nu,i}$ is the maximum element of $\hat W^h_{\mu,\nu,i}$ with

(3.80) $\hat W^h_{\mu,\nu,i}=\{w^h:\Omega^h\to R|\hat{\text{b}}^h),\text{c}^h)\}$,

$\hat{\text{b}}^h$) $w^h(x^h_s)\leqq\psi_{\mu,\nu,i}(x^h_s)$.

We remark that the sets (3.79) and (3.80) comprise finite linear elements.

Some properties of $\bar w^h_{\mu,\nu,i}$ *and* $\hat w^h_{\mu,\nu,i}$. After pointing out that using (3.75) we obtain

$$|\psi_{\mu,\nu,i}(x)-\psi_{\mu,\nu,i}(x')|\leqq(\max\{L_\phi, L_V, L_q+L_V(1+L_g)\})\cdot\|x-x'\|,$$

we can assume as the Lipschitz constant of $\psi_{\mu,\nu,i}$, independent of μ, ν, i

$$L_\psi=L_q+L_V(1+L_g). \tag{3.81}$$

So, using Lemmas 3.1 and 3.2 applied to the function $\hat w^h$ and the stopping time problem (3.74) we can ensure, with L_ψ given by (3.81),

$$\max_{x^h_s}|\hat w^h_{\mu,\nu,i}(x^h_s)-V_{\mu,\nu,i}(x^h_s)|\leqq G_4(\eta,\rho,\|h\|), \tag{3.82}$$

with $G_4(\eta,\rho,\|h\|)\to 0$, for $\eta=4\rho$, $\rho=\|h\|^{1/2}$, if $\|h\|\to 0$.

Now, as in Theorem 3.1,

$$\max_{x\in\Omega^h}|\hat w^h_{\mu,\nu,i}(x)-V_{\mu,\nu,i}(x)|\leqq G_4(\eta,\rho,\|h\|)+L_V\|h\|=G_5(\eta,\rho,\|h\|). \tag{3.83}$$

The following four propositions will be presented without proofs (for proofs see [12])

PROPOSITION 3.4. *The approximation solution* $\bar{w}^h$ *satisfies*

$$\bar{w}^h(x_s^h) \leqq \bar{w}^h_{\mu,\nu,i}(x_s^h) \quad \forall x_s^h, \quad \forall(\mu,\nu), \quad \forall i. \tag{3.84}$$

PROPOSITION 3.5.

$$\max_{x_s^h} (\bar{w}^h_{\mu,\nu,i}(x_s^h) - \hat{w}_{\mu,\nu,i}(x_s^h)) \leqq \max_{x_s^h} (\psi^h_{\mu,\nu,i}(x_s^h) - \psi_{\mu,\nu,i}(x_s^h))^+. \tag{3.85}$$

PROPOSITION 3.6. *For all* x_s^h *and for all* $i=1$, n_u *the positive part of* $\psi^h_{\mu,\nu,i} - \psi_{\mu,\nu,i}$ *has the following bounds*

(3.86)

$$\max_{x_s^h} (\psi^h_{\mu,\nu,i}(x_s^h) - \psi_{\mu,\nu,i}(x_s^h))^+$$

$$\leqq \max\left\{ \max_{x_s^h; j=1,n_u} (\bar{w}^h_{\mu,\nu-1,j}(x_s^h) - V_{\mu,\nu-1,j}(x_s^h))^+, \max_{x\in\Omega^h} (\bar{w}^h_{\mu-1,\nu,}i(x) - V_{\mu-1,\nu,i}(x))^+ \right\}$$

$$\text{if } \mu \geqq 1, \quad \nu \geqq 1,$$

$$\max_{x_s^h} (\psi^h_{0,\nu,i}(x_s^h) - \psi_{0,\nu,i}(x_s^h))^+$$

$$\leqq \max_{x_s^h, j=1,n_u} (\bar{w}^h_{0,\nu-1,j}(x_s^h) - V_{0,\nu-1,j}(x_s^h))^+ \quad \text{if } \mu = 0, \quad \nu \geqq 1,$$

$$\max_{x_s^h} (\psi^h_{\mu,0,i}(x_s^h) - \psi_{\mu,0,i}(x_s^h))^+$$

$$\leqq \max_{x\in\Omega^h} (\bar{w}^h_{\mu-1,\nu,i}(x) - V_{\mu-1,\nu,i}(x))^+ \quad \text{if } \mu \geqq 1, \quad \nu = 0,$$

$$\max_{x_s^h} (\psi^h_{0,0,i}(x_s^h) - \psi_{0,0,i}(x_s^h))^+ = 0 \quad \text{if } \mu = \nu = 0.$$

PROPOSITION 3.7.

$$\max_{\substack{x\in\Omega^h \\ i=1,n_u}} (\bar{w}^h_{\mu,\nu,i}(x) - V_{\mu,\nu,i}(x)) \leqq (1+\mu+\nu) G_5(\eta, \rho, \|h\|). \tag{3.87}$$

We are now able to conclude.

THEOREM 3.2. *The solution* $\bar{w}^h$ *of* (P_h) *converges uniformly to* $V(x)$.

To begin we remark that we can, as in the derivation of (3.20) use (3.64) and (3.65) to establish that for $\eta = 4\rho$, $\rho = \|h\|^{1/2}$, there exists $C_5 > 0$ such that

$$V(x) - C_5\|h\|^{1/2} \leqq \bar{w}^h(x) \quad \forall x \in \Omega, \quad \forall \|h\| \leqq \|h_0\|, \tag{3.88}$$

in which h_0 denotes a fixed "initial" triangulation.

Taking into account the affinity of functions $\bar{w}^h$, $\bar{w}^h_{\mu,\nu,i}$, we obtain from (3.84)

$$\bar{w}^h(x) \leqq \bar{w}^h_{\mu,\nu,i}(x) \quad \forall x \in \Omega^h,$$

which allows us, taking advantage of (3.87), to write

$$\bar{w}^h(x) \leqq V_{\mu,\nu,i}(x) + (1+\nu+\mu) G_5(\eta, \rho, \|h\|)$$

and, to ensure the existence of a positive constant C_6 such that

$$\bar{w}^h(x) \leqq V_{\mu,\nu,i}(x) + (1+\nu+\mu) C_6 \|h\|^{1/2} \quad \forall x \in \Omega, \quad \forall \|h\| < \|h_0\|. \tag{3.89}$$

Using (3.73) in (3.89), we have

$$\bar{w}^h(x) \leqq V(x) + \delta(\mu,\nu) + (1+\nu+\mu) C_6 \|h\|^{1/2} \tag{3.90}$$

with

$$\delta(\mu, \nu) \to 0 \quad \text{if } (\mu, \nu) \to \infty. \tag{3.91}$$

To finish, by (3.91), for all $\varepsilon > 0$ there exists $(\mu_\varepsilon, \nu_\varepsilon)$ such that $\delta(\mu_\varepsilon \nu_\varepsilon) < \varepsilon/2$; if we choose

$$\|h_\varepsilon\| = \min\left\{\left(\frac{\varepsilon}{C_6}\right)^2, \left(\frac{(\varepsilon/2)C_6}{1+\mu_\varepsilon+\nu_\varepsilon}\right)^2\right\},$$

we obtain, using (3.88), (3.90),

$$|V(x) - \bar{w}^h(x)| \leqq \varepsilon \quad \forall x \in \Omega^h, \quad \forall \|h\| \leqq \|h_\varepsilon\|,$$

that is, the desired conclusion.

3.4. An estimation of the rate of convergence. We reduce our control policy to stopping time and impulse controls. Because we do not have continuous control, we ignore the parameter ν in (3.73) and, following the same technique used in [9] we can show that

$$\delta(\mu) \leqq 2\, e\left(M_\phi + \frac{M_I}{\alpha}\right) e^{-\mu p}, \qquad p = \frac{q_0}{2\, e(M_I/\alpha + M_\phi)}.$$

So (3.90) becomes

$$\bar{w}^h(x) \leqq V(x) + 2\, e\left(M_\phi + \frac{M_I}{\alpha}\right)^{-\mu p} + (1+\mu) C_6 \|h\|^{1/2} \tag{3.92}$$

$$\forall \mu = 0, 1, \cdots, \forall \|h\| \leqq \|h_0\|.$$

Using for μ in (3.92) the integer part of

$$\frac{-1}{p} \log \frac{C_5 \|h\|^{1/2}}{2\, e(M_\phi + M_I/\alpha)},$$

we can ensure the existence of a constant C_7 such that

$$|\bar{w}^h(x) - V(x)| \leqq C_7 |\log \|h\|| \cdot \|h\|^{1/2} \quad \forall x \in \Omega_h, \quad \forall \|h\| \leqq \|h_0\|. \tag{3.93}$$

Remarks. In [13] we can see an estimation of the rate of convergence of value functions associated with a discrete-time approximation. The result applies using bang-bang controls.

We want to also note that in some applications we have solved exactly the fixed point problem using different types of "one-iteration convergent" algorithms (see [14], [15]).

REFERENCES

[1] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.

[2] ———, *Contrôle impulsionnel et inéquations quasi-variationnelles*, Dunod, Paris, 1982.

[3] W. FLEMING AND R. RISHEL, *Optimal Deterministic and Stochastic Control*, Springer-Verlag, New York, 1975.

[4] M. N. EL TARAZI, *Contraction et ordre partiel pour l'étude d'algorithmes synchrones et asynchrones en analyse numérique*, Thèse d'etat, Besançon, 1981.

[5] A. FRIEDMAN, *Differential Games*, Wiley Interscience, New York, 1971.

[6] R. GLOWINSKI, J. L. LIONS AND R. TREMOLIERES, *Analyse numérique des inéquations variationnelles*, Dunod, Paris, 1976.

[7] R. GONZALEZ, *Sur l'existence d'une solution maximale de l'équation de Hamilton-Jacobi*, CRAS Paris, 1976, Ser. A, pp. 1287-1290.

[8] R. GONZALEZ AND E. ROFMAN, *An algorithm to obtain the maximum solution of the Hamilton-Jacobi equation*, Lecture Notes in Control and Information Sciences, 6, Springer-Verlag, New York, 1978.

[9] R. GONZALEZ, *Sur la résolution de l'équation de Hamilton-Jacobi du contrôle déterministe*, Thèse, 3ème cycle, Université de Paris IX, 1980; Cahiers de Math. de la Décision, Ceremade, 8029 bis.

[10] R. GONZALEZ AND E. ROFMAN, *Iterative computation of the maximal solution of the Hamilton-Jacobi equation*, V. Symposium of Op. Res., Verlag Anton Hain, Part I, R. Burkard and T. Ellinger, eds., Univ. Koln, Aug. 25-27, 1980.

[11] R. GONZALEZ, *Optimal bang-bang control policies with restrictions of time between switchings*, Matematicae Notae, XXVII (1979), pp. 111-137.

[12] R. GONZALEZ AND E. ROFMAN, *Rapport de Recherche*, 151 INRIA.

[13] I. C. DOLCETTA AND M. FALCONE, *Discrete time approximation to the infinite horizon problem*, Raporto di Ricerca; Istituto Matematico, Univ. di Roma, to appear.

[14] J. C. MIELLOU, *Sur la résolution itérative d'un problème de point fixe: un algorithme monoitération*, Rapport ERA, Univ. de Besançon; CNRS, to appear.

[15] R. GONZALEZ AND E. ROFMAN, *A real-time algorithm for optimal energy production management*, Proc.International IASTED Symposium EES '83, Athens, Acta Press, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

006

ON DETERMINISTIC CONTROL PROBLEMS: AN APPROXIMATION PROCEDURE FOR THE OPTIMAL COST II THE NONSTATIONARY CASE*

R. GONZALEZ† AND E. ROFMAN‡

Abstract. We study deterministic optimal control problems having stopping time, continuous and impulse controls in each strategy.

We obtain the optimal cost, considered as the maximum element of a suitable set of subsolutions of the associated Hamilton-Jacobi equation, using an approximation method. A particular derivative discretization scheme is employed.

Convergence of approximate solutions is shown taking advantage of a discrete maximum principle which is also proved.

For the numerical solutions of approximate problems we use a method of relaxation type. The algorithm is very simple; it can be run on computers of small central memory.

In Part I [SIAM J. Control Optim., 23 (1985), pp. 242-266] we studied the stationary case; in Part II we study the nonstationary case and we apply our results to a short-run model of energy production management.

Key words. deterministic control, Hamilton-Jacobi-Bellman equations, finite elements, energy production systems

Introduction. In this part we consider the nonstationary case. There are not serious difficulties in extending the results of Part I (this issue, pp. 242-266) to the case in which the dynamics depends explicitly on the time t. For the most part we will limit ourselves merely to stating the results concerning the nonstationary case; we will comment on and analyse only those aspects for which there are some important differences.

As an application of the methodology described in Part II we give a solution to the problem of computing the optimal control of an electrical production system. Systems with a significant number of thermal and hydropower plants may be optimized in this way.

1. The theoretical approach.

The original problem and its equivalent formulation. In this case the system satisfies in absence of impulse controls the differential equation

$$\begin{aligned} &\frac{dy}{ds} = f(y, u, s), && x \in \Omega \subset \mathbb{R}^n, \\ &y(t) = x, && t \in [0, T]. \end{aligned} \tag{1.1}$$

$u(\cdot)$ is a measurable function of the time, with values in a compact set $U \subset \mathbb{R}^m$.

In a finite set of times θ_ν $(\nu = 1, 2, \cdots, \mu)$ impulses $z(\theta_\nu) \in Z$ are applied; the trajectory jumps are

$$y(\theta_\nu^+) = y(\theta_\nu^-) + g(y(\theta_\nu^-), z(\theta_\nu), \theta_\nu). \tag{1.2}$$

Z is a compact set in $\mathbb{R}^p$.

* Received by the editors October 7, 1982, and in final revised form April 15, 1984. This work was supported in part by the U.S. Department of Energy, Office of Electric Energy Systems, under contract 01-80RA-50154.

† Electronics Department, University of Rosario, Argentina. The work of this author was supported in part by CONICET, Argentina under grant 9977/81.

‡ INRIA, Domaine de Voluceau, BP 105-Rocquencourt, 78153 Le Chesnay Cedex, France.

We denote by $(u(\cdot), z(\cdot), \tau)$ a control strategy with the stopping time $\tau \in [0, T[$. The cost associated with each strategy is

$$J(x, t; u(\cdot), z(\cdot), \tau) = \int_t^\tau e^{-\alpha(s-t)} l(y(s), u(s), s)\, ds + \sum_\nu q(y(\theta_\nu^-), z(\theta_\nu), \theta_\nu)\, e^{-\alpha(\theta_\nu - t)} + e^{-\alpha(\tau - t)} \phi(y(\tau), \tau) \cdot \chi_{[t,T[}, \tag{1.3}$$

with $\chi_{[t,T[}(\cdot)$ a characteristic function of the interval $[t, T[$.

The optimal cost function is $\forall (x, t) \in Q$

$$V(x, t) = \inf J(x, t; u(\cdot), z(\cdot), \tau), \tag{1.4}$$

$$Q = \Omega \times [0, T]. \tag{1.5}$$

In the following we will suppose:

i) f, l, ϕ, g, q are continuous and bounded functions; they are Lipschitzian functions in (x, t).

ii) $\phi(x, T) \geqq 0 \quad \forall x \in \Omega$.

iii) $q(x, z, t) \geqq q_0 > 0 \quad \forall (x, t) \in Q, \forall z \in Z$.

iv) $\forall t$, $y(t) \in \Omega$, independent of the strategy.

We can give the following characterization of $V(x, t)$.

THEOREM 1.1. *$V(x, t)$ is the maximum element of the set W, with*

$$W = \{w(x, t) \to \mathbb{R} \,|\, (1.6)\text{-}(1.10)\}; \tag{1.6}$$

$w(x, t)$ Lipschitzian function in (x, t);

$$\frac{\partial w(x, t)}{\partial t} + \min_{u \in U} \left| \left[\frac{\partial w(x, t)}{\partial x} \cdot f(x, u, t) + l(x, u, t) - \alpha w(x, t) \right] \geqq 0 \right. \tag{1.7}$$

a.e. $(x, t) \in Q$;

$$w(x, t) \leqq \min_{z \in Z} (q(x, z, t) + w(x + g(x, z, t), t)) \quad \forall (x, t) \in Q; \tag{1.8}$$

$$w(x, t) \leqq \phi(x, t) \quad \forall (x, t) \in Q; \tag{1.9}$$

$$w(x, T) \leqq 0 \quad \forall x \in \Omega. \tag{1.10}$$

The proof follows the method used in [9, p. 29].

2. The discretized problem (P_h).

2.1. Introduction.

a) The set Q is approximated by a triangulation Q^h, a union of simplices of vertices (x_p, t_q); $p = 0, N_x$; $q = 0, N_T$, $t_q = q \cdot \delta$, $\delta = T/N_T$. This triangulation is "regular in t" in the following sense:

i) each simplex of Q^h has its vertices in two hyperplanes with equations $t = t_q$, $t = t_{q+1}$.

ii) If a face of a simplex of Q^h is contained in the hyperplane $\{t = t_q\}$ we will have a "mirror image" of that face in the hyperplanes $\{t = t_{q-1}\}$, $\{t = t_{q+1}\}$; they are themselves faces of simplices of Q^h. An example is shown in Fig. 1.

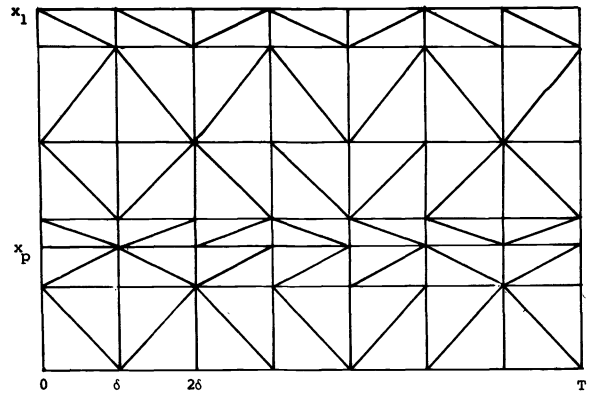


FIG. 1

b) In the set of linear finite elements w^h defined in Q^h we consider the set W^h:

$$W^h = \{w^h : Q^h \to R \,|\, (2.1), (2.3), (2.4), (2.5)\},$$

$$\frac{\partial w^h}{\partial t}(x_p, t_q; u) + \frac{\partial w^h}{\partial x_f}(x_p, t_p, u)\|f(x_p, u, t_q)\| + l(x_p, u, t_q) \tag{2.1}$$

$$-\alpha w^h(x_p, t_q) \geqq 0 \quad \forall u \in U^h, \quad \forall x_p, p = 0, N_x \quad \forall t_q, q = 0, N_T - 1.$$

For example, in the situation depicted in Fig. 2, the expression

$$\frac{\partial w^h}{\partial t}(x_p, t_q; u) + \frac{\partial w^h}{\partial x_f}(x_p, t_q; u)\|f(x_p, u, t_q)\|$$

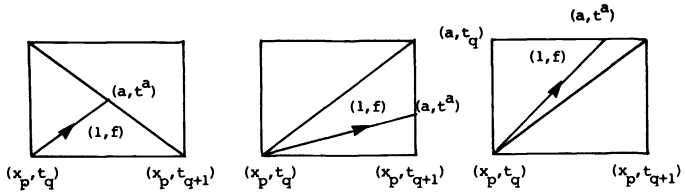


FIG. 2

is approximated by

$$\frac{w^h(a, t^a) - w^h(x_p, t_q)}{\Delta} \quad \text{with } \Delta = t^a - tq, \tag{2.2}$$

$$w^h(x_p, t_q) \leqq q(x_p, z, t_q) + w^h(x_p + g(x_p, z, t_q), t_q) \tag{2.3}$$

$$\forall z \in Z^h, \quad \forall x_p, p = 0, N_x, \quad \forall t_q, q = 0, N_T - 1,$$

$$w^h(x_p, t_q) \leqq \phi(x_p, t_q) \quad \forall p = 0, N_x \quad \forall q = 0, N_T - 1, \tag{2.4}$$

$$w^h(x_p, t_{N_T}) \leqq 0 \quad \forall p = 0, N_x. \tag{2.5}$$

Remarks. Similar observations apply as in d) of Part I, § 2.2.

c) We introduce the following partial order "$\leqq$":

$$w^h \leqq \tilde{w}^h \Leftrightarrow w^h(x_p, t_q) \leqq \tilde{w}^h(x_p, t_q) \quad \forall p = 0, N_x, \quad q = 0, N_T \tag{2.6}$$

and we pose the discretized problem:

(P_h): Find the maximum element $\bar{w}^h$ of the set W^h with respect to the partial order "$\leqq$".

2.2. The solution of (P_h) and its properties. Equations (2.1) and (2.3) will be transformed into equivalent and more useful relations.

As Fig. 3 shows we express the point (a, t^a) as a convex combination of points (x', t_q) and (a', t_{q+1}),

$$(a, t^a) = \frac{\Delta}{\delta}(a', t_{q+1}) + \left(1 - \frac{\Delta}{\delta}\right)(x', t_q). \tag{2.7}$$

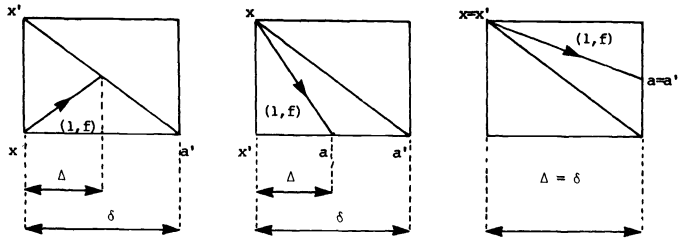


FIG. 3

Furthermore, taking into account that x' and a' are, in general, interior points of faces (or edges) of some simplex, we will express these points as convex combinations of the vertices of the faces to which they belong:

$$a'(u) = \sum_j \lambda_j(x_p, t_q, u)x_j, \tag{2.8}$$

$$x'(u) = \sum_j \hat{\lambda}_j(x_p, t_q, u)x_j. \tag{2.9}$$

So, because of (2.7) and the affinity of w^h, (2.1) becomes

$$w^h(x_p, t_q) \leqq \min_{u \in U^h} \frac{1}{(1+\alpha\Delta)} \left\{ \frac{\Delta}{\delta} \sum_j \lambda_j(x_p, t_q, u) \cdot w^h(x_j, t_{q+1}) + \left(1 - \frac{\Delta}{\delta}\right) \sum_j \hat{\lambda}_j(x_p, t_q, u) w^h(x_j, t_q) + \Delta l(x_p, u, t_q) \right\}. \tag{2.10}$$

In the same way we put

$$x_p + g(x_p, z, t_q) = \sum_j \lambda'_j(x_p, t_q, z)x_j \tag{2.11}$$

and (2.3) is rewritten in the equivalent form

$$w^h(x_p, t_q) \leqq \min_{z \in Z^h} \left(q(x_p, z, t_q) + \sum_j \lambda'_j(x_p, t_q, z) w^h(x_j, t_q)\right) \quad \forall p = 0, N_x, \quad q = 0, N_T - 1. \tag{2.12}$$

We will use (2.10) and (2.12) to define the real operator M (w^h denotes a linear finite element in Q^h):

if $q = N_T$ $(Mw^h)(x_p, t_q) = 0$,

if $q = 0, \cdots, N_T - 1$

$$(Mw^h)(x_p, t_q) = \min \left\{ \phi(x_p, t_q), \min_{z \in Z^h} \left(q(x_p, z, t_q) + \sum_j \lambda_j'(x_p, t_q, z) w^h(x_j, t_q) \right), \right.$$

$$\min_{u \in U^h} \frac{1}{1+\alpha\Delta} \left[\frac{\Delta}{\delta} \sum \lambda_j(x_p, t_q, z) w^h(x_j, t_{q+1}) \right. \tag{2.13}$$

$$\left. \left. + \left(1 - \frac{\Delta}{\delta}\right) \sum_j \hat{\lambda}_j(x_p, t_q, u) w^h(x_j, t_q) + l(x_p, u, t_q) \right] \right\}.$$

We define Mw^h at arbitrary points in Q^h by linear interpolations of the values given by (2.13) at the vertices of the triangulation.

Some properties of Mw^h which follow immediately are;

$$w^h \geqq \hat{w}^h \to Mw^h \geqq M\hat{w}^h, \tag{2.14}$$

$$w^h \in W^h \Leftrightarrow w^h \leqq Mw^h. \tag{2.15}$$

Remark. (2.15) gives us a characterization of W^h.

Finally the most important property is given by:

THEOREM 2.1. *There exists $\bar{w}^h$, maximum element of W^h; furthermore $\bar{w}^h$ is characterized by the condition $\bar{w}^h = M\bar{w}^h$, i.e.*

$$\bar{w}^h = M\bar{w}^h \leftrightarrow \bar{w}^h \geqq w^h \quad \forall w^h \in W^h. \tag{2.16}$$

Proof. We follow the proof of Theorem I.2.2, using the following new discrete maximum principle.

LEMMA 2.1. *Let us call S^h a subset of all vertices of Q^h and C^h its complement. Let w^h be some linear finite element defined in Q^h such that $w^h(x_p, t_{N_T}) \leqq 0$, and $r: \Omega \times U \times [0, T] \to \mathbb{R}$.*

If there exists $U_{p,q} \subset U^h$ such that for all $(x_p, t_q) \in C^h$,

$$\min_{u \in U_{p,q}} \left[\frac{\partial w^h(x_p, t_q; u)}{\partial t} + \frac{\partial w^h(x_p, t_q; u)}{\partial x} \| f(x_p, u, t_q) \| \right. \tag{2.17}$$

$$\left. + r(x_p, u, t_q) - \alpha w^h(x_p, t_q) \right] \geqq 0$$

then

$$w^h(x_p, t_q) \leqq M_{S^h}^+ + M_r^+ \cdot T \quad \forall (x_p, t_q) \in C^h, \tag{2.18}$$

with

$$M_{S^h}^+ = \max_{(x_p, t_q) \in S^h} (w^h(x_p, t_q) \vee 0), \tag{2.19}$$

$$M_r^+ = \max_{\substack{(x_p, t_q) \in C^h \\ u \in U_{p,q}}} (r(x_p, u, t_q) \vee 0). \tag{2.20}$$

We show the lemma. Let us define

$$M_{C_q} = \max_{\substack{(x_p, t_{q'}) \\ q \leqq q' \leqq N_T}} w^h(x_p, t_q); \tag{2.21}$$

(2.18) will be proved after showing that

$$M_{C_q} \leqq M_{S^h}^+ + M_r^+(T - t_q) \tag{2.22}$$

holds for any $q = 0, 1, \cdots, N_T$. We will proceed inductively.

We know after (2.5) that (for $q = N_T$) $M_{C_{N_T}} \leqq 0$; then

$$M_{S^h}^+ + M_r^+(T - t_{N_T}) = M_{S^h}^+ \geqq 0 \geqq M_{C_{N_T}}. \tag{2.23}$$

Now we will suppose that (2.22) holds for q and we will show that such a supposition implies that it holds for $q-1$. Let $(x_p, t_{q-1}) \in C^h$. From (2.17) used in its equivalent form (2.10), we have

$$w^h(x_p, t_{q-1}) \leqq \min_{u \in U_{p,q-1}} \frac{1}{1+\alpha\Delta} \left\{ \frac{\Delta}{\delta} \sum_j \lambda_j(x_p, t_{q-1}, u) w^h(x_j, t_q) + \left(1 - \frac{\Delta}{\delta}\right) \sum_j \hat{\lambda}_j(x_p, t_{q-1}, u) w^h(x_j, t_{q-1}) + r(x_p, u, t_{q-1})\Delta \right\}. \tag{2.24}$$

If $\forall (x_p, t_{q-1}) \in C^h$,

$$w^h(x_p, t_{q-1}) \leqq M_{S^h}^+ + M_r^+(T - q\delta), \tag{2.25}$$

it follows that, with $q-1$ in place of q in the second term, it is also true that $w^h(x_p, t_{q-1}) \leqq M_{S^h}^+ + M_r^+(T-(q-1)\delta)$, that is to say (2.22) holds for $q-1$. If (2.25) does not hold, there exists (x_p^*, t_{q-1}^*) such that

$$w^h(x_p^*, t_{q-1}^*) = M_{C_{q-1}} > M_{S^h}^+ + M_r^+(T - q\delta) \tag{2.26}$$

or, in other words

$$w^h(x_p, t_{q-1}) \leqq M_{C_{q-1}} \quad \forall p = 0, N_x. \tag{2.27}$$

On the other hand, as we have accepted that (2.22) holds for q:

$$w^h(x_p, t_q) \leqq M_{S^h}^+ + M_r^+(T - q\delta); \tag{2.28}$$

then, using (2.27), (2.28) in (2.24), we have

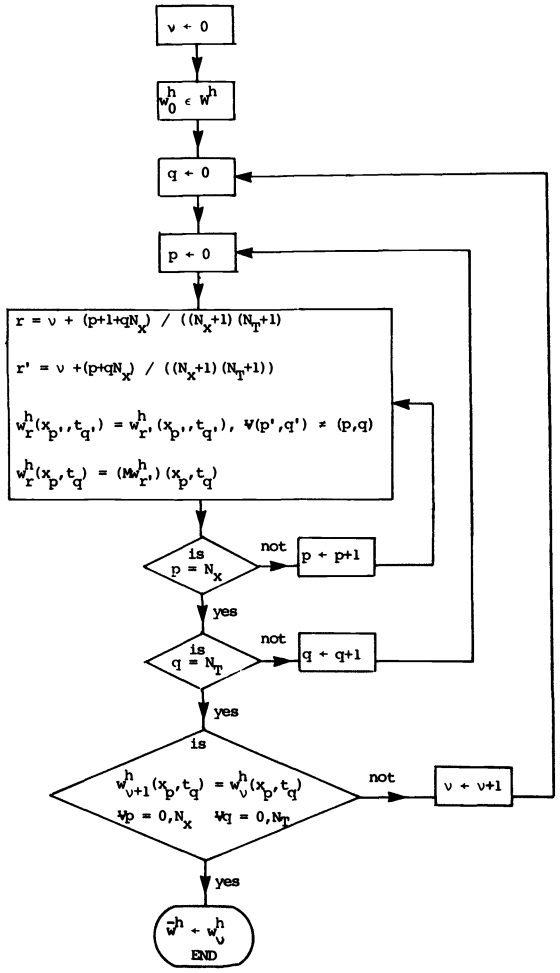
$$M_{C_{q-1}} \leqq \min_{u \in U_{p,q-1}} \frac{1}{1+\alpha\Delta} \left\{ \frac{\Delta}{\delta}(M_{S^h}^+ + M_r^+(T - q\delta)) + \left(1 - \frac{\Delta}{\delta}\right) M_{C_{q-1}} + M_r^+\Delta \right\}.$$

From here, as $1/(1+\alpha\Delta) < 1$, $\Delta \leqq \delta$ we have

$$M_{C_{q-1}} \leqq M_{S^h}^+ + M_r^+(T - (q-1)\delta). \qquad \square$$

2.3. Algorithms to compute w^h. To compute $\bar{w}^h$ we can use Algorithms 2.1 and 2.2. These algorithms, similar to Algorithm 2.1 in Part I used in the stationary case, define increasing elements $w_p^h \in W^h$, having $\bar{w}^h$ as limit.

ALGORITHM 2.1.



In fact, it is possible to show the following theorems.

THEOREM 2.2. *Algorithm* 2.1 *stops after a finite number of steps at the element* $w_{\bar{\nu}}^h = \bar{w}^h$ *or it gives a sequence* $\{w_\nu^h\}$ *convergent to* $\bar{w}^h$, *i.e.*

$$\lim_{\nu \to \infty} w_\nu^h(x_p, t_q) = \bar{w}^h(x_p, t_q) \quad \forall p = 0, N_x, \quad \forall q = 0, N_T.$$

THEOREM 2.3. *Algorithm* 2.2 *stops after a finite number of iterations at the element* $\bar{w}_\varepsilon^h$, *which has the following properties*;

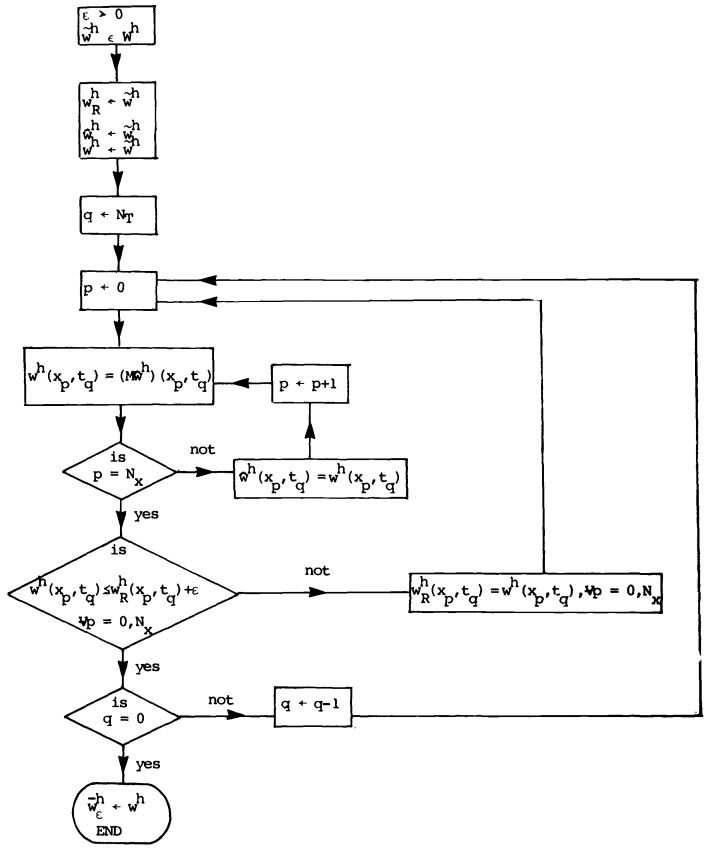
a) $\bar{w}_\varepsilon^h \in W^h$, $\forall \varepsilon > 0$,

b) $\varepsilon \leqq \varepsilon' \to \bar{w}_\varepsilon^h \geqq \bar{w}_{\varepsilon'}^h$,

c) $\lim_{\varepsilon \to 0} \bar{w}_\varepsilon^h = \bar{w}^h$.

Remark. Algorithm 2.2 is an improvement on Algorithm 2.1 which takes advantage of the particular structure of nonstationary problems (we use backward solutions).

ALGORITHM 2.2.



2.4. The convergence of the approximate solutions. It is possible to prove a theorem similar to Theorem I.3.2.

THEOREM 2.4. *The approximate solutions* $\bar{w}^h$ *converge uniformly to* $V(x, t)$, *i.e.*

$$\lim_{\|h\|\to 0} \max_{(x,t)\in Q^h} |\bar{w}^h(x, t) - V(x, t)| = 0.$$

3. An application to the management of energy production.

3.1. Model of the problem (short-run model). The energy production system consists of two thermal power plants (P_1, P_2 being their level of production) and a dam (x_h: hydropower stock, P_h: hydropower production). D is the demand of electricity and we denote by P_3 the production of an additional source which is available if it is required:

$$D = P_1 + P_2 + P_h + P_3. \tag{3.1}$$

The cost of the operation is given by

$$J = \int_0^T (c_1P_1(t) + c_2P_2(t) + c_h(x_h(t))P_h(t) + c_3P_3(t))\, dt + n_1\bar{k}_1 + n_2\bar{k}_2. \tag{3.2}$$

n_1, n_2 is the number of starts of plants 1, 2 in the interval $[0, T]$; k_1, k_2 the costs of each start. We suppose c_1, c_2, c_3 constants and $c_h(x_h)$ is a shadow price obtained after

a long-run optimization (about one year). In our problem we will consider $[0, T]$ one day or one week (cf. [12], [14], [15]).

We will suppose that there are no delays between the start of a thermal plant and the instant in which it begins to produce energy. The methodology to be used here can be easily modified to take into account these delays (cf. [11], [12], [16]).

In this form the system will be modeled by its internal state: a discrete variable E (showing if the plants 1 and 2 are working or not) and a continuous variable x_h whose evolution equation is

$$\frac{dx_h}{dt} = A(t) - P_h(t), \qquad 0 \leqq x_h \leqq x_{h,\max}. \tag{3.3}$$

$A(t)$ is the input of water in the dam.

$$E = \begin{cases} 1, & \text{plants 1 and 2 do not operate,} \\ 2, & \text{plant 1 operates; plant 2 does not operate,} \\ 3, & \text{plant 1 does not operate; plant 2 operates,} \\ 4, & \text{plants 1 and 2 operate.} \end{cases}$$

Our aim is to obtain the control strategy giving the minimum of J. The optimal policy is a decision concerning when plants 1 and 2 must operate and at what level of production. We look for optimal feedback policies acting on the instantaneous state $(E(t), x_h(t))$ of the system.

3.2. Optimal feedback policies. Let us consider as parameters the initial state x and the initial time t of the system and let us introduce the optimal cost functions

$$V_i(x, t),\ i = 1, 4,\ (x, t) \in Q = [0, x_{h,\max}] \times [0, T],$$

$$V_i(x, t) = \inf_{P_1(\cdot), P_2(\cdot), P_h(\cdot)} J(x_h, i, t, P_1(\cdot), P_2(\cdot), P_h(\cdot)) \tag{3.4}$$

with

$$\begin{aligned} &J(x_h, i, t, P_1(\cdot), P_2(\cdot), P_h(\cdot)) \\ &\quad = \int_t^T (c_1 P_1(s) + c_2 P_2(s) + c_h(x_h(s)) P_h(s) + c_3 P_3(s))\, ds + n_1 \bar{k}_1 + n_2 \bar{k}_2 \end{aligned} \tag{3.5}$$

cost related to the policy $P_1(\cdot)$, $P_2(\cdot)$, $P_h(\cdot)$ in the interval $[t, T]$ with the initial data $(E(t), x_h(t)) = (i, x_h)$.

From $V_i(x, t)$ it is possible to define the optimal feedback policies (cf. [2], [3], [9]). So, our problem is to compute $V_i(x, t)$. We recall for that the following.

3.3. Quasi-variational inequalities (QVI) associated with the control problem and characterization of V_i. It is possible to show (cf. [5], [9]) that V_i's are differentiable in a.e.$(x, t) \in Q$. Furthermore they verify (cf. [2], [9], [12]) the system of QVI:

$$\begin{aligned} &\frac{\partial V_i}{\partial t} + \min_{(P_1, P_2, P_h) \in \Gamma_i(x_h)} \left(\frac{\partial V_i}{\partial x_h} \cdot (A - P_h) + c_1 P_1 + c_2 P_2 \right. \\ &\qquad \left. + c_3 (D - P_1 - P_2 - P_h)^+ + c_h(x_h) P_h \right) \geqq 0, \end{aligned} \tag{3.6}$$

$$V_i(x_h, t) \leqq V_j(x_h, t) + k_j^i \quad \forall j \neq i, \tag{3.7}$$

$$V_i(x_h, T) = 0, \tag{3.8}$$

with $\Gamma_i(x_h)$ the set of admissible levels of production related to the state i and the stock x_h; k_j^i the cost for passing from state i to state j.

Remark. In a.e. $(x, t) \in Q$, one, at least, of (3.6), (3.7) becomes an equality.

The following characterization of $V_i(x, t)$ will allow us to compute it using the method introduced in § 2.

$V_i(x, t)$ is the maximum element of the set,

$$W_i = \{w_i \in H^{1,\infty}(Q) / w_i \text{ verifies (3.6), (3.7), (3.8)}\}, \qquad i = 1, 2, 3, 4, \tag{3.9}$$

i.e.

$$w_i(x, t) \leqq V_i(x, t) \quad \forall (x, t) \in Q, \quad \forall w_i \in W_i, \quad i = 1, 2, 3, 4.$$

3.4. Discretization of (3.9) and the discrete problem. Let us introduce in Q a triangulation Q^a as shown in Fig. 4, and let us consider in it linear finite elements

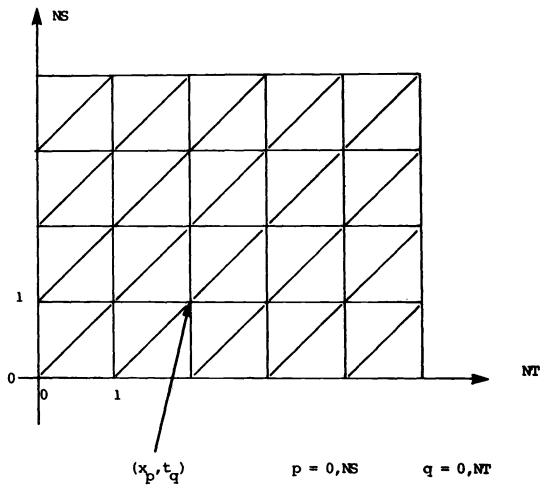


FIG. 4

with vertices (x_p, t_q). The set W is replaced by the approximate set W^a having as elements, linear finite elements $w^a = (w_i^a)$ satisfying suitable discretizations of (3.6), (3.7), (3.8):

If $A - P_h \geqq 0$,

$$\frac{w_i^a(x_p, t_{q+1}) - w_i^a(x_p, t_q)}{t_{q+1} - t_q} + \frac{w_i^a(x_{p+1}, t_{q+1}) - w_i^a(x_p, t_{q+1})}{x_{p+1} - x_p}(A - P_h)$$

$$+ c_1 P_1 + c_2 P_2 + c_h(x_p) P_h + c_3 (D - P_1 - P_2 - P_h)^+ \geqq 0, \tag{3.10}$$

$$\forall (P_1, P_2, P_h) \in \Gamma_i^a(x_p), \quad \forall p = 0, \cdots, NS - 1, \quad \forall q = 0, \cdots, NT - 1.$$

If $A - P_h < 0$,

$$\frac{w_i^a(x_p, t_{q+1}) - w_i^a(x_p, t_q)}{t_{q+1} - t_q} + \frac{w_i^a(x_{p-1}, t_q) - w_i^a(x_p, t_q)}{x_{p-1} - x_p}(A - P_h)$$

$$+ c_1 P_1 + c_2 P_2 + c_h(x_p) P_h + c_3 (D - P_1 - P_2 - P_h)^+ \geqq 0 \tag{3.10'}$$

$$\forall (P_1, P_2, P_h) \in \Gamma_i^a(x_p), \quad \forall p = 1, \cdots, NS, \quad \forall q = 0, \cdots, NT - 1.$$

$$(3.11)\quad w_i^a(x_p, t_q) \leqq w_j^a(x_p, t_q) + k_j^i \quad \forall p = 0, NS, \quad \forall q = 0, NT-1, \quad \forall i, \quad \forall j \neq i,$$

$$(3.12)\quad w_i^a(x_p, t_{NT}) = 0, \quad p = 0, NS, \quad i = 1, 4.$$

We pose the following discrete problem:

(P^a): Find the maximum element $\bar{w}^a$ of W^a.

By Theorem I.2.2 and Theorem 2.1., we know that $\bar{w}^a$ exists and it is unique. We can also introduce an algorithm with the properties remarked in § 3. In fact, after a suitable transformation of (3.10), (3.10′) into relations of the following type

$$(3.13)\quad w_i^a(x_p, t_q) \leqq \phi_i^a(w_i^a(x_p, t_{q+1}), w_i^a(x_{p-1}, t_q), w_i^a(x_{p+1}, t_{q+1}), x_p, t_q)$$

we define:

ALGORITHM 3.1.
Step 0; $\tilde{w}_i^a(x_p, t_q) = w_i^a(x_p, t_q) = 0, \forall i = 1, 4, \forall p = 0, NS, \forall q = 0, NT.$
Step 1: $q = NT - 1$
Step 2: $p = 0$
Step 3: $i = 1$
Step 4: $w_i^a(x_p, t_q) = \min\{w_j^a(x_p, t_q) + k_j^i (j \neq i); \phi_i^a(w_i^a, \cdots, x_p, t_q)\}$
Step 5: if $i = 4$ go to 6; if not, $i = i + 1$ and go to 4
Step 6: if $\tilde{w}_i^a(x_p, t_q) = w_i^a(x_p, t_q), \forall i = 1, 4$, go to 7; if not do $\tilde{w}_i^a(x_p, t_q) = w_i^a(x_p, t_q)$, $\forall i = 1, 4$ and go to 3
Step 7: if $p = NS$ go to 8; if not do $p = p + 1$ and go to 3
Step 8: if $q > 0$, do $q = q - 1$ and go to 2; if not, do $\bar{w}_i^a(x_p, t_q) = w_i^a(x_p, t_q), \forall i = 1$, 4, $\forall p = 0$, NS; $\forall q = 0$, NT and stop.

Remark. The algorithm is very easy to program; it uses only "local" information (i.e. to compute $w^a(x_p, t_q)$ it uses only $w^a(x_{p-1}, t_q)$, $w^a(x_p, t_{q-1})$, $w^a(x_{p+1}, t_{q+1})$). So it is possible to implement it on computers of small central memory.

We recall also that the algorithm converges in a finite number of iterations to $\bar{w}^a$ and $\lim_{\|a\| \to 0} \max_{(x,t) \in Q} |\bar{w}_i^a(x, t) - V_i(x, t)| = 0$, with $\|a\|$ the norm of the triangulation Q^a.

3.5. Some remarks preceding the presentation of numerical results. We shall solve a problem involving a simplified model and the real demand data considered in [12]. The demand will have the following form shown in Fig. 5. The hydraulic cost will be

$$(3.14)\quad \begin{aligned} &c_h(x_h) = c_{h_1} + (c_{h_2} - c_{h_1}) x_h / x_{h\max}, \\ &c_{h_1} = 0.1, \quad c_{h_2} = 0.06, \quad x_{h\max} = 5000 \text{ MWh}. \end{aligned}$$

Other dates are:

$$(3.15)\quad P_1 \in [P_{1\min}, P_{1\max}] = [250, 500] \quad \text{(MW)},$$

$$(3.16)\quad P_2 \in [P_{2\min}, P_{2\max}] = [125, 250] \quad \text{(MW)},$$

$P_h \in [0, P_{h\max}(x_h)]$ with

$$(3.17)\quad P_{h\max} = \begin{cases} P_{h_1} & \forall s \cdot x_{h\max} \leqq x_h \leqq x_{h\max},\ s = 0.3,\ P_{h_1} = 250 \text{ MW}, \\ \dfrac{x_h}{s \cdot x_{h\max}} \cdot P_{h_1} & \forall 0 \leqq x_h \leqq s \cdot x_{h\max}, \end{cases}$$

$$(3.18)\quad A = 0 \text{ MW}.$$

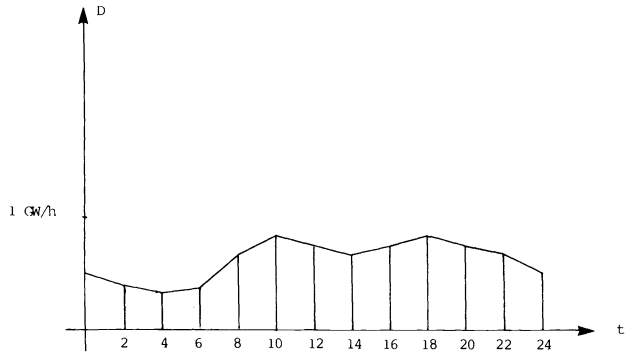


FIG. 5

Knowing $\bar{w}_i^a$, we obtain the approximate optimal feedback policies as follows:

a) We change our policy (that is to say, we start or we stop the operation of a thermic generator) in the regions where (3.11) becomes an equality. For example, if our system is in the state $(2, x_p, t_q)$ and $\bar{w}_2^a(x_p, t_q) = \bar{w}_4^a(x_p, t_q) + \bar{k}_2$ holds, we must pass to state 4 (starting with central 2). In this form we obtain tables of future states showing where and how we must switch.

b) When (3.11) are strict inequalities, we must define the production level of thermopower plants. The optimal level is that giving equality in (3.10) or (3.10′).

c) Finally with the tables concerning future states and optimal production levels we obtain the optimal trajectories solving the differential equation (3.3).

3.6. Numerical results. The interval of time $[0, T]$ is taken as 8 days. We use $NT = 192$; so the length of the discretized time is 1 hour. We use $NS = 18$; as $X_{h\max} = 5000$ MWh we obtain 277 MWh for each step. We divide each interval of thermal power production in six values ($P_1 = 250, 300, 350, 400, 450, 500$ MW; $P_2 = 125$, 150, 175, 200, 225, 250 MW). The values of the cost to start a power plant are $\bar{k}_1 = 0.5775 \times 10^5$, $\bar{k}_2 = 0.325 \times 10^5$.

In the algorithm the iterative part concerns steps 3 to 6. It converges in a finite number of iterations. An example is given in Table 3.1 (in which the values are divided by 10^5).

TABLE 3.1

Number of iterations	w_1	w_2	w_3	w_4
0	0	0	0	0
1	.325	.325	.325	.325
2	.650	.55917	.650	.55917
3	.975	.55917	.71641	.55917
4	.98241	.55917	.71641	.55917
5	.98241	.55917	.71641	.55917

The program gives as output suboptimal policies (future state E_f and production level to be generated) as functions of the state E_p, x_h of the system, as shown in Table 3.2.

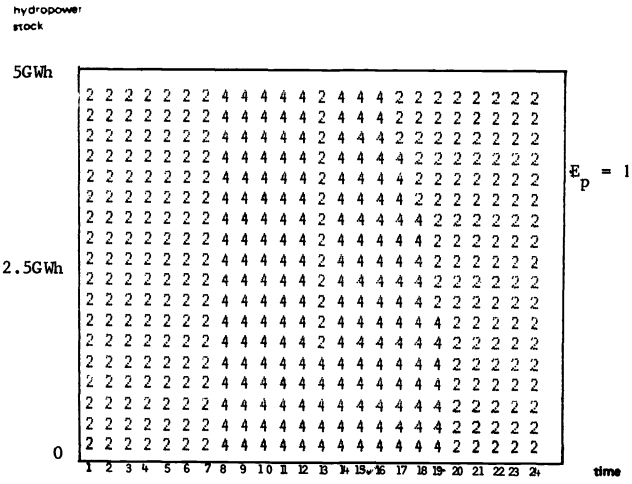
With these values we prepare the operation tables (as functions of (x_h, t)) that are shown in the following pages. We have chosen those referring to Wednesday. In one table of each pair we give the future state; in the other the power level to produce.

TABLE 3.2

E_p	E_f	P_h	P_1	P_2	P_3	W
1	2	.00000D+01	.40000D+06	.00000D+01	.11670D+05	.91130D+06
2	2	.00000D+01	.40000D+06	.00000D+01	.11670D+05	.85355D+06
3	4	.00000D+01	.30000D+06	.12500D+06	.00000D+01	.90549D+06
4	4	.00000D+01	.30000D+06	.12500D+06	.00000D+01	.84774D+06

TABLE 3.3

a. *Future states.*



b. *Power level to produce.*

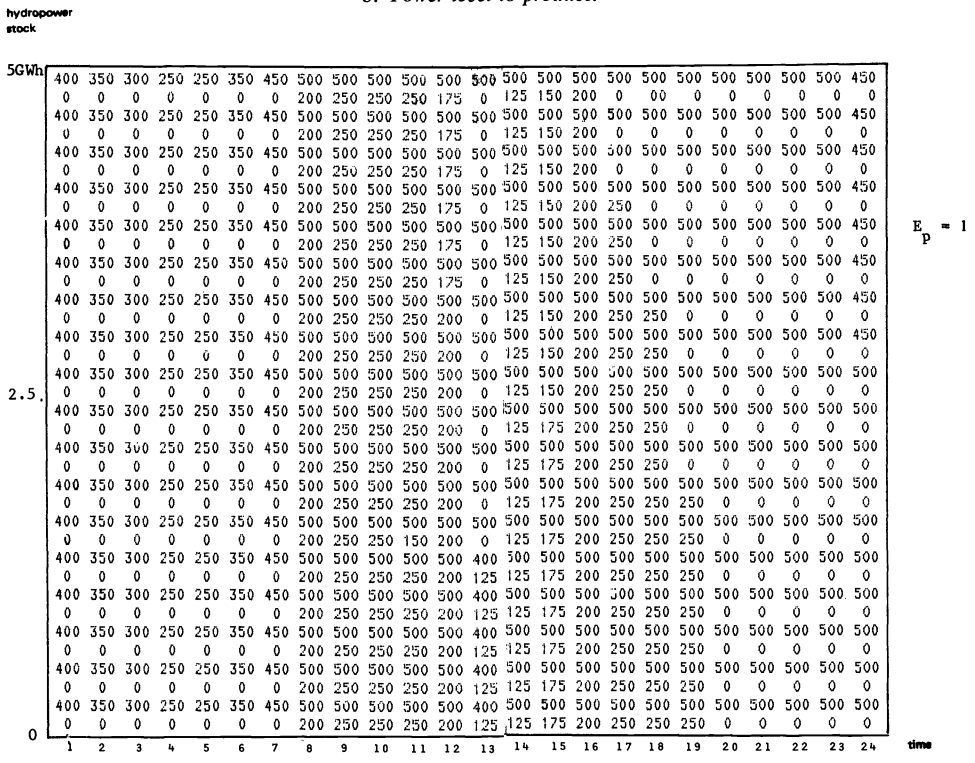
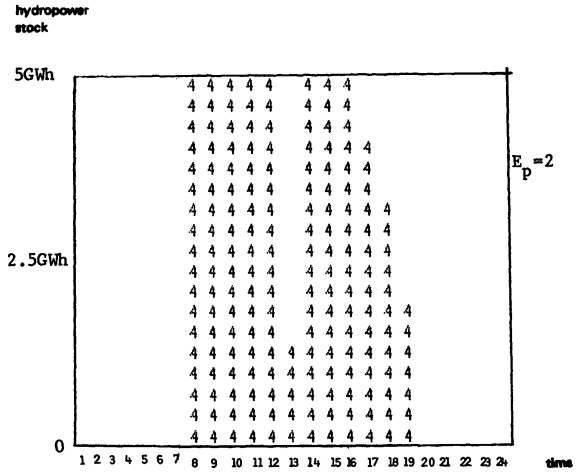


TABLE 3.4

a. *Future states.*



b. *Power level to produce.*

hydropower stock

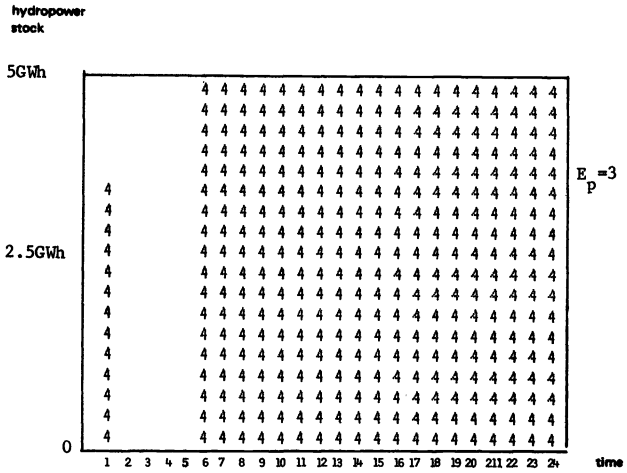
5GWh	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	450
	0	0	0	0	0	0	0	200	250	250	250	175	0	125	150	200	0	0	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	450
	0	0	0	0	0	0	0	200	250	250	250	175	0	125	150	200	0	0	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	450
	0	0	0	0	0	0	0	200	250	250	250	175	0	125	150	200	250	0	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	300	500	500	500	500	500	500	500	500	500	500	500	450
	0	0	0	0	0	0	0	200	250	250	250	175	0	125	150	200	250	0	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	450
	0	0	0	0	0	0	0	200	250	250	250	175	0	125	150	200	250	0	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	450
	0	0	0	0	0	0	0	200	250	250	250	175	0	125	150	200	250	250	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	0	125	150	200	250	250	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	0	125	150	200	250	250	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	0	125	150	200	250	250	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
2.5GWh	0	0	0	0	0	0	0	200	250	250	250	200	0	125	175	200	250	250	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	0	125	175	200	250	250	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	0	125	175	200	250	250	250	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	0	125	175	200	250	250	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	0	125	175	200	250	250	0	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	125	125	175	200	250	250	250	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	125	125	175	200	250	250	250	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	500
	0	0	0	0	0	0	0	200	250	250	250	200	125	125	175	200	250	250	250	0	0	0	0
	400	350	300	250	250	350	450	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	500
0	0	0	0	0	0	0	0	200	250	250	250	200	125	125	175	200	250	250	250	0	0	0	0
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

E_p=2

time

TABLE 3.5

a. *Future states.*



b. *Power level to produce.*

hydropower stock

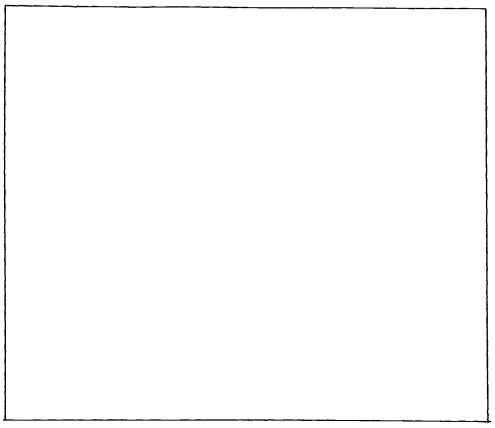
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
5GWh	0	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	250	250	250	250	250	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	0	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	250	250	250	250	250	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	0	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	250	250	250	250	250	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	0	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	250	250	250	250	250	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	0	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	250	250	250	250	250	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	250	250	250	250	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	250	250	250	250	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	150	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
2.5GWh	125	250	250	250	250	125	125	200	250	250	250	200	125	125	150	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	200	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	400	350
	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	0	0	0	0	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	400	350
0	125	250	250	250	250	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125

E_p=3

time

TABLE 3.6.

a. *Future states.*



b. *Power level to produce.*

hydropower stock																							
5GWh	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	175	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
2.5GWh	125	125	125	125	125	125	125	200	250	250	250	200	125	125	150	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	500	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	300	250	250	250	250	250	350	500	500	500	500	500	400	500	500	500	500	500	500	500	500	450	400	350
	125	125	125	125	125	125	125	200	250	250	250	200	125	125	175	200	250	250	250	250	175	125	125	125
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

$E_p = 4$

time

Example. If we are in table $E_p = 1$ of future states and in the intersection of lines (x_h, t) we read $E_f = 4$; meaning that at time t, if the stock of hydraulic energy is x_h, both plants 1, 2, must work. To know the production level we obtain from the table $E_p = 1$ of power production the two values; i.e. if we read $\frac{350}{125}$, that means $P_1 = 350$, $P_2 = 125$.

If in the table of future states there is no entry we must continue with the same state E_p.

4. Time of computation and final remarks. The algorithm converges independently of the choice of $(w^h)_0 \in W^h$ and of the order of numeration of the vertices. Nevertheless for each problem it is possible to analyse if some special choices give an improvement of the convergence (and of course, a reduction of the computation time).

In our problem the vertices were ordered in the sense of x increasing but in decreasing sense for the time.

On the other hand a choice for $(w^h)_0$ can always be the trivial choice $w_0^h \equiv -(M_l/\alpha + M_\phi)$; looking for better results, we made two choices:

first $(w_i^h)_0 \equiv 0$;

$$\text{second} \quad (\tilde{w}_i^h(x_p^h, t_q^h))_0 = \min_{\substack{i=1,4 \\ p=0,NS}} \bar{w}_i^h(x_p^h, t_{q+1}^h) \quad \forall q = 0, NT-1.$$

The latter choice was possible because $\bar{w}_i^h(x_p^h, t_{NT}^h) \equiv 0$ and functions $\bar{w}_i^h(\cdot, t_{q+1})$ can be computed, in our algorithm, after knowing $\bar{w}_i^h(\cdot, t_{q'})$, $q' = q+2, \cdots, NT$. Table 4.1 shows that for T (time of operation in hours of our electric system) increasing, the second choice is much better. Figure 6 points out the (linear and parabolic)

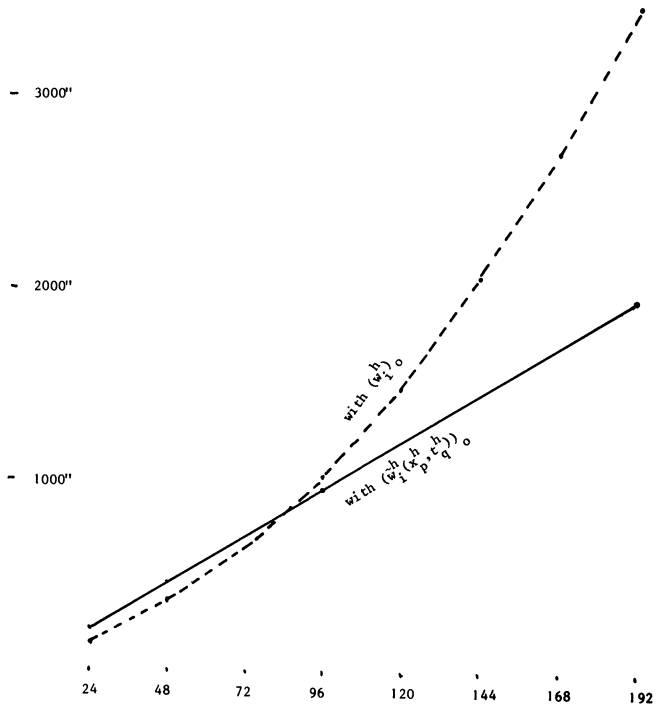


FIG. 6

TABLE 4.1

T	$w_0^h \equiv 0$	$(\tilde{w}_i^h)_0$
	computing time in seconds	
24	154	221
48	365	472
72	645	713
96	1,013	953
192	3,469	1,943

behaviour of the time of computation. We have used a PDP 11/23 (128 K, operative system RT 11/XM). As a final piece of information, using a minicomputer COMPUSYST 2000 (64 K, operating system CP/M, central proc. ZIL06-Z80) for $T = 24$ hours we need 380 seconds. (On the other hand, using a HB/68 DPS/Multics computer, we need, for $T = 192$ hours, only 48, 72 seconds.)

The example here presented required at run-time 2,500 memory positions. For one-day simulations 10,944 file positions are used.

Actually, the numerical approximation method (described in § 3.4) needs only 36 central memory positions for the calculation of w_h, at each discretization point.

These numbers show that with this method it is possible to increase the number of thermal power stations admitted in the system. First results have been recently obtained in this sense by M. C. Bancora-Imbert at INRIA.

With the use of large scale memory and special programming and simulation techniques further analysis is pursued to establish the extent and advantages of this procedure in applications.

REFERENCES

[1] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.

[2] ———, *Contrôle impulsionnel et inéquations quasi-variationnelles*, Dunod, Paris, 1982.

[3] W. FLEMING AND R. RISHEL, *Optimal Deterministic and Stochastic Control*, Springer-Verlag, New York, 1975.

[4] M. N. EL TARAZI, *Contraction et ordre partiel pour l'étude d'algorithmes synchrones et asynchrones en analyse numérique*, Thèse d'etat, Besançon, 1981.

[5] A. FRIEDMAN, *Differential Games*, Wiley Interscience, New York, 1971.

[6] R. GLOWINSKI, J. L. LIONS AND R. TREMOLIERES, *Analyse numérique des inéquations variationnelles*, Dunod, Paris, 1976.

[7] R. GONZALEZ, *Sur l'existence d'une solution maximale de l'équation de Hamilton-Jacobi.* CRAS Paris 1976, Ser. A, pp. 1287-1290.

[8] R. GONZALEZ AND E. ROFMAN, *An algorithm to obtain the maximum solution of the Hamilton-Jacobi equation*, Lectures Notes in Control and Information Sciences, 6, Springer-Verlag, New York, 1978.

[9] R. GONZALEZ, *Sur la résolution de l'équation de Hamilton-Jacobi du contrôle déterministe*, Thèse 3ème cycle, Université de Paris IX, 1980; Cahiers de Math. de la Décision, Ceremade 8029, 8029 bis.

[10] R. GONZALEZ AND E. ROFMAN, *Iterative computation of the maximal solution of the Hamilton-Jacobi equation*, V. Symposium of Op. Res., Verlag Anton Hain, Part I, R. Burkard and T. Ellinger, eds., Univ. Koln, Aug. 25-27, 1980.

[11] R. GONZALEZ, *Optimal bang-bang control policies with restrictions of time between switchings*, Matematicae Notae, XXVII (1979), pp. 111-137.

[12] C. LEGUAY, *Application du contrôle stochastique à un problème de gestion optimale d'énergie*, Thèse de Docteur-Ingénieur, Université de Pairs IX, 1975.

[13] P. L. LIONS AND B. MERCIER, *Approximation numérique des équations de Hamilton-Jacobi-Bellman*, RAIRO Anal. Numér., 14 (1980), pp. 369-393.

[14] J. P. QUADRAT AND F. DELEBECQUE, *Contribution of stochastic control singular perturbation averaging and team theories to an example of large scale systems: management of hydropower production*, IEEE. Trans. Automat. Control, AC-23 (1978), pp. 209-222.

[15] J. P. QUADRAT, F. DELEBECQUE, P. COLLETER AND F. FALGARONE, *Application du contrôle stochastique à la gestion des moyens de production d'énergie en Nouvelle Calédonie*, Proc. Recent Methods in Non linear Analysis, Pitagora editrice Bologna, 1978, pp. 465-497.

[16] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, Thèse d'etat, Université de Paris IX, 1977.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

STRUCTURAL STABILIZATION OF UNCERTAIN SYSTEMS: NECESSITY OF THE MATCHING CONDITION*

IAN R. PETERSEN†

Abstract. This paper investigates one aspect of the problem of stabilizing an uncertain linear system. That is, the systems under consideration contain uncertain parameters which are unknown but bounded. The question arises as to whether such a system can be stabilized via feedback control. In some of the previous papers in this area, the system is assumed to satisfy a so-called "matching-condition;" this type of assumption is used to assure that the uncertain system can be stabilized. It is known, however, that the matching condition is not a necessary condition for stabilizability. This paper introduces a strengthened notion of stabilizability referred to as *structural stabilizability via a nominally determined quadratic Lyapunov function.* It is then shown that if a system is to have this stronger property, the matching condition must necessarily be satisfied.

Key words. uncertain systems, stabilization, Lyapunov functions, matching conditions

1. System and introduction. This paper is concerned with a stabilization problem for an uncertain linear dynamical system described by

$$(\Sigma) \qquad \dot{x}(t) = [A_0 + \Delta A(r(t))]x(t) + B_0 u(t), \qquad r(t) \in \mathcal{R}$$

where $x(t) \in R^n$ is the *state*, $u(t) \in R^m$ is the *control*, $r(t) \in R^p$ is the *vector of uncertain parameters* and $\mathcal{R} \subset R^p$ is a compact *uncertainty bounding set.* It is assumed that the matrix function $\Delta A(\cdot)$ is continuous and that $\Delta A(\tilde{r}) = 0$ for some $\tilde{r} \in \mathcal{R}$. The function $r(\cdot)$ is restricted to be a Lebesgue measurable function such that $r(t) \in \mathcal{R}$ for all $t \geqq 0$ and it is furthermore assumed that $m < n$ and rank $B_0 = m$.

Remark. The analysis can also be extended to handle a class of uncertainties in the input matrix; i.e., in § 4, we consider the case when B_0 is replaced by "$B_0 + \Delta B$."

Associated with (Σ) above is a known system obtained when $\Delta A(r) \equiv 0$. This system, described by the state equation

$$\text{NOM } (\Sigma) \qquad \dot{x}(t) = A_0 x(t) + B_0 u(t)$$

is henceforth referred to as the *nominal system*, NOM (Σ).

When dealing with systems of the form (Σ), it is of interest to know whether asymptotic stability can be guaranteed via the application of a feedback control. That is, can a feedback control be found such that the closed loop system is asymptotically stable for any admissible uncertainty? In [1], NOM (Σ) is assumed to be stabilizable and it is shown that the satisfaction of a "matching condition" by $\Delta A(\cdot)$ is sufficient for stabilizability. More precisely, $\Delta A(\cdot)$ is said to satisfy the *matching condition* if there exists a continuous matrix function $D(\cdot): R^p \to R^{m \times n}$ such that

$$(1.1) \qquad \Delta A(r) = B_0 D(r)$$

for all $r \in R^p$. It should be noted that the matching condition above is only sufficient for stabilizability and not necessary. That is, there exist systems which fail to satisfy the matching condition and yet are stabilizable. Examples of such systems can be

* Received by the editors December 16, 1982, and in revised form March 12, 1984. This work was supported by the National Science Foundation under grant ECS-8108804. A preliminary version of this paper was presented at the 20th Allerton Conference on Communication, Control and Computing, University of Illinois.

† Department of Systems Engineering, Australian National University, Canberra, ACT 2601, Australia. This work was done while the author was at the Department of Electrical Engineering, University of Rochester.

found in [2]-[4]. This paper introduces a stronger notion of stabilizability for which the matching condition on $\Delta A(\cdot)$ is shown to be necessary.

In order to relate this work to existing results, it is important to point out one salient feature of [1]; that is, if (Σ) satisfies the matching condition, then stabilization can be achieved given any uncertainty bounding set $\mathscr{R}$ which is arbitrarily large but finite. The ability to tolerate arbitrarily large perturbations motivates our definition of *structural stabilizability.* This concept will be formally defined in the next section.

In [1]-[6], a quadratic Lyapunov function of the form $V(x)=x'Px$ is used to establish the stability of the closed loop system. Furthermore, the positive-definite matrix P is also used in the construction of the desired stabilizing feedback control law. The procedure given in [1] for the construction of the matrix P is as follows:

(i) Find an $m\times n$ matrix K such that $\bar{A}_0 \triangleq A_0+B_0K$ is a stability matrix.

(ii) Choose any positive-definite matrix Q and solve the Lyapunov equation

$$\bar{A}_0'P+P\bar{A}_0=-Q$$

to obtain the matrix P. Hence, we see that the Lyapunov function obtained by (i) and (ii) is solely determined by the system NOM (Σ). In this paper, we consider the problem of characterizing the class of systems (Σ) for which this "nominally determined" Lyapunov function will work; that is, for what class of systems (Σ) can one use a Lyapunov function generated from NOM (Σ)? Uncertain systems having this property have the advantage that a suitable Lyapunov function is straightforward to find. In the sequel, systems having this property are formally said to be *structurally stabilizable via a nominally determined quadratic Lyapunov function.* The main result of this paper, Theorem 3.1, can now be paraphrased in a rather compact manner: Namely, (Σ) is structurally stabilizable via a nominally determined quadratic Lyapunov function if and only if $\Delta A(\cdot)$ satisfies the matching condition and NOM (Σ) is stabilizable. In contrast to the earlier work dealing only with sufficiency of the matching condition, this paper establishes the fact that the matching condition on $\Delta A(\cdot)$ is also necessary.

2. Definitions and notation. The uncertain system (Σ) is said to be *quadratically stabilizable* if there exists a continuous feedback control function $p(\cdot): R^n \to R^m$ and an $n\times n$ positive-definite matrix P leading to the satisfaction of the following condition: The closed loop system

$$\dot{x}(t)=[A_0+\Delta A(r(t))]x(t)+B_0p(x(t)), \qquad r(t)\in\mathscr{R}$$

with Lyapunov function

$$V(x)=x'Px,$$

is uniformly asymptotically stable in the so-called *guaranteed sense.* That is, there exists a constant $\beta>0$ such that for any admissible vector uncertainty function $r(\cdot)$, the Lyapunov derivative admits the bound

$$\begin{aligned}-\beta\|x\|^2 \geqq \mathscr{L}(x,t) &\triangleq [\nabla V(x)]'[A_0x+\Delta A(r(t))x+B_0p(x)]\\ &=2x'P[A_0+\Delta A(r(t))]x+2x'PB_0p(x)\end{aligned}$$

for all pairs $(x,t)\in R^{n+1}$; see also [5].

If P is an $n\times n$ positive-definite symmetric matrix such that the above condition is satisfied, then the uncertain system (Σ) is said to be *quadratically stabilizable with Lyapunov function* $V(x)=x'Px$.

The uncertain system (Σ) is said to have *linear uncertainty structure* if the matrix function $\Delta A(\cdot)$ is linear. That is, $\Delta A(r)$ can be written in the form

$$\Delta A(r) = \sum_{i=1}^{p} A_i r_i$$

where $A_1, A_2, \cdots, A_p$ are constant $n \times n$ matrices and r_i is the ith component of the vector r.

2.1. Systems with linear uncertainty structure. The definitions presented in this subsection will apply only to systems with linear uncertainty structure. These definitions will be generalized in the next subsection.

For the linear case, the uncertain system (Σ) is said to be *structurally stabilizable* if for each compact set $\mathcal{R} \subset R^p$, the system (Σ), with this uncertainty bounding set, is quadratically stabilizable.

We now define a set $\mathcal{S}_{\text{nom}}$ which characterizes the set of nominally determined quadratic Lyapunov functions. For reasons which will become apparent in the next section, it is convenient to deal with the inverse of the Lyapunov matrix rather than the Lyapunov matrix itself. Hence, we define $\mathcal{S}_{\text{nom}}$ as follows: A positive-definite symmetric matrix S is in the set $\mathcal{S}_{\text{nom}}$ if there exists an $m \times n$ gain matrix K such that $\bar{A}_0 \triangleq A_0 + B_0 K$ has strict left half-plane eigenvalues and $\bar{A}_0' S^{-1} + S^{-1}\bar{A}_0$ is negative-definite. (Note that if NOM (Σ) is stabilizable, then the set $\mathcal{S}_{\text{nom}}$ will be nonempty.)

The uncertain system (Σ) is said to be *structurally stabilizable via a nominally determined quadratic Lyapunov function* if for each compact $\mathcal{R} \subset R^p$ and each $S \in \mathcal{S}_{\text{nom}}$, the system (Σ) with uncertainty bounding set $\mathcal{R}$, is quadratically stabilizable with Lyapunov function $V(x) = x' S^{-1} x$.

2.2. Systems with nonlinear uncertainty structure. We now concern ourselves with systems which do not necessarily have linear uncertainty structure. The motivation for our more general definition of structural stabilizability stems from the fact that it is more convenient to work with perturbations in the matrix $\Delta A(r)$ rather than perturbations in the parameter vector r, especially when $\Delta A(r)$ depends nonlinearly on r. Given these possible nonlinear dependencies on the uncertain parameters, there is a possibility that the matrix function $\Delta A(\cdot)$ may be such that $\Delta A(r)$ remains bounded even if $\|r\| \to \infty$. Therefore, an arbitrarily large bound on the uncertain parameter vector r, may not result in arbitrarily large perturbations $\Delta A(r)$. To circumvent this technical difficulty, we associate the system (Σ) with a class of systems with linear uncertainty structure. This class will be denoted SPAN (Σ) and the definition is motivated by one simple fact: Instead of dealing with the admissible set of uncertain parameters $\mathcal{R}$, it is equivalent to deal with the set

$$\Delta A(\mathcal{R}) \triangleq \{\Delta A(r) : r \in \mathcal{R}\} \subset R^{n \times n}.$$

We now provide a definition.

Given any set of basis matrices $\{A_1, A_2, \cdots, A_k\}$ for the linear space[1] span $\Delta A(R^p) \subset R^{n \times n}$, we can generate a system (Σ_l) in the class SPAN (Σ) as follows: Define a linear matrix function $\Delta A_l(\cdot) : R^k \to R^{n \times n}$ by

$$\Delta A_l(\alpha) \triangleq \sum_{i=1}^{k} A_i \alpha_i$$

[1] Given any set $G \subset R^n$, span G is defined by span $G \triangleq \{x \in R^n : x = \sum_{i=1}^{k} \alpha_i g_i$, where $\{g_i\}_{i=1}^{k} \subset G$ and $\alpha_i \in R$ for $i = 1, 2, \cdots, k\}$.

where α_i is the ith component of the vector $\alpha \in R^k$. Then the system (Σ_l) is described by the state equation

$$(\Sigma_l) \qquad \dot{x}(t) = [A_0 + \Delta A_l(\alpha(t))]x(t) + B_0 u(t).$$

Given any compact uncertainty bounding set $\mathscr{A} \subset R^k$, the vector function $\alpha(\cdot)$ is restricted to be a measurable function such that $\alpha(t) \in \mathscr{A}$ for all $t \geqq 0$. We now make use of the class SPAN (Σ) in the following definition.

The system (Σ) is said to be *structurally stabilizable* if each system (with linear uncertainty structure) $(\Sigma_l) \in$ SPAN (Σ) is structurally stabilizable according to the definition given in § 2.1.

It follows from the above definition of SPAN (Σ) that the set $\Delta A_l(R^k)$ is independent of the system $(\Sigma_l) \in$ SPAN (Σ). Hence, it is straightforward to verify that if there exists one system $(\Sigma_l) \in$ SPAN (Σ) which is structurally stabilizable, then every system $(\Sigma_l) \in$ SPAN (Σ) will be structurally stabilizable.

The system (Σ) is said to be *structurally stabilizable via a nominally determined quadratic Lyapunov function* if each system (with linear uncertainty structure) $(\Sigma_l) \in$ SPAN (Σ) is structurally stabilizable via a nominally determined quadratic Lyapunov function. Again it is straightforward to verify that if there exists one system $(\Sigma_l) \in$ SPAN (Σ) which is structurally stabilizable via a nominally determined quadratic Lyapunov function, then every system $(\Sigma_l) \in$ SPAN (Σ) will have this property.

2.3. Notation to be used in the sequel. We let Θ be any matrix whose columns form a set of basis vectors for the linear space

$$\mathscr{N}[B_0'] \triangleq \{x \in R^n : B_0' x = 0\}.$$

Let $\mathscr{V}_k$ denote the inner-product space of symmetric $k \times k$ matrices; e.g., see [6]. Given any matrix $M \in \mathscr{V}_k$, $\lambda_{\max}[M]$ will denote the maximum eigenvalue of M.

3. The main result. In this section we present the main result of this paper.

THEOREM 3.1. *The uncertain system* (Σ) *is structurally stabilizable via a nominally determined quadratic Lyapunov function if and only if the following conditions hold*: $\Delta A(\cdot)$ *satisfies matching condition* (1.1) *and* NOM (Σ) *is stabilizable.*

Before proving this theorem, we first establish some preliminary results.

Our basic concern is to establish the necessity of the matching condition. As far as sufficiency is concerned, we shall use the following lemma which is an immediate consequence of Theorem 4.1 of [3].

LEMMA 3.1. *Suppose that* $\Delta A(\cdot)$ *satifies matching condition* (1.1) *and* NOM (Σ) *is stabilizable. Then the system* (Σ) *is structurally stabilizable via a nominally determined quadratic Lyapunov function.*

LEMMA 3.2. *The matrix function* $\Delta A(\cdot)$ *satisfies the matching condition* (1.1) *if and only if every system* (*with linear uncertainty structure*) $(\Sigma_l) \in$ SPAN (Σ) *is such that the corresponding matrix function* $\Delta A_l(\cdot)$ *satisfies this same matching condition.*

Proof (*Necessity*). If $\Delta A(\cdot)$ satisfies the matching condition, then there exists a continuous matrix function $D(\cdot)$ such that

$$\Delta A(r) = B_0 D(r)$$

for all $r \in R^p$. Therefore

$$\text{span } \Delta A(R^p) = \text{span } B_0 D(R^p) = B_0 \text{ span } D(R^p).$$

Hence, if $\{A_1, A_2, \cdots, A_k\}$ is any set of basis matrices for the space span $\Delta A(R^p)$, we must have

$$A_i \in B_0 \text{ span } D(R^p)$$

for $i = 1, 2, \cdots, k$. Therefore, there exist matrices $D_1, D_2, \cdots, D_k$ in span $D(R^p)$ such that

$$A_i = B_0 D_i$$

for $i = 1, 2, \cdots, k$. This implies that, for the system $(\Sigma_l) \in \text{SPAN}(\Sigma)$ corresponding to this set of basis matrices,

$$\Delta A_l(\alpha) = \sum_{i=1}^{k} B_0 D_i \alpha_i = B_0 \sum_{i=1}^{k} D_i \alpha_i$$

for all vectors $\alpha \in R^k$. Therefore, if we define the continuous matrix function

$$D_l(\alpha) \triangleq \sum_{i=1}^{k} D_i \alpha_i,$$

it is apparent that

$$\Delta A_l(\alpha) = B_0 D_l(\alpha)$$

for all vectors $\alpha \in R^k$. That is, $\Delta A_l(\cdot)$ satisfies the matching condition (1.1).

(*Sufficiency*). We assume that each system (Σ_l) in the class SPAN (Σ) is such that the corresponding matrix function $\Delta A_l(\cdot)$ satisfies matching condition (1.1). Now let $(\Sigma_l) \in \text{SPAN}(\Sigma)$ be one such system and note that the matching condition implies that there exists a continuous matrix function $D_l(\cdot)$ such that

$$\Delta A_l(\alpha) = B_0 D_l(\alpha)$$

for all vectors $\alpha \in R^k$. However, given any matrix $\Delta A \in \text{span } \Delta A(R^p)$ there exists a vector $\alpha \in R^k$ such that

$$\Delta A = \Delta A_l(\alpha).$$

Therefore, any matrix $\Delta A \in \text{span } \Delta A(R^p)$ can be written in the form

$$\Delta A = B_0 D$$

where D is a constant $m \times n$ matrix. That is, for each $r \in R^p$, there exists an $m \times n$ matrix $D(r)$ such that

$$\Delta A(r) = B_0 D(r).$$

Furthermore, using the fact that rank $B_0 = m$, it is clear that the matrix function $D(\cdot)$ is defined uniquely by the expression

$$D(r) = (B_0' B_0)^{-1} B_0' \Delta A(r).$$

The required continuity of $D(\cdot)$ is a consequence of the continuity of $\Delta A(\cdot)$ and moreover,

$$\Delta A(r) = B_0 D(r)$$

for all $r \in R^p$. Hence the matching condition is satisfied. □

We shall exploit the following lemma which is established in [2]; for the sake of completeness we include the proof.

LEMMA 3.3. *Suppose that the uncertain system* (Σ) *is quadratically stabilizable with Lyapunov function* $V(x) = x'S^{-1}x$. *Then*

$$\eta'\Theta'(A_0S + SA_0')\Theta\eta + \eta'\Theta'(\Delta A(r)S + S\Delta A'(r))\Theta\eta < 0 \tag{3.1}$$

for all nonzero vectors $\eta \in R^{n-m}$ *and all* $r \in \mathcal{R}$.

Proof. Using the assumed quadratic stabilizability (defined in the previous section), there exists a constant $\beta > 0$ such that

$$2x'S^{-1}[A_0 + \Delta A(r)]x + 2x'B_0p(x) \leqq -\beta\|x\|^2$$

for all $x \in R^n$ and all $r \in \mathcal{R}$. If we let $y = S^{-1}x$, it is clear that

$$\begin{aligned} -\beta\|Sy\|^2 &\geqq 2y'[A_0 + \Delta A(r)]Sy + 2y'B_0p(Sy) \\ &= y'[A_0S + SA_0']y + y'[\Delta A(r)S + S\Delta A'(r)]y + 2y'B_0p(Sy) \end{aligned}$$

for all $y \in R^n$ and all $r \in \mathcal{R}$. In particular, this inequality must hold for all vectors $y \in \mathcal{N}[B_0']$ such that $y \neq 0$; that is, it must be true that

$$y'[A_0S + SA_0']y + y'[\Delta A(r)S + S\Delta A'(r)]y < 0$$

for all nonzero vectors $y \in \mathcal{N}[B_0']$ and for all $r \in \mathcal{R}$. Now, we make use of the fact that any vector $y \in \mathcal{N}[B_0']$ can be represented as $y = \Theta\eta$ for some appropriate $\eta \in R^{n-m}$. (Recall that the columns of Θ form a basis $\mathcal{N}[B_0']$.) Replacing y by $\Theta\eta$ in the preceding inequality yields the desired result. □

LEMMA 3.4. *The set* $\mathcal{S}_{\text{nom}}$ *is an open set in the space* $\mathcal{V}_n$.

Proof. Suppose $S_0 \in \mathcal{S}_{\text{nom}}$. Then according to the definition of $\mathcal{S}_{\text{nom}}$, there exists an $m \times n$ matrix K such that $\bar{A}_0 = A_0 + B_0K$ has strict left half-plane eigenvalues and the matrix $\bar{A}_0'S_0^{-1} + S_0^{-1}\bar{A}_0$ is negative-definite. Therefore, the matrix $\bar{A}_0S_0 + S_0\bar{A}_0'$ is also negative-definite and it follows that

$$\lambda_{\max}[\bar{A}_0S_0 + S_0\bar{A}_0'] < 0.$$

Now, using the continuity of the $\lambda_{\max}[\cdot]$ function, it follows that there exists a constant $\delta_1 > 0$ such that if $S \in \mathcal{V}_n$ satisfies $\|S - S_0\| < \delta_1$, then $\bar{A}_0S + S\bar{A}_0'$ is negative-definite. Furthermore, we note that the matrix S_0 is positive-definite and the set of positive-definite matrices is an open set in the space $\mathcal{V}_n$; e.g. see [6]. Therefore, there exists a second constant $\delta_2 > 0$ such that if $S \in \mathcal{V}_n$ satisfies $\|S - S_0\| < \delta_2$, then S is a positive-definite matrix. Hence, $\|S - S_0\| < \min\{\delta_1, \delta_2\}$ implies that $S \in \mathcal{S}_{\text{nom}}$. Therefore, $\mathcal{S}_{\text{nom}}$ is an open set in the space $\mathcal{V}_n$. □

LEMMA 3.5. *Suppose that the system* (Σ) *is structurally stabilizable via a nominally determined quadratic Lyapunov function. Then*

$$\text{span } \mathcal{S}_{\text{nom}} = \mathcal{V}_n.$$

Proof. We recall that $\Delta A(r) = 0$ for some $r \in \mathcal{R}$. This, together with the fact that the uncertain system (Σ) is structurally stabilizable implies that the system NOM (Σ) is stabilizable. Therefore, the set $\mathcal{S}_{\text{nom}}$ is nonempty. Lemma 3.4 states that the set $\mathcal{S}_{\text{nom}}$ is an open set. Now, given any matrix $W \in \mathcal{V}_n$ and a matrix $X \in \mathcal{S}_{\text{nom}}$, one can choose $\delta > 0$ sufficiently small so as to ensure that

$$Z \triangleq X + \delta W \in \mathcal{S}_{\text{nom}}.$$

Therefore,

$$W = \frac{Z}{\delta} - \frac{X}{\delta} \in \text{span } \mathcal{S}_{\text{nom}}.$$

It now follows that

$$\text{span } \mathscr{S}_{\text{nom}} = \mathscr{V}_n. \qquad \square$$

LEMMA 3.6. *Suppose that the system* (Σ) *is structurally stabilizable via a nominally determined quadratic Lyapunov function. Then, given any system* $(\Sigma_l) \in \text{SPAN}(\Sigma)$, *the corresponding set of basis matrices* $\{A_1, A_2, \cdots, A_k\}$ *satisfies the following condition*:

$$\Theta'[A_i S + S A_i']\Theta = 0 \tag{3.2}$$

for all $S \in \mathscr{V}_n$ *and* $i = 1, 2, \cdots, k$.

Proof. We will first establish (3.2) for all matrices $S \in \mathscr{S}_{\text{nom}}$. Subsequently, it will be shown that (3.2) must hold for all matrices $S \in \mathscr{V}_n$.

Since the system (Σ) is structurally stabilizable via a nominally determined quadratic Lyapunov function, it is apparent that all systems in the class SPAN (Σ) must also have this property. Indeed, let (Σ_l) be any system in the class SPAN (Σ). Then, for each compact set $\mathscr{A} \subset R^k$ and each matrix $S \in \mathscr{S}_{\text{nom}}$, the system (Σ_l) with uncertainty bounding set $\mathscr{A}$ is quadratically stabilizable with Lyapunov function $V(x) = x'S^{-1}x$. We now apply Lemma 3.3 and infer the following: Given any compact set $\mathscr{A} \subset R^k$ and matrix $S \in \mathscr{S}_{\text{nom}}$

$$\eta'\Theta'[A_0 S + S A_0']\Theta\eta + \eta\Theta'[\Delta A_l(\alpha)S + S\Delta A_l'(\alpha)]\Theta\eta < 0$$

for all $\alpha \in \mathscr{A}$ and all nonzero vectors $\eta \in R^{n-m}$. Equivalently

$$\eta'\Theta'[A_0 S + S A_0']\Theta\eta + \max_{\alpha \in \mathscr{A}} \{\eta'\Theta'[\Delta A_l(\alpha)S + S\Delta A_l'(\alpha)]\Theta\eta\} < 0 \tag{3.3}$$

for all nonzero vectors $\eta \in R^{n-m}$. In particular, if we take the set $\mathscr{A}$ to be the hypercube $\{\alpha \in R^k\colon \alpha = (\alpha_1, \alpha_2, \cdots, \alpha_k)', \max_i |\alpha_i| \leqq \bar{\alpha}\}$ and use the linear structure of $\Delta A_l(\cdot)$, inequality (3.3) leads to the following statement: Given any matrix $S \in \mathscr{S}_{\text{nom}}$, any nonzero vector $\eta \in R^{n-m}$ and any $\bar{\alpha} > 0$,

$$\eta'\Theta'[A_0 S + S A_0']\Theta + \bar{\alpha} \sum_{i=1}^{k} |\eta'\Theta'[A_i S + S A_i']\Theta\eta| < 0 \tag{3.4}$$

where $\{A_1, A_2, \cdots, A_k\}$ is the set of basis matrices corresponding to the system (Σ_l). Since $\bar{\alpha}$ can be arbitrarily large, the only way inequality (3.4) can hold is if

$$\Theta'[A_i S + S A_i']\Theta = 0 \tag{3.5}$$

for all $S \in \mathscr{S}_{\text{nom}}$ and $i = 1, 2, \cdots, k$.

The next part of this proof involves extending (3.5) to all matrices $S \in \mathscr{V}_n$. For each integer $i = 1, 2, \cdots, k$, we define a linear operator $\mathscr{A}_i\colon \mathscr{V}_n \to \mathscr{V}_{n-m}$ by

$$\mathscr{A}_i(S) \triangleq \Theta'[A_i S + S A_i']\Theta.$$

Let $\mathscr{N}[\mathscr{A}_i]$ denote the nullspace of the operator $\mathscr{A}_i$. Using this notation and (3.5), it is clear that

$$\mathscr{S}_{\text{nom}} \subseteq \mathscr{N}[\mathscr{A}_i]$$

for $i = 1, 2, \cdots, k$. Hence,

$$\mathscr{S}_{\text{nom}} \subseteq \bigcap_{i=1}^{k} \mathscr{N}[\mathscr{A}_i].$$

Note, however, that $\cap_{i=1}^{k} \mathcal{N}[\mathcal{A}_i]$ is a subspace of $\mathcal{V}_n$ and hence

$$\text{span } \mathcal{S}_{\text{nom}} \subseteq \bigcap_{i=1}^{k} \mathcal{N}[\mathcal{A}_i]. \tag{3.6}$$

Invoking Lemma 3.5, we can replace (3.6) by

$$\mathcal{V}_n \subseteq \bigcap_{i=1}^{k} \mathcal{N}[\mathcal{A}_i].$$

Therefore,

$$\Theta'[A_i S + S A_i]\Theta = 0$$

for all $S \in \mathcal{V}_n$ and $i = 1, 2, \cdots, k$. □

LEMMA 3.7. *Suppose that the system* (Σ) *is structurally stabilizable via a nominally determined quadratic Lyapunov function. Then given any system* $(\Sigma_l) \in \text{SPAN}(\Sigma)$, *the corresponding set of basis matrices* $\{A_1, A_2, \cdots, A_k\}$ *satisfies the following condition*:

$$\Theta' A_i = 0 \tag{3.7}$$

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

Proof. We proceed by contradiction. Suppose that (3.7) does not hold for some $i \in \{1, 2, \cdots, k\}$. Then, if we let θ_j denote the jth column of matrix Θ, there exists some $j \in \{1, 2, \cdots, n-m\}$ such that

$$c_j \triangleq \theta_j' A_i \neq 0.$$

Let

$$\mathcal{N}[c_j'] \triangleq \{x \in R^n : c_j' x = 0\}$$

and

$$\mathcal{N}[\theta_j'] \triangleq \{x \in R^n : \theta_j' x = 0\}.$$

Since $c_j \neq 0$ and $\theta_j \neq 0$, the two sets $\mathcal{N}[c_j']$ and $\mathcal{N}[\theta_j']$ are $(n-1)$-dimensional subspaces of R^n. Therefore, the set $\mathcal{N}[c_j'] \cup \mathcal{N}[\theta_j'] \neq R^n$. Hence, there exists a vector $v \in R^n$ such that $v' c_j \neq 0$ and $\theta_j' v \neq 0$. We now define the symmetric matrix

$$S^* \triangleq \frac{v v'}{v' c_j}.$$

Notice that this matrix has the property that

$$v = S^* c_j$$

and moreover, the (j, j)th element of the matrix $\Theta'[A_i S^* + S^* A_i']\Theta$ is

$$\theta_j'[A_i S^* + S^* A_i']\theta_j = 2\theta_j' S^* c_j = 2\theta_j' v \neq 0.$$

This implies that the matrix $\Theta'[A_i S^* + S^* A_i']\Theta$ is nonzero which contradicts Lemma 3.6. □

We are now in a position to prove the main result of this paper.

Proof of Theorem 3.1 (*Sufficiency*). To establish sufficiency, we assume that $\Delta A(\cdot)$ satisfies the matching condition (1.1) and that NOM (Σ) is stabilizable; it must be shown that the system (Σ) is structurally stabilizable via a nominally determined quadratic Lyapunov function. Indeed, let (Σ_l) be any system in the class SPAN (Σ). Since $\Delta A(\cdot)$ satisfies the matching condition, Lemma 3.2 states that $\Delta A_l(\cdot)$ will also

satisfy the matching condition. Since NOM (Σ_l) = NOM (Σ), stabilizability of NOM (Σ) implies stabilizability of NOM (Σ_l). Invoking Lemma 3.1, it follows that the system (Σ_l) is structurally stabilizable via a nominally determined quadratic Lyapunov function. However, recalling that (Σ_l) is an arbitrary system in the class SPAN (Σ), it must follow that the system (Σ) is structurally stabilizable via a nominally determined quadratic Lyapunov function.

(*Necessity*). To establish necessity, we suppose that the system (Σ) is structurally stabilizable via a nominally determined quadratic Lyapunov function; it must be shown that $\Delta A(\cdot)$ satisfies the matching condition (1.1) and that NOM (Σ) is stabilizable. Taking note of Lemma 3.5, it is clear that structural stabilizability of (Σ) implies that the system NOM (Σ) is stabilizable.

Next, let $(\Sigma_l) \in$ SPAN (Σ) be a given system with linear uncertainty structure. It follows from Lemma 3.7 that

$$\Theta' A_i = 0 \tag{3.8}$$

for $i = 1, 2, \cdots, k$. Recalling that rank $B_0 = m$ and that the columns of Θ span $\mathcal{N}[B_0']$, it is clear that the $n \times n$ matrix $[B_0 \vdots \Theta]$ is nonsingular. Let the inverse of this matrix be partitioned as

$$\left[\frac{F}{G}\right] \triangleq [B_0 \vdots \Theta]^{-1}$$

where F is an $m \times n$ matrix and G is an $(n-m) \times n$ matrix. As a consequence of this partitioning, it is apparent that

$$B_0 F + \Theta G = I \tag{3.9}$$

and

$$GB_0 = 0. \tag{3.10}$$

Letting g_j' denote the jth row of G, (3.10) is tantamount to

$$B_0' g_j = 0$$

for $j = 1, 2, \cdots, n-m$. Therefore, $g_j \in \mathcal{N}[B_0']$ for $j = 1, 2, \cdots, n-m$ and hence, each g_j can be written in the form $g_j = \Theta h_j$ for appropriate $h_j \in R^{n-m}$. This implies that the matrix G can be expressed as

$$G = H\Theta' \tag{3.11}$$

where H has the vectors h_j, $j = 1, 2, \cdots, n-m$ as its rows.

For each $i \in \{1, 2, \cdots, k\}$, we define the matrix

$$D_i \triangleq FA_i.$$

We now claim that $A_i = B_0 D_i$ for $i = 1, 2, \cdots, k$. Indeed, let $i \in \{1, 2, \cdots, k\}$ be fixed. It now follows from (3.9) that

$$B_0 FA_i + \Theta GA_i = A_i$$

and combining this with (3.8) and (3.11), we obtain

$$A_i = B_0 FA_i = B_0 D_i.$$

Therefore, for the system (Σ_l), the function $\Delta A_l(\cdot)$ is given by

$$\Delta A_l(\alpha) = \sum_{i=1}^{k} A_i \alpha_i = B_0 \sum_{i=1}^{k} D_i \alpha_i \triangleq B_0 D_l(\alpha).$$

where α_i denotes the ith component of the vector $\alpha \in R^k$.

We can now conclude that $\Delta A_l(\cdot)$ satisfies the matching condition. By applying Lemma 3.2 it follows that $\Delta A(\cdot)$ also satisfies the matching condition. □

4. Input connection uncertainty. If one includes input connection uncertainty in the system (Σ), the state equations become

$$(\Sigma_*) \quad \dot{x}(t) = [A_0 + \Delta A(r(t))]x(t) + [B_0 + \Delta B(s(t))]u(t); \qquad r(t) \in \mathcal{R}, \qquad s(t) \in \mathcal{S}$$

where $s(t) \in R^q$ is the vector of *input connection uncertainty parameters* and $\mathcal{S} \subset R^q$ is a compact bounding set. It is assumed that the matrix function $\Delta B(\cdot)$ is continuous and that $s(\cdot)$ is a Lebesgue measurable function such that $s(t) \in \mathcal{S}$ for all $t \geqq 0$.

Previous authors dealing with systems of the form (Σ_*) have also required that $\Delta B(\cdot)$ satisfy a matching condition; e.g., see [1] and [7]. This condition is described as follows: there exists a continuous matrix function $E(\cdot): \mathcal{S} \to R^{m \times m}$ such that for all $s \in \mathcal{S}$

$$\Delta B(s) = B_0 E(s)$$

and

$$\|E(s)\| < 1.$$

To extend the analysis of §§ 1-3 to handle (Σ_*), we make some observations:

(i) The presence of $\Delta B(\cdot)$ only influences the proof of Lemma 3.3.

(ii) Under the strengthened hypothesis that $\Delta B(\cdot)$ satisfies the above matching condition, Theorem 3.1 remains valid; the proof of this theorem remains the same.

5. Illustrative example. To illustrate the results of this paper, we consider the RLC electrical circuit shown in Fig. 1.

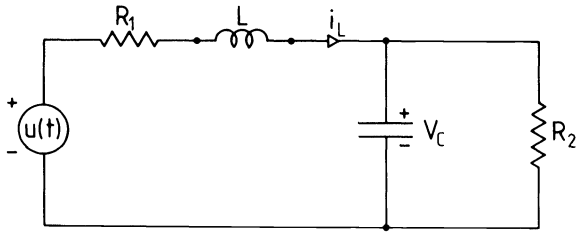


FIG. 1

We define state variables $x_1 \triangleq v_c$ and $x_2 \triangleq i_L$ where v_c is the voltage across the capacitor and i_L is the current through the inductor. Using Kirchhoff's laws, we arrive at the state equations

$$\begin{aligned} \dot{x}(t) &= \frac{-1}{R_2 C} x_1(t) + \frac{1}{C} x_2(t), \\ \dot{x}(t) &= \frac{-1}{L} x_1(t) - \frac{R_1}{L} x_2(t) + \frac{1}{L} u(t). \end{aligned} \tag{5.1}$$

Suppose now that the resistance R_1 is regarded as being an uncertain parameter and the values of the other parameters are fixed at $R_2 = L = C = 1$. Furthermore, assume that $0 \leqq R_1 \leqq \bar{R}$ where $\bar{R} > 0$ is pre-specified. Hence, (4.1) can be viewed as an uncertain system described by

$$(\Sigma_1) \qquad \begin{aligned} \dot{x}_1(t) &= -x_1(t) + x_2(t), \\ \dot{x}_2(t) &= -x_1(t) - R_1(t)x_2(t) + u(t), \qquad R_1(t) \in [0, \bar{R}]. \end{aligned}$$

The nominal system associated with this system is described by the state equations

$$\text{NOM } (\Sigma_1) \qquad \begin{aligned} \dot{x}_1(t) &= -x_1(t) + x_2(t), \\ \dot{x}_2(t) &= -x_1(t) + u(t). \end{aligned}$$

It is straightforward to verify that the matrix function $\Delta A(\cdot)$ associated with the system (Σ_1) satisfies matching condition (1.1). Furthermore, since the system NOM (Σ_1) is stabilizable, it follows from Theorem 3.1 that the uncertain system (Σ_1) is structurally stabilizable via a nominally determined quadratic Lyapunov function. Consequently, when stabilizing the system (Σ_1), the Lyapunov function to be used may be obtained by considering only the nominal system NOM (Σ_1). On the other hand, if R_2 happens to be an uncertain parameter, it is straighforward to verify that matching condition (1.1) will be violated. In this situation, the system is not structurally stabilizable via a nominally determined Lyapunov function.

Acknowledgments. Discussions with Professor B. R. Barmish and Mr. C. V. Hollot are gratefully acknowledged.

REFERENCES

[1] G. LEITMANN, *Guaranteed asymptotic stability for some linear systems with bounded uncertainties*, ASME J. Dynamical Systems, Measurement and Control, 101 (1979), pp. 212-216.

[2] C. V. HOLLOT AND B. R. BARMISH, *Optimal quadratic stabilizability of uncertain linear systems*, Proc. 18th Allerton Conference on Communications, Control and Computing, Univ. Illinois, Monticello, 1980.

[3] J. S. THORP AND B. R. BARMISH, *On guaranteed stability of uncertain linear systems via linear control*, J. Optim. Theory Appl., 35 (1981), pp. 559-579.

[4] B. R. BARMISH, *Necessary and sufficient conditions for quadratic stabilizability of an uncertain linear system*, J. Optim. Theory Appl., to appear.

[5] ———, *Stabilization of uncertain systems via linear control*, Proc. IEEE Asilomar Conference on Computers, Circuits and Systems, Monterey, CA, 1981.

[6] A. BERMAN, *Cones, Matrices and Mathematical Programming*, Springer-Verlag, Berlin, 1973.

[7] B. R. BARMISH, I. R. PETERSEN AND A. FEUER, *Linear ultimate boundedness control of uncertain dynamical systems*, Automatica, 19 (1983), pp. 523-532.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

008

ON FINITE VOLTERRA SERIES AND A THEOREM OF P. CROUCH*

I. A. K. KUPKA†

Abstract. In this paper, we give a new proof of Crouch's theorem on the realisation of finite Volterra series and extend it to the C^∞ case.

Key words. finite Volterra series, nilpotent Lie groups, representations, Lie algebras

AMS(MOS) subject classifications. Primary 93B15, B20; secondary 22E25, E27

Introduction. In [C], P. Crouch proved an interesting theorem about the reduction of real analytic input-output systems having a finite Volterra series. In this paper, we give a new proof of Crouch's theorem and extend it to the infinitely differentiable case.

Our method, related to the one introduced in [F-K], uses the "observation space" ([F-K]), that is the functional space generated by the output function under the action of the Lie algebra of the system. We show that when the Volterra series is finite, this action factors through the representation of a linear solvable Lie algebra on a finite-dimensional vector space. The theorem then follows easily from elementary results on the structure of orbits of linear unipotent groups (see [P, pp. 90–91]). We get a concrete representation of the minimal realisation in the observation space and avoid the use of the deep and difficult theorem of Malcev on the structure of nilmanifolds in Crouch's proof (see Raghunathan, *Discrete subgroups of Lie groups*, Springer Ergebnisse Bd 68, Chap. II).

Finally, we get a canonical form for the system on the minimal realisation which is "real upper triangular", that is, it is upper triangular except for some 2×2 blocks on the diagonal corresponding to the complex roots of the representation of the solvable algebra.

1. Statement of the main results. In this paper we shall consider C^∞ or C^ω (real analytic) input-output systems whose state space is a C^∞ or C^ω connected manifold M, whose control and output spaces are respectively the euclidean spaces R^m and R^c and whose dynamic is described by the following control-affine system:

$$\text{(I)}\qquad \begin{aligned} &\frac{dx(t)}{dt} = X[x(t)] + \sum_{j=1}^{m} u_j(t)\, Y_j[x(t)], \\ &x(0) = x_0, \\ &y(t, x_0) = r[x(t)], \end{aligned}$$

where

1) $X, Y, \cdots, Y_n$ are C^∞ or C^ω complete vector fields on M,
2) the outputs $u(t) = [u_1(t), \cdots, u_m(t)]$ are assumed to be piecewise constant,
3) the output function $r: M \to R^c$ is C^∞ or C^ω.

In the real analytic case, the output $y(t, x_0)$ of such a system can be represented, at least for small enough inputs, as a series expansion $y(t, x_0) = \sum_{n=0}^{\infty} W_n(u)[t, x_0]$ called the Volterra series (see [B], [C], [K-L]).

In the infinitely differentiable case the functionals $W_n(u)$ can still be defined and can be used to give partial expansions representing $y(t, x_0)$ similar to the Taylor formula.

* Received by the editors July 5, 1983, and in revised form December 15, 1983. This work was supported in part by A.T.P. under grant 040 228-18.

† Mathématiques Pures, Institute Fourier, BP74-38402 St. Martin d'Hères, France.

Each $W_n(u)$ is a functional of the input u that can be represented as an iterated integral:

$$W_n(u)[t, x_0] = \sum_{\substack{1<\alpha_1,\alpha_2,\cdots,\alpha_n<m \\ \alpha_1,\alpha_2,\cdots,\alpha_n \text{ integers}}} \int_0^t \int_0^{t_1} \cdots \int_0^{t_{n-1}} u_{\alpha_1}(t_1) \cdots u_{\alpha_n}(t) W_n^{\alpha}(t, t_1, \cdots, t_n, x_0)\, dt_n \cdots .$$

Each W^{α} is a C^{∞} or C^{ω} function $R^{n+1} \times M \to R^c$ having the following expression:

$$W_n^{\alpha}(t_0, \cdots, t_n, x) = [\theta(Z_{\alpha_n}(t_n)) \cdots \theta(Z_{\alpha_1}(t_1))r](e^{t_0 X}(x)),$$

where

1) e^{tX} denotes the flow generated by X and e^{-tadX} the adjoint action of this flow on vector-fields,

2) $Z_j(t) = e^{-tadX}(Y_j)$, $1 < j < m$,

3) $\theta(V)$ denotes the derivation operation associated to the vector field V.

DEFINITION. For any $x \in M$, we call $w(x)$ the smallest integer q such that $W_n^{\alpha}(t_0, \cdots, t_n, x) = 0$ for all $n > q$, all $\alpha = (\alpha_1, \cdots, \alpha_n)$ and all $(t_0, \cdots, t_n) \in R^{n+1}$ if such an integer exists and $+\infty$ otherwise.

Notation. L will denote the Lie algebra of vector fields on M generated by $\{X, Y_1, \cdots, Y_m\}$. S will denote the ideal in L generated by $\{Y_1, \cdots, Y_n\}$.

Basic assumptions.

(i) H1(x) will denote the statement

"at $x \in M$, $w(x) < +\infty$".

(ii) H2(x) will denote the statement

"at $x \in M$, $S(x) = T_x M$".

Now we are ready to state our main results.

THEOREM 1. *Under either of the two following assumptions*:

1) *we are in the real analytic case,* H1(x_0) *is satisfied at one point* $x_0 \in M$ *and* H2(x) *is satisfied at all* $x \in M$;

2) *we are in the infinitely differentiable case and for all* $x \in M$, H1(x) *and* H2(x) *hold*;

there exist:

α) *a finite-dimensional vector space* H,

β) *a Lie algebra representation* $\rho: L \to \mathrm{End}(H)$,

γ) *a linear mapping* $1: H \to R^c$,

δ) *a* C^{ω} *(resp.* C^{∞}*) mapping* $\Psi: M \to H$

having the following properties:

(i) *All elements from* $\rho(S)$ *are nilpotent.*

(ii) $\rho(L)$ *is a solvable subalgebra of* End (H) *whose nilpotent radical is* $\rho(S)$.

(iii) $\Psi(M)$ *is an orbit of the unipotent subgroup of* Aut (H) *generated by* $\rho(S)$ *and the mapping* $\Psi: M \to \Psi(M)$ *is a locally trivial fibration.*

(iv) *Any* $Z \in L$ *is* Ψ*-projectable and* $\Psi(Z)$ *is the linear field induced by* $\rho(Z)$ *(this means that for all* $m \in M$, $d\Psi(m)Z(m) = \rho(Z)\Psi(m)$).

(v) $l \circ \Psi = r$.

(vi) $\Psi(M)$ *endowed with the input-output system*

$$\text{(II)} \qquad \begin{aligned} \frac{dx'}{dt} &= \rho(X)x' + \sum_{j=1}^{m} u_j(t)\rho(Y_j)x', \\ y'(t) &= l[x'(t)] \end{aligned}$$

is a minimal realization of system I.

As an immediate corollary of Theorem 1 we get the following.

COROLLARY 1. *For any* $v \in \Psi(M)$ *there exists a sequence* $\{Z_1, \cdots, Z_e\}$ *of elements of* S *such that, if* Φ_v *denotes the mapping*

$$R^e \to H, (\lambda_1, \cdots, \lambda_e) = e^{\lambda_1 \rho(Z)} \cdots e^{\lambda_e \rho(Z_e)}(v),$$

then:

1) Φ *is a proper embedding having* $\Psi(M)$ *as image.*

2) *There exist a linear coordinate system* $(h_1, \cdots, h_\sigma)$ *on* H *and a sequence of integers* $j(0) = 0 < j(1) < j(2) < \cdots < j(e) < \sigma = \dim H$ *such that*:

(i) $h_j \circ \Phi(\lambda_1, \cdots, \lambda_e)$ *is a polynomial in* $\lambda_1, \cdots, \lambda_s$ *if* $j(s) < j < j(s+1)$.

(ii) $h_j \circ \Phi(\lambda) - \lambda_s$ *is a polynomial in* $\lambda_1, \cdots, \lambda_{s-1}$ *if* $j = j(s)$.

(iii) *In the coordinates* $(\lambda_1, \cdots, \lambda_e)$, *system* II *has the following "triangular" form*:

1) *For each* k, $1 < k < m$, *the field induced on* $\Psi(M)$ *by* $\rho(Y_k)$ *is of the form* $\sum_{s=1}^{e} P_{ks}\, \partial/\partial\lambda_s$, *where* P_{ks} *is a polynomial in* $(\lambda_1, \cdots, \lambda_{s-1})$.

2) *The field induced on* $\Psi(M)$ *by* $\rho(X)$ *is of the form* $\sum_{s=1}^{e} (A_s + B_s)\partial/\partial\lambda_s$ *where*

a) B_s *is a polynomial in* $(\lambda_1, \cdots, \lambda_{s-1})$;

b) *there exists a sequence* $1 < s(1) < s(2) \cdots < s(p) < e$ *with* $s(t) - s(t-1) > 2$ *for all* $t > 2$, *such that*:

* *if* s *and* $s-1$ *do not belong to* $\{s(1), \cdots, s(p)\}$ $A_s(\lambda) = \alpha_s \lambda_s$ *with a scalar* α_s,

* *if* s *belongs to* $\{s(1), \cdots, s(p)\}$

$$\begin{aligned} A_s(\lambda) &= \alpha_s \lambda_s + \beta_s \lambda_{s+1}, \\ A_{s+1}(\lambda) &= -\beta_s \lambda_s + \alpha_s \lambda_{s+1}. \end{aligned}$$

The scheme of the proof is as follows: in § 2 we prove some auxiliary lemmas whose purpose is to show that a certain vector space (i.e. *Ur*) that plays a crucial role in our construction is finite-dimensional. In § 3 we set up the basic constructions and prove Theorem 1. In § 4 we deduce Corollary 1 from Theorem 1. This proof is independent of § 2 and § 3. Finally in § 5 we give a simple example.

2. Some auxiliary lemmas. First we introduce a list of notation used in the rest of the paper:

(i) *Notation.*

$M :=$ C^∞ or C^ω (real analytic) connected manifold.

$TM :=$ its tangent space.

$T \times M :=$ its tangent space at $x \in M$.

$C^\infty(M; E)$ (resp. $C^\omega(M; E)$) $:=$ vector space of all C^∞ (resp. C^ω) mappings from M into the real finite-dimensional vector space E.

$C^\infty_x(M; E)$ (resp. $C^\omega_x(M; E)$) $:=$ space of all germs at x of C^∞ (resp. C^ω) mappings from a neighbourhood of x into E.

$M_x :=$ maximal ideal of all germs zero at x.

$J^\infty(M: E) := \infty$-jet bundle of all jets of C^∞ or C^ω mappings from open sets of M into E.

$J^\infty_x(M; E) :=$ its fiber at x.

$VF^\infty(M)$ (resp. $VF^\omega(M)$) := Lie algebra of all C^∞ (resp. C^ω) vector fields on M.
$VF^\infty_x(M)$ (resp. $VF^\omega_x(M)$) := Lie algebra of germs at x of C^∞ (resp. C^ω) vector fields.
e^Z of $\exp Z$:= exponential mapping of a complete vector field Z.
$\theta(Z)$:= Lie derivative induced by the vector field Z either on $C^\infty(M; E)$ or $C^\omega(M; E)$, or $C^\infty_x(M; E)$ or $C^\omega_x(M; E)$ or $J^\infty_x(M; E)$ according to the case.
U := associative R-algebra of differential operators on M generated by $\theta(L)$.
A := associative subalgebra of U generated by $\theta(S)$
U is generated as a vector space by the monomials $\{\theta(V_1)\cdots\theta(V_k)\theta(X)^n | V_1, \cdots, V_k \in S\}$ not necessarily distinct, n integer $>0\}$.
Finally $J^\infty(M; E)$ is a differentiable vector bundle on M if we endow the fibers with the topology of the convergence of coefficients.

(ii) *To begin with, let us state the following trivial but useful remarks.*

LEMMA 1. *Let $\Gamma \subset VF_x(M)$ be a subset such that $\Gamma(x) = T_xM$. For any $f \in C^\infty_x(M, R^c)$ (resp. $C^\omega_x(M; R^c)$) such that $f \in M^p_x - M^{p+1}_x$ for some integer p, there is a $V \in \Gamma$ such that*

$$[\theta(V)^p f](x) \neq 0.$$

COROLLARY 2. *Given an integer $q \geqq 0$, call E_q the subspace of all $f \in C^\infty_x(M; R^c)$ (resp. $C^\omega_x(M; R^c)$) such that for any sequence $V_1, \cdots, V_n$ of elements of Γ, with $n \geqq q+1$, $[\theta(V_1)\theta(V_2)\cdots\theta(V_n)f](x) = 0$. Then the canonical projection $\pi_q: J^\infty_x(M, E) \to J_x(M, E)$ is injective on $j^\infty_x E$. In particular* $\dim j^\infty_x E_q < \dim J^q_x(M, R^c) = c\binom{d+q}{q}$, $d = \dim M$.

Proof of Corollary 2. Assume there is an $f \in E_q$ such that $j^\infty_x f \in \operatorname{Ker} \pi_q$ and $j^\infty_x f = 0$. Then there is an integer $p > q+1$ such that $f \in M^p_x - M^{p+1}_x$. By Lemma 1 there is a $V \in \Gamma$ such that $[\theta(V)^p f](x) = 0$. This is a contradiction.

LEMMA 2. *Assume* H1(x) *and* H2(x) *hold at an $x \in M$. Then*

1) *For any integer $k \geqq 0$, A_{q+1} ideal of A generated by $\theta(V_1)\cdots\theta(V_k)$, $k \geqq q+1$ $V_1, \cdots, V_k \in S$.*

$$[A_{q+1}\theta(X)^k r](x) = 0 \quad \text{with } q = w(x) - 1.$$

2) $\dim j^\infty_x(Ur)$ *is finite and* $\leqq c\binom{d+q}{q}$.

3) *In the real analytic case,* $\dim Ur$ *is finite.*

Proof of Lemma 2.

1) Take any $n > q$, any $\alpha = (\alpha_1, \cdots, \alpha_n)$ $\alpha_j \in \{1, 2, \cdots, m\}$ and any $\beta = (\beta_0, \cdots, \beta_n) \in N^{n+1}$. Since $W^\alpha_n(t_0, \cdots, t_n, x) = 0$ for all $(t_0, \cdots, t_n) \in R^{n+1}$ it follows that

$$\frac{\partial^{|\beta|} W^\alpha_n}{\partial t_0^{\beta_0} \cdots \partial t_n^{\beta_n}}(0 \cdot x) = 0.$$

But this derivative is up to sign $[\theta(ad^{\beta_1}X(Y_{\alpha_1}))\cdots(ad^{\beta_n}X(Y_{\alpha_n}))\theta(X)^{\beta_0} r](x)$.

2) This follows from 1) and Corollary 2 applied to $\Gamma = S$.

3) This follows from the fact that j^∞_x: $C^\omega(M, R^c) \to J^\infty_x(M, R^c)$ is injective.

The next lemma proves property 3 of Lemma 2 under stronger assumptions. It is contained in the results proved in [F-K]. For the convenience of the reader we prove it here.

LEMMA 3. *Let E be a vector subspace of $C^\infty(M; R^c)$ and Γ a subset of $VF^\infty(M)$, having the following properties*:

1) *for all $x \in M$. $\Gamma(x) = T_xM$,*

2) *for all $V \in \Gamma$, $\theta(V)E \subset E$,*

3) *for all $x \in M$, $\dim j^\infty_x E < -\infty$.*

Then $\dim E < +\infty$ *and for any $x \in M$, j^∞_x: $E \to J^\infty_x(M; R^c)$ is injective.*

Proof of Lemma 3. Call $\nu: M \to N$ the function $\nu(x) = \dim j_x^\infty E$. It is obvious that ν is lower semicontinuous. We shall consider a special case first.

Assume ν is constant on M. Then $\hat{E} = \bigcup_{x \in M} j_x^\infty E$ has a natural structure of a subvector bundle of $J^\infty(M: R^c)$. $J^\infty(M; R^c)$ is endowed with a natural connection (see [M, p. 504]) whose associated derivation we denote by D. Since $Dj^\infty = 0$, E is stable under this connection and the induced connection on E is flat. Hence the space of all global horizontal sections has at most dimension ν. Since j^∞ maps E into the horizontal sections of $\hat{E}$, $\dim E \leqq \nu$. Since $\dim E \geqq \dim j_x^\infty E$ for any $x \in M$, $\dim E = \nu$ and j_x^∞ is injective on E.

Let us consider the general case now. Baire's theorem shows that the set $\mathcal{O}$ of all $x \in M$ at which ν is locally constant is an open dense subset of M. Denote the connected components of $\mathcal{O}$ by $\mathcal{O}_i$, $i \in I$. Let E_i be the restriction of E to $\mathcal{O}_i$. On each $\mathcal{O}_i$, ν is constant and we can apply our first result to E_i: $\dim E_i < +\infty$ and for each $x \in \mathcal{O}_i$, j_x^∞ is injective on E_i. It is sufficient to prove that $\mathcal{O} = M$.

We are going to show the following property: let $V \in \Gamma$ and let $y \in \mathcal{O}_i$. Assume that for all t, $0 < t < a$, $e^{tV}(y) \in \mathcal{O}_i$ but $x = e^{aV}(y) \notin \mathcal{O}_i$. Then $\dim j_x^\infty E = \dim j_y^\infty E$ and j_x^∞ is injective on E. Assuming this property it is easy to finish the proof: by Baire's theorem again, the set ω of all $x \in M - \mathcal{O}$ such that the restriction of ν to $M - \mathcal{O}$ is locally constant at x (on $M - \mathcal{O}$!) is open dense in $M - \mathcal{O}$. Take any $x \in \omega$. On can choose elements $V_1, \cdots, V_d \in \Gamma$ and a neighborhood Λ of x such that:

(i) ν is constant on $\Lambda \cap (M - \mathcal{O})$,

(ii) $V_1(x), \cdots, V_d(x)$ form a basis of $T_x M$,

(iii) The mapping $B^d = \{\lambda \in R^d \lambda_1^2 + \cdots + \lambda_d^2 < 1 \to \exp(\lambda_1 V_1 + \cdots + \lambda_d V_d)(x)$ is a diffeomorphism onto Λ.

For any $\lambda \in B^d$ call γ_λ the arc $\{\exp[t(\lambda_1 V_1 + \cdots + \lambda_d V_d)](x) | 0 < t < 1\}$. Since $\gamma_\lambda \subset \Lambda$, ν is equal to $\nu(x)$ on $\gamma_\lambda \cap M - \mathcal{O}$. $\gamma_\lambda \cap \mathcal{O}$ is a disjoint union of open arcs $\{\gamma^k | k \in N\}$. By the property above, ν is constant on the closure of each γ^k in γ. Hence ν is constant and equal to $\nu(x)$ on γ_λ. This shows that ν is constant and equal to $\nu(x)$ on Λ. $x \in \mathcal{O} \cap (M - \mathcal{O}) = 0$. This is a contradiction.

It remains to prove the above assertion about the flows of the V's in Γ. Since $\theta(V) E_i \subset E_i$, and E_i is finite dimensional. $\theta(V)$ induces a linear transformation A of E_i. e^{tV} induces a continuous family of isomorphisms $\phi_t: J_{x_t}(M; R^c) \to J_y(M; R^c)$, $x_t = e^{tV}(x)$. It is easy to see that for all $t \in [0, a[$ $\phi_t(j_{x_t}^\infty E) \subset j_y^\infty E$ and for all $f \in E$, $\phi_t(j_{x_t}^\infty f) = j_y^\infty(e^{tA} f_i)$, $f_i = f/\mathcal{O}_i$. By continuity it follows that for all $f \in E$, $\phi_a(j_x^\infty f) = j_y^\infty(e^{tA} f_i)$. Hence $\dim j_x^\infty E = \dim \phi_a(j_x^\infty E) = \dim j_y^\infty E$.

COROLLARY 3. *If* H1(x) *and* H2(x) *for all* $x \in M$, *Ur is finite dimensional, and w is constant equal to q say, and* $A_q(Ur) = \{0\}$.

Proof of Corollary 3. Lemma 3, applied to $\Gamma = S$ and $E = Ur$, shows that Ur is finite dimensional and for all $x \in M$. j_x^∞ is injective on Ur. Let $P \in A_{w(x)}$. For any $f \in Ur$, $j_x^\infty(Pf) = P(j_x^\infty f) = 0$. Hence $Pf = 0$. This proves the corollary.

3. The basic construction and the proof of Theorem 1. In this part we assume that either of the assumptions of Theorem 1 is satisfied. By Lemma 2 and Corollary 2, Ur is of finite dimension. The Lie derivative operator θ induces a Lie algebra representation, $\theta_L: L \to \mathrm{End}(Ur)$. Let us call H the space $\mathrm{Lin}(Ur, R^c)$ of all linear mappings $Ur \to R^c$. θ_L induces a contragradient representation $\rho: L \to \mathrm{End}(H)$ since H is just $\mathrm{Dual}(Ur) \otimes R^c$. M can be mapped into H in a natural fashion as follows: for any $x \in M$ denote by $\Psi(x)$ the linear mapping $Ur \to R^c$ defined as follows: $\Psi(x)[f] = f(x)$ if $f \in Ur$. Finally, let $l; H \to R^c$ be the linear map: $l(\alpha) = \alpha(r)$. We have the following lemma.

LEMMA 4.

(i) *Any $V \in L$ is Ψ projectable and $\Psi(V)$ is the linear field generated by $\rho(V)$.*

(ii) *All elements from $\rho(S)$ are nilpotent.*

(iii) *$\rho(L)$ is solvable and $\rho(S)$ is its nilpotent radical.*

(iv) *If N denotes the unipotent subgroup of Aut (H) generated by $\rho(S)$, $\Psi(M)$ is an orbit of N.*

(v) $\Psi^*(l) = r$.

Proof of Lemma 4.

(i) Let $V \in L$. For any $x \in M$ and any $f \in Ur$,

$$(\rho(V)\Psi(x))[f] = \Psi(x)[\theta(V)f] = (\theta(V)f)(x),$$

$$[d\Psi(x) \cdot V(x)][f] = \lim_{t \to 0} \frac{\Psi[e^{tV}(x)] - \Psi(x)}{t}[f],$$

$$[d\Psi(x) \cdot V(x)][f] = \lim_{t \to 0} \frac{f[e^{tV}(x)] - f(x)}{t} = df(x) \cdot V(x)$$
$$= [\theta(V)f](x).$$

(ii) By Corollary 1 $\theta_L(V)^q = 0$ on Ur, for all $V \in S$, and hence $\rho(V)^q = 0$.

(iii) Follows from (ii).

(iv) N is just the set $\{e^{\rho(V_1)} \cdots e^{\rho(V_n)} | V_1, \cdots, V_n \in S\}$.

(v) Since S is transitive on M, N is transitive on $\Psi(M)$, and (v) is trivial.

Lemma 4 covers the claims (i), (ii), (iv) and (v) of Theorem 1. To complete the proof of part (iii) one has to show that Ψ is a locally trivial fibration. This we leave to the reader. It remains to prove that II is minimal.

Assume it is not. Then it is easy to see that there exist two points $\xi, \eta \in \Psi(M)$ such that $l(g\, e^{tX}\xi) = l(g\, e^{tX}\eta)$ for all g in an open subset of N and all t, $|t| < \varepsilon$. Since N is connected and l linear, $l(g\, e^{tX}\xi) = l(g\, e^{tX}\eta)$ for all $g \in N$ and all $t \in R$. Deriving this relation, we get $\xi(f) = \eta(f)$ for all $f \in Ur$. Hence $\xi = \eta$.

4. Proof of Corollary 1. The proof is similar to the one in [P] p with an additional twist. We shall sketch it briefly, stressing only the points which are different.

Let $v \in \Psi(M)$. By adding to X an appropriate $Z \in S$, we can assume that $\rho(X)v = 0$. Since $\rho(L)$ is solvable, by Lie's theorem there exists a flag of subspaces $H = H'_0 \supset H'_1 \supset \cdots \supset H'_a = \{0\}$ such that:

(i) $\rho(L)H'_k \subset H'_k$ for all k,

(ii) $\dim (H'_{k-1}/H'_k)$ is 1 or 2 and if $\dim (H'_{k-1}/H'_k) = 2$, the spectrum of $\rho(x)$ in the quotient H'_{k-1}/H'_k is complex.

Using Engel's theorem, this flag can be refined in a flag $H = H_0 \supset H_1 \supset \cdots \supset H_\sigma = \{0\}$ such that:

(i) $\rho(S)H_k \subset H_{k+1}$ for all k.

Denote by $0 < k(1) < k(2) < \cdots < k(p) < \sigma$ those k for which H_k is not $\rho(X)$-invariant. For simplicity we call the set $\{k(1), \cdots, k(p)\}$, K. Then H_{k-1} and H_{k+1} are $\rho(X)$-invariant and the spectrum of $\rho(X)$ in H_{k-1}/H_{k+1} is complex.

Let ϕ: $S \to H$ be the evaluation mapping: $\phi(V) = \rho(V)v$. Set $S_k = \phi^{-1}(H_k)$ for all k. Then $S = S_0 \supset S_1 \supset \cdots \supset S_\sigma$ and since H_k is $\rho(S)$ stable, the S_k are subalgebras of S. It also follows that $\dim S_{k-1}/S_k \leqq 1$.

Since $\rho(X)v = 0$, $\phi \circ (adX) = \rho(X) \circ \phi$. This implies that if $k \notin \{k(1), \cdots, k(p)\}$, S_k is adX-stable. Moreover if $k \in \{k(1), \cdots, k(p)\}$ and if $S_{k-1}/S_{k+1} \neq \{0\}$, then $\dim S_{k-1}/S_{k+1} = 2$ and the spectrum of adX on S_{k-1}/S_{k+1} is complex since the adX module S_{k-1}/S_{k+1} is isomorphic via ϕ to the $\rho(X)$-module H_{k-1}/H_{k+1}.

Now it is possible to construct a basis $(\varepsilon_1, \cdots, \varepsilon_\sigma)$ of H and a family of vectors $\{Z_{(k)} | 1 < k < \sigma, S_{k-1} \neq S_k\}$ of S satisfying the following conditions:

a) For any $k \geqq 1$, $(\varepsilon_k, \cdots, \varepsilon_\sigma)$ is a basis of H_{k-1}.

b) If $k \in \{k(1), \cdots, k(p)\}$, $\varepsilon_{k-1} + \sqrt{-1}\, \varepsilon_k$ is an eigenvector of adX modulo H_{k+1}.

c) If $\dim S_{k-1}/S_k = 1$ and $k \notin \{k(1), \cdots, k(p)\}$, $\rho[Z_{(k)}]v = \varepsilon_k$ modulo H_k.

d) If $k \in \{k(1), \cdots, k(p)\}$ and $S_{k-1} \neq S_{k+1}$ $\rho[Z_{(k)}]v = \varepsilon_k$ modulo H_{k+1} and $\rho[Z_{(k+1)}]v = \varepsilon_{k+1}$ modulo H_{k+1} and $Z_{(k)} \in S_{k-1}$, $Z_{(k+1)} \in S_k$. It is clear that in case d), $Z_{(k)} + \sqrt{-1}\, Z_{(k+1)}$ is an eigenvector for adX modulo S_{k+1}.

Now denote by $(h_1, \cdots, h_\sigma)$ the basis dual to $(\varepsilon_1, \cdots, \varepsilon_\sigma)$. Reorder the $\{Z_{(k)} | S_{k-1} \neq S_k\}$ in a sequence $(Z_1, \cdots, Z_e)$ in such a way that, if $Z_n \in S_k$ then $Z_t \in S_k$ for all $t \geqq n$. I claim that the vectors $(Z_1, \cdots, Z_e)$ and the linear forms $(h_1, \cdots, h_\sigma)$ satisfy the condition of Corollary 1.

In fact, as in Pukansky [P, p. 50-64], it can be shown that conditions (i) and (ii) or Corollary 1 hold. The j_s are those j such that $S_{j-1} \neq S_j$. As for condition (iii), let us denote for simplicity $h_j \circ \Phi$ by Φ_j. Let $Z \in S$. Then by the choice of the ε, $\rho(Z)\varepsilon_k = \sum_{j=k+1}^{\sigma} b_{jk}\varepsilon_j$ where the b_{jk} are scalars. If the field induced on $\Psi(M)$ by $\rho(Z)$ is $\sum_{k=1}^{e} P_k(\lambda)\, \partial/\partial\lambda_k$, then

$$\sum_{k=1}^{e} \frac{\partial \Phi}{\partial \lambda_k}(\lambda) P_k(\lambda) = \rho(Z)\Phi(\lambda).$$

Componentwise;

$$\sum_{k=1}^{e} \frac{\partial \Phi_i}{\partial \lambda_k}(\lambda) P_k(\lambda) = \sum_{j=1}^{i-1} b_{ij}\Phi_j(\lambda).$$

Choosing $i = j_s$, we get

$$P_s(\lambda) + \sum_{k=1}^{s-1} \frac{\partial \Phi_{j_s}}{\partial \lambda_k}(\lambda) P_k(\lambda) = \sum_{j=1}^{j_s-1} a_{j_s j}\Phi_j(\lambda).$$

By induction on s we see that P_{s-1} is a polynomial in $(\lambda_1, \cdots, \lambda_s)$ since $(\partial\Phi_{j_0}/\partial\lambda_k)(\lambda)$ depends only on $(\lambda_1, \cdots, \lambda_{s-1})$.

As for X, $\rho(X) = \rho(X)_1 + \rho(X)_2$ where for all k, $1 < k < \sigma$,

$$\rho(X)_2 \varepsilon_k = \sum_{j=k+1}^{e} a_{jk}\varepsilon_j$$

a_{jk} are scalars, and $\rho(X)_1 \varepsilon_k = a_k \varepsilon_k$ if $k \neq k(i)$ or $k \neq k(i) - 1$ for some $k(i)$, $1 \leqq i \leqq p$ and

$$\rho(X)_1 \varepsilon_{k-1} = a_k \varepsilon_{k-1} - b_k \varepsilon_k,$$
$$\rho(X)_1 \varepsilon_k = b_k \varepsilon_{k-1} + a_k \varepsilon_k,$$

for some scalars a_k, b_k with $b_k \neq 0$, if $k \in \{k(1), \cdots, k(p)\}$. Let $\sum_{s=1}^{e} R_s(\lambda)\, \partial/\partial\lambda_s$ and $\sum_{s=1}^{e} Q_s(\lambda)\, \partial/\partial\lambda_s$ be the fields induced on $\Psi(M)$ by $\rho(X)_1$ and $\rho(X)_2$ respectively. A computation similar to the one above shows that R_s is a polynomial in $(\lambda_1, \cdots, \lambda_{s-1})$. If $j(s)$ and $j(s+1)$ do not belong to $\{k(1), \cdots, k(p)\}$, $Q_s = a_{j(s)}\lambda_s + \text{polynomial in } (\lambda_1, \cdots, \lambda_{s-1})$. If $j(s)$ belongs to $\{k(1), \cdots, k(p)\}$ then $S_{j(s)} \neq S_{j(s)+1}$ and $S_{j(s)-1} \neq S_{j(s)}$. Hence $j(s+1) = j(s) + 1$ and

$$Q_s = a_{j(s)}\lambda_s + b_{j(s)}\lambda_{s+1} + \text{polynomial in } \lambda_1, \cdots, \lambda_{s-1},$$
$$Q_{s+1} = -b_{j(s)}\lambda_s + a_{j(s)}\lambda_{s+1} - \text{polynomial in } \lambda_1, \cdots, \lambda_{s-1}.$$

5. An example. Let

$$M = R^3, \quad m = 2, \quad c = 1, \quad Y_1 = \frac{\partial}{\partial x_1}, \quad Y_2 = \frac{\partial}{\partial x_2} + x_i \frac{\partial}{\partial x_2},$$

$$X = (\alpha x_1 + \beta x_2)\, \partial/\partial x_1 + (\gamma x_1 - \alpha x_2)\, \partial/\partial x_2 + \tfrac{1}{2}(\gamma x_1^2 + \beta x_2^2)\, \partial/\partial x_3$$

where α, β, γ are constants.

$r = x_1$.

Then $L = RX + RY_1 + RY_2 + R\,\partial/\partial x_3$, $S = RY_1 + RY_2 + R\,\partial/\partial x_3$. Ur is the space of all affine functions in the variables x_1 and x_2. H is the dual of Ur. Denote by 1, $\hat{x}_1$, $\hat{x}_2$ the basis of H dual to the basis 1, x_1, x_2 of Ur. On this basis,

$$\rho(X) = \begin{vmatrix} 0 & 0 & 0 \\ 0 & \alpha & \beta \\ 0 & \gamma & \alpha \end{vmatrix}, \quad \rho(Y_1) = \begin{vmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{vmatrix}, \quad \rho(Y_2) = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{vmatrix},$$

$\rho[\partial/\partial x_3] = 0$. It is clear that the subspace H' of H generated by $\hat{x}_1$ and $\hat{x}_2$ is $\rho(L)$-invariant.

The mapping $\Psi\colon M \to H$ is defined as follows:

$$\Psi(\xi_1, \xi_2, \xi_3) = \hat{1} + \xi_1 \hat{x}_1 + \xi_2 \hat{x}_2.$$

Hence $e = \dim \Psi(M) = 2$. We choose 0 as the point v. $\Psi(0) = \hat{1}$. In order to pursue our discussion we need to distinguish two cases.

Case 1. $\rho(X)$ has a real spectrum. Then the space $RY_1 + RY_2$ contains an eigenvector Z_1 of adX and we can choose a vector $Z_2 \in (RY_1 + RY_2)$, linearly independent from Z_2 which is an eigenvector of adX modulo RZ_2. We define the basis $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ of H as follows:

$$\varepsilon_1 = 1, \quad \varepsilon_2 = \rho(Z_1)(\hat{1}), \quad \varepsilon_3 = \rho(Z_2)(\hat{1}).$$

If we parametrize $\Psi(M)$ by $\lambda = (\lambda_1, \lambda_2) \in R^2 \to \exp(\lambda_1 Z_1) \exp(\lambda_2 Z_2)(\hat{1})$, the minimal system can be expressed as follows: A, B, C, D are constants,

$$\frac{d\lambda_1}{dt} = a\lambda_1 - Au_1 - Bu_2, \qquad \frac{d\lambda_1}{dt} = \lambda_2 + Au_1 + Bu_2,$$

$$\frac{d\lambda_2}{dt} = -a\lambda_2 + Cu_1 + Du_2, \qquad \frac{d\lambda_2}{dt} = Cu_1 + Du_2,$$

$$\text{if } \alpha^2 - \beta\gamma \neq 0; \qquad \text{if } \alpha^2 + \beta\gamma = 0;$$

a an eigenvalue of $\left|\begin{matrix} \alpha & \beta \\ \gamma & -\alpha \end{matrix}\right|$.

Case 2. $\rho(\lambda)$ has a complex spectrum. The space $CY_1 + CY_2$ contains an eigenvector $Z_1 + (\sqrt{-1})Z_2$ where $Z_1, Z_2 \in (RY_1 + RY_2)$. We define the basis $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ of H as follows: $\varepsilon_1 = 1$, $\varepsilon_2 = \rho(Z_1)(\hat{1})$, $\varepsilon_3 = \rho(Z_2)(\hat{1})$. If we parametrize $\Psi(M)$ by $\lambda(\lambda_1, \lambda_2) \to \exp(\lambda_1 Z_1) \cdot \exp(\lambda_2 Z_2)(\hat{1})$ the minimal system can be expressed as follows: A, B, C, D constants

$$\frac{d\lambda_1}{dt} = a\lambda_1 - b\lambda_2 + Au_1 + Bu_2, \qquad a + \sqrt{-1}\, b \text{ an eigenvalue of } Z_1 + (\sqrt{-1})Z_2,$$

$$\frac{d\lambda_2}{dt} = b\lambda_1 + a\lambda_1 + Cu_1 + Du_2.$$

REFERENCES

[B] R. W. BROCKETT, *Volterra series and geometric control theory*, Automatica, 12 (1976), pp. 167-176.

[C] P. CROUCH, *Dynamical realisations of finite Volterra series* this Journal, 19 (1981), pp. 177-202.

[G] E. G. GILBERT, *Functional expansions for nonlinear differential systems*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 900-921.

[F-K] M. FLIESS AND I. KUPKA, *A finiteness criterion for nonlinear input-output differential systems*, this Journal, 21 (1983), pp. 721-728.

[K-L] A. J. KRENER AND C. M. LESIAK, *The existence and uniqueness of Volterra series for nonlinear systems*, IEEE Trans. Automat. Control, AC-23 (1978), pp. 1090-1095.

[M] B. MALGRANGE, *Lie equations* I *and* II J. Differential Geometry, 6 (1971-1972), pp. 501-522.

[P] L. PUKANSZKY, *Leçons sur les représentations des groupes*, Monographe de la Soc. Math. de France no. 2, Dunod, Paris, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 2, March 1985

009

STOCHASTIC CONTROL AND EXIT PROBABILITIES OF JUMP PROCESSES*

SHUENN-JYI SHEU†

Abstract. A stochastic control problem is formulated for some problems related to Markov process. This formulation is in some sense a generalization of one used in [2], [3], [4], [8] for diffusion case. We apply this to study the asymptotic behavior of exit probabilities of a family of jump processes depending on a small parameter ε as $\varepsilon \to 0$.

Key words. jump process, exist probability, stochastic control, large deviation

Introduction. In [2], [3], [4], [8], some problems related to exit probabilites for diffusions were considered. They found certain control problems were naturally associated to these problems. This observation is basic for their solutions to the original problems. We may conjecture the similar phenomenon exists for some other problems. Indeed, in this paper we will show it is possible to formulate a stochastic control problem for some problems. This observation is applied to provide another approach to the problem considered by Ventsel' in [7] about large deviations concerning the exit probability of jump processes depending on small parameter $\varepsilon > 0$. This is the main body of the paper.

In § 1, we present briefly the properties of jump processes.

In § 2, some preliminaries about exit probabilities for jump processes are given.

In § 3, a stochastic control problem and a determinisitc control problem are formulated. Some basic analysis for these two control problems is also given in this section.

In § 4, we give the assumptions and main results. The proof of these results will follow the same ideas as in the aforementioned work. However, here we need to formulate the control problems mentioned above. In the rest of this introduction, we will give the basic idea for this.

The main idea is as follows. We have a certain family of functions $\phi^\varepsilon(t, x)$, $t \in R^+$, $x \in E$ a certain metric space, $\varepsilon > 0$. For each fixed $\varepsilon > 0$, $\phi(t, x) = \phi^\varepsilon(t, x)$ satisfies an equation

$$\frac{d\phi}{dt}(t, x) = L_\varepsilon \phi = L\phi_t(x), \qquad t > 0 \tag{1}$$

where L_ε is an operator on a certain function space B on E. Usually B is a dense subspace of $C(E)$-collection of bounded continuous functions on E. Now we assume

$$\phi > 0 \quad \text{on } R^+ \times E. \tag{2}$$

The first step is to apply a logarithmic transform to ϕ and get ψ, i.e.

$$\psi = -\log \phi. \tag{3}$$

We have the following equation for ψ

$$\frac{d\psi}{dt} = -e^{\psi} L\, e^{-\psi}. \tag{4}$$

* Received by the editors January 31, 1983, and in revised form. This work was supported by the National Science Foundation under grant MCS-7903554 and the Air Force Office of Scientific Research under grant AFOSR-81-0116.

† Mathematics Department, Brown University, Providence, Rhode Island 02912. Current Address: Institute of Mathematics, Academia Sinica, Nankang, Taipei, Taiwan, Republic of China.

We hope that there are operators L^u on a certain subspace of $C(E)$ and function k^u on E depending on a certain parameter u, such that $-e^{\psi}L\,e^{-\psi}=\inf_u[L^u+k^u]$. Then for fixed $T>0$, $\tilde{\psi}_t=\psi_{T-t}$ satisfies the following equation

$$\frac{d\tilde{\psi}_t}{dt}+\inf_u[L^u\tilde{\psi}_t+k^u]=0. \tag{5}$$

We recognize this as the dynamic programming equation of a certain control problem. The concept of control method appears in this sense.

To simplify the argument we assume $L: C(E)\to C(E)$ is bounded. The following condition is crucial for the existence of L^u and k^u.

(6) L satisfies a positive maximum principle, i.e., if $f\in C(E)$, take a nonnegative maximal at x_0; then $Lf(x_0)\leqq 0$.

The following example will indicate how one chooses L^u and k^u.

Example. Consider $E=\{1,\cdots,N\}$, $C(E)$ is nothing but R^N. Let L be given by

$$(Ly)_i=\sum q_{ij}y_j,\qquad y=(y_1,\cdots,y_N),\qquad i=1,\cdots,N.$$

$$q_{ij}\geqq 0,\quad i\neq j,\qquad q_{ii}=-\sum_{i\neq j}q_{ij}.$$

L is just a generator of N-state continuous Markov chain

Let

$$v=(v_1,\cdots,v_N)\in R^N,\qquad e^v=(e^{v_1},\cdots,e^{v_N}),$$

$$H(v)=-e^vLe^{-v},\quad \text{i.e.,}$$

$$H(v)_i=-\sum q_{ij}\,e^{v_i-v_j}=-\left\{q_{ii}+\sum_{i\neq j}q_{ij}\,e^{v_i-v_j}\right\}.$$

We want $H(v)_i=\min[k_i^w+\sum_{j\neq i}q_{ij}^w(v_j-v_i)]$. We note, by duality for convex functions, that $e^r=\max_{w>0}[w-w\log w+wr]$, $-e^r=\inf_{w>0}[w\log w-w-wr]$.

If we let

$$\tilde{q}_{ij}^w=q_{ij}w_{ij},\qquad i\neq j,\qquad w=(w_{ij})_{i\neq j}\geqq 0,$$

$$\tilde{k}_i^w=-q_{ii}+\sum_{i\neq j}(w_{ij}\log w_{ij}-w_{ij})q_{ij}$$

$$=\sum_{i\neq j}(w_{ij}\log w_{ij}-w_{ij}+1)q_{ij},$$

then $H(v)_i=\min_w[\tilde{k}^w+\sum_{i\neq j}\tilde{q}_{ij}^w(v_j-v_i)]$ where the minimum is attained at $w^*=(w_{ij}^*)$, $w_{ij}^*=e^{v_i^*-v_j^*}=u_i^*/u_j^*$, $u_i^*=e^{v_i^*}$. Hence

$$H(v)_i=\min_u\left[k_i^u+\sum_{i\neq j}q_{ij}^u(v_j-v_i)\right]=\min_u[k_i^u+(L^uv)i],$$

$$u=(u_1,\cdots,u_N)\geqq 0,$$

$$q_{ij}^u=q_{ij}\frac{u_i}{u_j},$$

$$k_i^u=-q_{ii}+\sum_{i\neq j}\left(\frac{u_i}{u_j}\log\frac{u_i}{u_j}-\frac{u_i}{u_j}\right)q_{ij}=\sum_{i\neq j}\left(\frac{u_i}{u_j}\log\frac{u_i}{u_j}-\frac{u_i}{u_j}+1\right)q_{ij},$$

$$(L^uv)_i=\sum_{i\neq j}q_{ij}\frac{u_i}{u_j}(v_j-v_i).$$

We then observe that

$$(L^u v)_i = \sum_{i \neq j} q_{ij} \frac{u_i}{u_j}(v_j - v_i) = u_i \left(L\frac{v}{u}\right)_i - u_i v_i L\left(\frac{1}{u}\right)_i.$$

Make a change $u \to 1/u$. Then

$$(L^u v)_i = \frac{1}{u_i}(L(uv)_i - (vL(u))i),$$

$$k_i^u = L^u(\log u)_i - \left(\frac{L(u)}{u}\right)_i = L^u(\log u)_i - \left(\frac{Lu}{u}\right)_i.$$

For our operator L we can still define L^u and k^u for u being a positive continuous function analogously as follows

$$L^u f = \frac{1}{u}(L(uf) - fL(u)), \qquad k^u = L^u(\log u) - \frac{1}{u}L(u).$$

Then under condition (6) we have the following

LEMMA. $\inf_u [L^u f + k^u] = -e^f L e^{-f}$.

Proof. Fix x and consider the function

$$F(y) = \frac{1}{u(x)}((uf)(y) - f(x)u(y) + (u \log u)(y) - (\log u(x))u(y) - u(y))$$
$$+ \exp(f(x) - f(y))$$
$$= \frac{u(y)}{u(x)}(f(y) - f(x)) + \frac{u(y)}{u(x)} \log \frac{u(y)}{u(x)} - \frac{u(y)}{u(x)} + \exp(f(x) - f(y)).$$

We have $F \geqq 0$ and $F(x) = 0$, i.e x is the nonnegative maximal point of $-F$. Therefore $LF(x) \geqq 0$ by (6). This is equivalent to

$$L^u f(x) + k^u(x) \geqq -e^f L e^{-f}(x).$$

Now by taking $u(y) = \exp(-f(y))$, we get $F \equiv 0$, then $LF \equiv 0$. This again is equivalent to

$$L^u f + k^u = -e^f L e^{-f}.$$

Example. We consider the operator

$$Lf(x) = \tfrac{1}{2}a(x)f''(x) + b(x)f'(x) - c(x)f(x), \qquad c \geqq 0, \quad a > 0$$

which is the generator of a one-dimensional diffusion with killing rate c. $E = R$ in this case and L satisfies the condition (6).

$$L^u f(x) = \frac{1}{2}a(x)f''(x) + \left(b(x) + \frac{au'}{u}\right)f'(x),$$

$$k^u(x) = \frac{1}{2}a\left(\frac{u'}{u}\right)^2 + c.$$

Therefore L^u corresponds to a change in the drift. Usually it takes the following form, which corresponds to changing the drift

$$\tilde{L}^v f = \tfrac{1}{2}a(x)f''(x) + v(x)f'(x).$$

If $v = b + au'u^{-1}$, then $L^u = \tilde{L}^v$ and k^u takes the familiar form

$$k^u = \tilde{k}^v = \tfrac{1}{2}a(a^{-1}(v - b))^2 + c.$$

Remark. Although L in the previous example is not bounded as was required in the lemma, the argument can still be applied to L for $f \in C_b^2(R)$, the set of all functions f with f', f'' bounded and continuous.

Remark. The change of generator from L to L^u means to change the probability measure from P to p^u. We refer the reader to [10] for the explanation.

1. Preliminaries. In this section we will briefly discuss some results about jump processes. For more detail on the properties of jump processes we refer to the book [5].

Let $\pi(x, dy)$ be a kernel on R^n, i.e. for each $x \in R^n$, $\pi(x, \cdot)$ is a probability measure on R^n and for each Borel set $B \in R^n$, $\pi(\cdot, B)$ is Borel measurable. We assume also $\pi(x, \{0\}) = 0$. $a(x)$ is a bounded Borel function. Then we have the following theorem for existence and uniqueness of a jump process corresponding to (π, a).

THEOREM 1. *There is a unique homogeneous Markov process $(x(\cdot), F_t, P_x)$ with state space R^n such that*

(a) *$x(\cdot)$ is a jump process,* i.e. *$x(\cdot)$ are right continuous step functions for a.s.*

(b) *For bounded Borel function f on R^n, we define $T_t f(x) = E_x[f(x_t)] = \int f(y) P(t, x, dy)$. Then $T_t f$ is a continuous differentiable function of t and satisfies the following differential equation*

$$\frac{dT_t f(x)}{dt} = a(x) \int (T_t f(x+y) - T_t f(x)) \pi(x, dy). \tag{1.1}$$

Proof. For the uniqueness, we note that if we take $f = I_B$ the characteristic function of a Borel set B, then $T_t f(x) = P(t, x, B)$. Therefore the transition probability $P(t, x, B)$ satisfies equation (1.1) with initial condition $T_0 f(x) = I_B(x)$. It has a unique solution. This means $P(t, x, B)$ is unique (cf. [5, p. 27]).

For the existence, we need the following more general construction theory for jump processes. It will be useful in the following. We present it here briefly. For more detail we refer to [5, Chap. III].

Let $\pi(t, x, B)$ be a kernel with $t \in R^+$, $x \in R^n$, B Borel set on R^n such that for each fixed B it is a Borel function on (t, x) and for each (t, x) it is a probability measure on R^n. Assume $\pi(t, x, \{0\}) = 0$ for $t > 0$ and $a(t, x)$ is a bounded Borel function. Define $(\tau_n, X(\tau_n))_{n=1}^{\infty}$ to be a Markov chain with state space $R^+ \times R^n$ and transition probability $Q(z, \tilde{B})$, $z \in R^+ \times R^n$, $\tilde{B}$ Borel set on $R^+ \times R^n$. $Q(z, \tilde{B})$ is defined by (1.2) for $z = (s, x)$, $\tilde{B} = [\alpha, \beta) \times B$, $0 < s$, $\alpha < \beta$. We can extend this to general $\tilde{B}$ as usual.

$$Q(z, \tilde{B}) = \begin{cases} \int_\alpha^\beta \exp\left(-\int_s^t a(\theta, x)\, d\theta\right) a(t, x) \pi(t, x, B - x)\, dt & \text{if } s \leqq \alpha, \\ 0 & \text{if } s > \beta. \end{cases} \tag{1.2}$$

We can see that this Markov chain $(\tau_n, x(\tau_n))_{n=1}^{\infty}$ has the properties

$$\tau_n < \tau_{n+1}, \qquad \tau_n \to \infty \quad \text{as } n \to \infty.$$

From this a continuous time process $x(\cdot)$ can be constructed as follows.

$$x(t) = x(\tau_n), \qquad \tau_n \leqq t < T_{n+1}.$$

This process is a jump process in the sense that for a.s. and any $t > 0$, $\exists h > 0$ depending on sample such that $x(t) = x(s)$ for $s \in [t, t+h]$. Also it is a Markov process, as stated in the following theorem.

THEOREM 2. *$\{x(\cdot), P_{s,x}\}$ is a Markov process where the $P_{s,x}$ are defined as follows:*

$$P_{s,x}(t, B) \equiv P(s, x, t, B) = \sum_{n=0}^{\infty} \int_s^t \int_B q(y, v, t) \cdot Q^{(n)}(s, x, dv, dy),$$

$$Q^{(n+1)}(z, \tilde{B}) = \int Q^{(n)}(z, d\tilde{z}) Q(\tilde{z}, B), \qquad Q^{(1)} = Q, \tag{1.3}$$

$$q(y, s, t) = \exp\left(-\int_s^t a(\theta, y)\, d\theta\right).$$

The proof of this theorem can be found in [5, Chap. III].

Remark 1. τ_n is the nth jump time of the process and $x(\tau_n)$ is the state of $x(\cdot)$ following immediately after the nth jump. $1/a(\cdot, \cdot)$ is the expected time the jump occurred. $\pi(t, x, \cdot)$ describe the probability of where the process will go after the jump.

Remark 2. The reason we construct a process in this way is intuitively clear as we explain in Remark 1. Theoretically it follows from the theorem [5, p. 191] that any jump process can have a Markov chain naturally associated to it. The construction above is just the reverse procedure of this. But the relation and (1.1) is not very clear from this construction. We will show below that the probability measure of $x(\cdot)$ above is a solution of a "martingale problem", in modern terminology.

In the following, $x(\cdot)$ will be the Markov process constructed as above. $E_{s,x}$ denotes the expectation with respect to the probability measure of $x(\cdot)$ starting from x at $s \cdot \mu(t, x, B) = a(t, x)\pi(t, x, B)$.

We begin with defining the operator $L^t : B(R^n) \to B(R^n)$ as follows:

$$L^t f(x) = a(t, x) \int (f(x+y) - f(x))\pi(t, x, dy). \tag{1.4}$$

LEMMA 3. *$f(x_t) - \int_s^t L^\theta f(x_\theta)\, d\theta$ is a martingale for $f \in B(R^n)$.*

Proof.

$$\begin{aligned} E_{s\cdot x}\left[\int_s^t L^\theta f(x_\theta)\, d\theta\right] &= \int_s^t E_{s,x}[L^\theta f(x_\theta)]\, d\theta \\ &= \int_s^t \int_{y \in R^n} L^\theta f(y) P(s, x, \theta, dy)\, d\theta \qquad (1.5) \\ &= \int_s^t \int_y \int_z f(y+z)\mu(\theta, y, dz) P(s, x, \theta, dy)\, d\theta \\ &\quad - \int_s^t \int_y f(y) a(\theta, y) P(s, x, \theta, dy)\, d\theta. \end{aligned}$$

Using the decomposition

$$P(s, x, \theta, dy) = \sum_{n=0}^{\infty} \int_s^\theta \exp\left(-\int_v^\theta a(\alpha, y)\, d\alpha\right) Q^n(s, x, dv, dy)$$

and the relations

$$
\begin{aligned}
&\int_s^t \int_y \int_z f(y+z)\mu(\theta, y, dz) \int_s^\theta \exp\left(-\int_v^\theta a(\alpha, y)\, d\alpha\right) Q^n(s, x, dv, dy)\, d\theta \\
&\quad = \int_s^t \int_y Q^{(n)}(s, x, dv, dy) \int_v^t \int_z f(y+z) \exp\left(-\int_v^\theta a(\alpha, y)\, d\alpha\right) \mu(\theta, y, dz)\, d\theta \\
&\quad = \int_s^t \int_y Q^{(n)}(s, x, dv, dy) E_{v,y}[f(x_{\tau_1}); \tau_1 \leqq t] \\
&\quad = E_{s,x}[E_{\tau_n, x_{\tau_n}}[f(x_{\tau_1}); \tau_1 \leqq t]; \tau_n \leqq t] \\
&\quad = E_{s,x}[E[f(x_{\tau_{n+1}}); \tau_{n+1} \leqq t | (\tau_n, x_{\tau_n})]; \tau_n \leqq t] \\
&\quad = E_{s,x}[f(x_{\tau_{n+1}}); \tau_{n+1} \leqq t],
\end{aligned} \tag{1.6}
$$

$$
\begin{aligned}
&\int_s^t \int_y f(y) a(\theta, y)\, d\theta \int_s^\theta \exp\left(-\int_v^\theta a(\alpha, y)\, d\alpha\right) Q^{(n)}(s, x, dv, dy) \\
&\quad = \int_s^t Q^n(s, x, dv, dy) \left(\int_v^t a(\theta, y) \exp\left(-\int_v^\theta a(\alpha, y)\, d\alpha\right) d\theta\right) f(y) \\
&\quad = \int_s^t Q^n(s, x, dv, dy) f(y) E_{v,y}[\tau_1 \leqq t] \\
&\quad = E_{s,x}[f(x_{\tau_n}) E_{\tau_n, \tau_n}[\tau_1 \leqq t]; \tau_n \leqq t] \\
&\quad = E_{s,x}[E[f(x_{\tau_n}); \tau_{n+1} \leqq t] | [\tau_n, x_{\tau_n}]] \\
&\quad = E_{s,x}[f(x_{\tau_n}); \tau_{n+1} \leqq t],
\end{aligned} \tag{1.7}
$$

we have

$$
\begin{aligned}
&E_{s,x}\left[\int_s^t L^\theta f(x_\theta)\, d\theta\right] \\
&\quad = \sum_{n=0}^\infty E_{s,x}[f(x_{\tau_{n+1}}); \tau_{n+1} \leqq t] - E_{s,x}[f(x_{\tau_n}); \tau_{n+1} \leqq t] \\
&\quad = \sum_{n=0}^\infty E_{s,x}[f(x_{\tau_n}); \tau_n \leqq t] - E_{s,x}[f(x_{\tau_n}); \tau_{n+1} \leqq t] - E_{s,x}[f(x_s); \tau_0 \leqq t] \\
&\quad = \sum_{n=0}^\infty E_{s,x}[f(x_{\tau_n}); \tau_n \leqq t < \tau_{n+1}] - f(x) \\
&\quad = \sum_{n=0}^\infty E_{s,x}[f(x_t); \tau_n \leqq t < \tau_{n+1}] - f(x) \\
&\quad = E_{s,x}[f(x_t)] - f(x).
\end{aligned}
$$

LEMMA 4.

$$
f(t, x_t) - \int_s^t \left\{\frac{\partial f(\theta, x_\theta)}{\partial \theta} + L^\theta f_\theta(x_\theta)\right\} d\theta
$$

are martingales for f, bounded, absolutely continuous in θ with $\partial f(\theta, x)/\partial\theta$ bounded.

Proof. We first consider the case where f is continuously differentiable with bounded derivative.

$$
f(t, x_t) - f(s, x_s) = \sum f(t_i, x_{t_i}) - f(t_{i-1}, x_{t_{i-1}}), \quad \text{where } s = t_1 < t_2 < \cdots < t_n = t
$$

is a partition of $[s, t]$.

$$f(t_i, x_{t_i}) - f(t_{i-1}, x_{t_{i-1}}) = f(t_i, x_{t_i}) - f(t_i, x_{t_{i-1}}) + \int_{t_{i-1}}^{t_i} \frac{\partial f(\theta, x_{t_{i-1}})}{\partial \theta}\, d\theta.$$

By applying Lemma 3 we have

$$E_{s,x}[f(t_i, x_{t_i}) - f(t_{i-1}, x_{t_{i-1}})]$$
$$= E_{s,x}\left[\int_{t_{i-1}}^{t_i} L^\theta f_{t_i}(x_\theta)\, d\theta\right] + E_{s,x}\left[\int_{t_{i-1}}^{t_i} \frac{\partial f(\theta, x_{i-1})}{\partial \theta}\, d\theta\right],$$

$$E_{s,x}\left[f(t, x_t) - f(s, x_s) - \int_s^t \frac{\partial f}{\partial \theta}(\theta, x) + L^\theta f_\theta(x_\theta)\, d\theta\right]$$
$$= \sum E_{s,x}\left[\int_{t_{i-1}}^{t_i} (L^\theta f_{t_i} - L^\theta f_\theta)(x_\theta)\, d\theta + E_{s,x}\int_s^t \left(\frac{\partial f}{\partial \theta}(\theta, x_\theta^{(n)}) - \frac{\partial f}{\partial \theta}(\theta, x_\theta)\right) d\theta\right]$$

where $x_\theta^{(n)} = x_{t_{i-1}}$ for $t_{i-1} \leqq \theta < t_i$. Since $x(\cdot)$ is right continuous, $x^{(n)}(\theta) \to x(\theta)$ as $\max(t_i - t_{i-1}) \to 0$. Passing to the limit, we have

$$E_{s,x}\left[f(t, x_t) - f(s, x_s) - \int_s^t \frac{\partial f}{\partial \theta}(\theta, x_\theta) + L^\theta f_\theta(x_\theta)\, d\theta\right] = 0.$$

For the general case we choose $g^{(n)}$, $f^{(n)}$ as follows.

$$g^{(n)} \to \frac{\partial f}{\partial \theta} \quad \text{bounded pointwise in } (\theta, x),$$

$$f^{(n)}(t, x) = f^{(n)}(s, x) + \int_s^t g^{(n)}(\theta, x)\, d\theta,$$

with $f^{(n)}(s, x) \to f(s, x)$ bounded pointwise, $g^{(n)}$ bounded continuous.

The above discussion guarantees that

$$E_{s,x}\left[f^{(n)}(t, x_t) - f^{(n)}(s, x_s) - \int_s^t \left(\frac{\partial f^{(n)}}{\partial \theta}(\theta, x_\theta) + L^\theta f_\theta^{(n)}(x_\theta)\right) d\theta\right] = 0.$$

Letting $n \to \infty$, we have

$$E_{s,x}\left[f(t, x_t) - f(s, x_s) - \int_s^t \left(\frac{\partial f}{\partial \theta}(\theta, x_\theta) + L^\theta f_\theta(x_\theta)\right) d\theta\right] = 0.$$

Remark 3. For the case that L^t is independent of t, we can get (1.1) easily from Lemma 4.

2. Exit probability. We will consider the exit probabilities of a family of jump processes depending on a small parameter $\varepsilon > 0$. Our model is as follows. We note that this has been studied in [7] by another method.

Let $a(\cdot)$ be a positive Borel bounded function on R^n; $\pi(x, dy)$ is a kernel on R^n satisfying $\pi(x, \{0\}) = 0$. Define operators Lf, $L_\varepsilon f$,

$$Lf(x) = a(x) \int (f(x+y) - f(x))\pi(x, dy),$$

$$L_\varepsilon f(x) = \frac{a(x)}{\varepsilon} \int (f(x+\varepsilon y) - f(x))\pi(x, dy).$$

We know that these operators generate jump processes $x(\cdot)$, $x_\varepsilon(\cdot)$. So does L_ε^u. u is a positive Borel function such that u, u^{-1} are bounded, and L_ε^u is given by

$$L_\varepsilon^u f(x) = \frac{1}{u(x)}(L_\varepsilon(uf)(x) - fL_\varepsilon(u)(x))$$
$$= \frac{a(x)}{\varepsilon}\int \frac{u(x+\varepsilon y)}{u(x)}(f(x+\varepsilon y)-f(x))\pi(x, dy).$$

Therefore $L_\varepsilon = L_\varepsilon^u$ with $u \equiv 1$. In addition we need $L_\varepsilon^{u_t} f$,

$$L_\varepsilon^{u_t} f(x) = \frac{1}{u_t(x)}(L_\varepsilon(u_t f)(x) - fL_\varepsilon(u_t)(x)),$$

for a positive function $u_t(x) \equiv u(t, x)$ on $R^+ \times R^n$. We denote $E_{x,s}^{\varepsilon,u}[\cdot]$ as the expectation with respect to the probability measure of the Markov process $x(\cdot)$ constructed from L_ε^u according to Theorem 2 with $x(s) = x$; here $u = u(t, x)$, a function as above. The abbreviation $E_{s,x}^u[\cdot]$ will be used occasionally if we fix ε in the discussion. $E_x^\varepsilon[\cdot]$ means $E_{s,x}^{\varepsilon,u}$ when $s = 0$, $u \equiv 1$. i.e., $E_x^\varepsilon[x(\cdot) \in B] = E_x[x_\varepsilon(\cdot) \in B]$.

DEFINITION. We fix a bounded open subset Ω of R^n. τ, called the exit time from Ω, is defined as $\tau = \inf\{t; x(t) \notin \Omega\}$ if $\{t; x(t) \notin \Omega\} \neq \varnothing$ and ∞ if it is empty. Let

$$\phi^\varepsilon(t, x) = \begin{cases} E_x^\varepsilon[\tau \leqq t] & \text{for } t > 0 \text{ and } x \in \Omega, \\ 1, & x \notin \Omega. \end{cases}$$

We are interested in the behavior of $\phi^\varepsilon(t, x)$ as $\varepsilon \to 0$. For the case where $x_\varepsilon(\cdot)$ are diffusions, the behavior of $\phi^\varepsilon(\cdot, \cdot)$ as $\varepsilon \to 0$ has been studied in [2]. We will follow the ideas in [2] together with the general formulation of control method presented in the introduction.

In the rest of this section, we will mention some properties of ϕ^ε.

LEMMA 5. *$d\phi^\varepsilon(t, x)/dt = L_\varepsilon \phi_t^\varepsilon(x)$ for $x \in \Omega$.*

Proof. We only give the proof for $\varepsilon = 1$.

$$\phi(t, x) = \sum_{n=1}^{\infty} P_s[\tau_1 + \cdots + \tau_n \leqq t, x(\tau_i) \in \Omega, i = 1, \cdots, n-1, x(\tau_n) \notin \Omega]$$
$$= P_x[\tau_1 \leqq t, x(\tau_1) \notin \Omega]$$
$$+ \sum_{n=2}^{\infty} E_x[E_{x(\tau_1)}[\tilde{\tau}_1 + \cdots + \tilde{\tau}_{n-1} \leqq t - \tau_1, x(\tilde{\tau}_i) \in \Omega,$$
$$i = 1, \cdots, n-2, x(\tilde{\tau}_{n-1}) \notin \Omega]; x(\tilde{\tau}_1) \in \Omega, \tau_1 \leqq t]$$
$$= P_x[\tau_1 \leqq t, x(\tau_1) \notin \Omega] + E_x[\phi(t - \tau_1, x(\tau_1)); x(\tau_1) \in \Omega, \tau_1 \leqq t]$$
$$= \int_0^t a(x)\, e^{-sa(x)} \int_{x+y \notin \Omega} \pi(x, dy)$$
$$+ \int_0^t a(x)\, e^{-sa(x)} \int_{x+y \in \Omega} \phi(t-s, x+y)\pi(x, dy)\, ds$$
$$= \int_0^t a(x)\, e^{-(t-s)a(x)} \int \phi(s, x+y)\pi(x, dy)\, ds.$$

Note that $\phi(s, z) = 1$ if $z \notin \Omega$. From this relation, $\phi(s, x)$ is a continuously differentiable function in s. Differentiating both sides of the above equation, we get

$$\frac{d\phi}{dt}(t, x) = a(x) \int \phi(t, x+y)\pi(x, dy) - a(x)\phi(t, x) = L\phi_t(x). \tag{2.1}$$

As was mentioned in the introduction, we are trying to apply a logarithmic transformation to ϕ^ε, i.e., we shall consider $\psi^\varepsilon(t, x) = -\varepsilon \log \phi^\varepsilon(T-t, x)$ where T is fixed and $t < T$. The reason why we consider $\phi^\varepsilon(T-t, \cdot)$ instead of $\phi^\varepsilon(t, \cdot)$ will be apparent later. Also we will see why we consider $\varepsilon \log \phi^\varepsilon$ instead of $\log \phi^\varepsilon$. Another problem concerning us is that we do not know whether $\phi^\varepsilon > 0$. The following device is commonly used for applying this procedure. For each $d > 0$, a small number, let $\tilde{\phi}^\varepsilon(s, x)$ be the solution of the differential equation

$$\begin{aligned} &\frac{d\tilde{\phi}}{dt}(t, x) = L_\varepsilon \tilde{\phi}^\varepsilon_t(t, x), && x \in \Omega, \\ &\tilde{\phi}^\varepsilon(0, x) = d^{1/\varepsilon}, \\ &\tilde{\phi}^\varepsilon(s, x) = 1, && x \notin \Omega. \end{aligned} \tag{2.2}$$

LEMMA 6. *$\tilde{\phi}^\varepsilon(t, x) \geqq d^{1/\varepsilon}$ for all $t \geqq 0$ and $\|\tilde{\phi}^\varepsilon - \phi^\varepsilon\| \leqq d^{1/\varepsilon}$.*

Proof. We note that $\tilde{\phi}^\varepsilon(t, x) = E_x[f_1(x_t)]$, $\phi^\varepsilon(t, x) = E_x[f_0(x_t)]$ where

$$f_1(y) = \begin{cases} 1, & x \notin \Omega, \\ 1/d^\varepsilon, & x \in \Omega, \end{cases} \qquad f_0(y) = \begin{cases} 1, & x \notin \Omega, \\ 0, & x \in \Omega. \end{cases}$$

Then the result follows easily.

3. Control problem. The basic approach is the following. First we show under some conditions $\tilde{\psi}^\varepsilon(t, x) = -\varepsilon \log \tilde{\phi}^\varepsilon(T-t, x)$ converges to a limit $I(t, x) \geqq 0$ which is independent of d for d being small enough. This means $\tilde{\phi}^\varepsilon(t, x) \sim \exp(-I(T-t, x)/\varepsilon)$. Then by Lemma 6, the difference between $\tilde{\phi}^\varepsilon(t, x)$ and $\phi^\varepsilon(t, x)$ is $d^{1/\varepsilon}$ which is very small compared to $\exp(-I(T-t, x)/\varepsilon)$ for d small enough. This will imply $\phi^\varepsilon(t, x) > 0$ for ε small and that $\psi^\varepsilon(t, x)$ converges to $I(t, x)$.

Before we give necessary conditions such that the above procedure works, we state one more property of $\tilde{\phi}^\varepsilon$.

$$\frac{d\tilde{\psi}^\varepsilon}{dt}(t, x) = \varepsilon e^{\tilde{\psi}^\varepsilon/\varepsilon} L_\varepsilon e^{-\tilde{\psi}^\varepsilon/\varepsilon} = -\varepsilon \min\left[L^u_\varepsilon\left(\frac{\tilde{\psi}^\varepsilon}{\varepsilon}\right) + \frac{1}{\varepsilon}k^u_\varepsilon\right] = -\min\left[L^u_\varepsilon(\tilde{\psi}^\varepsilon) + k^u_\varepsilon\right],$$

where

$$L^u_\varepsilon f(x) = \frac{a(x)}{\varepsilon} \int \frac{u(x+\varepsilon y)}{u(x)}(f(x+\varepsilon y) - f(x))\pi(x, dy).$$

$$k^u_\varepsilon(x) = a(x) \int \left(\frac{u(x+\varepsilon y)}{u(x)} \log \frac{u(x+\varepsilon y)}{u(x)} - \frac{u(x+\varepsilon y)}{u(x)} + 1\right)\pi(x, dy) \tag{3.1}$$

where $u \in F$ if u, u^{-1} are positive bounded Borel functions on R^n.

For the time being we drop ε; then the equation for $\tilde{\psi}^\varepsilon$ becomes

$$\frac{d\tilde{\psi}}{dt} + \min_{u \in F}[L^u(\tilde{\psi}) + k^u] = 0, \qquad t \in [0, T]. \tag{3.2}$$

Let

$$\tilde{\psi}(T, x) = W(x) = \begin{cases} 0 & \text{if } x \notin \Omega, \\ -\log d & \text{if } x \in \Omega. \end{cases}$$

We recognize that this is just the dynamic programming equation of the following stochastic control problem:

$$\tilde{J}(s, x) = \inf_{u \in \tilde{F}} E^u_{s,x}\left[\int_s^{\tau \wedge T} k^{u_t}(x_t)\, dt + W(x_{t \wedge T})\right], \tag{3.3}$$

where $u \in \tilde{F}$ if u, u^{-1} are positive bounded Borel functions on $[0, T] \times R^n$.

We have the following result describing the relation for $\tilde{\psi}$ and $\tilde{J}$, similar to the so-called "verification theorem."

LEMMA 7. $\tilde{\psi}(t, x) = \tilde{J}(t, x)$.

Proof. For each $u \in \tilde{F}$, we apply Lemma 4 to get

$$E^u_{s,x}\left[\tilde{\psi}(\tau \wedge T, x_{\tau \wedge T}) - \int_s^{\tau \wedge T} \left(L^{u_\theta}\tilde{\psi}_\theta(x_\theta) + \frac{\partial \tilde{\psi}}{\partial t}(\theta, x_\theta)\right) d\theta\right] = \tilde{\psi}(s, x).$$

By (3.2), $d\tilde{\psi}/d\theta + L^{u_\theta}\tilde{\psi} + k^{u_0} \geqq 0$. These two relationships imply

$$\tilde{\psi}(s, x) \leqq E^u_{s,x}\left[\int_s^{\tau \wedge T} k^{u_\theta}(x_\theta)\, d\theta + W(x_{\tau \wedge T})\right],$$

i.e. $\tilde{\psi}(s, x) \leqq \tilde{J}(s, x)$ by taking the infimum over $u \in \tilde{F}$.

On the other hand, $u^* = \exp(\tilde{\psi})$ gives us the optimal control as shown in the following.

$$\tilde{\psi}(s, x) = E^{u^*}_{s,x}\left[\tilde{\psi}(\tau \wedge T, x_{\tau \wedge T}) + \int_s^{\tau \wedge T} k^{u^*}(x_\theta)\, d\theta\right] \geqq \tilde{J}(s, x)$$

holds as above. This time we use $d\tilde{\psi}/d\theta + L^{u^*_\theta}\tilde{\psi} + k^{u^*_\theta} = 0$. We then get $\tilde{\psi}(s, x) = \tilde{J}(s, x)$.

Remark 4. Following the same argument we can show that

$$\tilde{J}(s, x) \leqq E^u_{s,x}\left[\int_s^{\tau \wedge T} k^{u_t}(x_t)\, dt + W(x_{\tau \wedge T})\right]$$

for those u such that $k^{u_t}(x)$ is bounded as a function of (t, x).

Now we plan to see how $J = \tilde{\psi}$ depends on ε; therefore rewrite (3.3) again as

$$\begin{aligned} J^\varepsilon(s, x) &= \inf_{u \in \tilde{F}} E^{\varepsilon, u}_{s,x}\left[\int_s^{\tau \wedge T} k^u_\varepsilon(x_t)\, dt + W(x_{\tau \wedge T})\right] \\ &\equiv \inf E^\varepsilon_{s,x}\left[\int_s^{\tau \wedge T} k^u(x_t)\, dt + W\right]. \end{aligned} \tag{3.4}$$

Here, in order to simplify the notation, we agree to use the second form on the right if it will not give any confusion. Otherwise we will write down all necessary parameters to make the argument clear.

Associated with these control problems, there is a deterministic control problem

$$J(s, x) = \inf\left\{\int_s^{\tau \wedge T} k^u(x_t)\, dt + W(x_{\tau \wedge T})\right\} \tag{3.5}$$

where x_t satisfies

$$\frac{dx(t)}{dt} = v(u_t, x_t), \qquad v(u, z) = a(z) \int u(y) y \pi(z, dy),$$

$$x(s) = x, \tag{3.6}$$

$$k^u(x) = a(x) \int (u(y) \log u(y) - u(y) + 1) \pi(x, dy).$$

There is a little difference between k^u and k^u_ε. The choice of k^u in this form is indicated by the following example.

Example. Consider the one-dimensional case with $a(x)$, $\pi(x, B)$ described as follows

$$a(x)\pi(x, \{1\}) = a^+(x),$$
$$a(x)\pi(x, \{-1\}) = a^-(x),$$
$$\pi(x, \cdot) \text{ concentrated on } \{1, -1\},$$
$$a^+, a^- \text{ positive, bounded.}$$

Under $E^{\varepsilon,u}_{s,x}$, $x(\cdot)$ is a process with state space $\{x + n\varepsilon, n \text{ is integer}\} = D$.

$$L^w_\varepsilon f(y) = \frac{1}{\varepsilon}\left(a^+(y)\frac{w^+(y)}{w(y)}(f(y+\varepsilon) - f(y)) + a^-(y)\frac{w^-(y)}{w(y)}(f(y-\varepsilon) - f(y)) \right),$$

$y \in D$, f is a function defined on D. $w^+(y) = w(y+\varepsilon)$, $w(y)^- = w(y-\varepsilon)$ in terms of original notation. Take $f(y) = y$; the above takes the form

$$L^u_\varepsilon f(y) = (a^+(y)u^+(y) - a^-(y)u^-(y)),$$
$$k^u_\varepsilon(y) = a^+(y)(u^+(y) \log u^+(y) - u^+(y) + 1) + a^-(y)(u^-(y) \log u^-(y) - u^-(y) + 1).$$

Here we change the notation $u^+(y) \leftrightarrow w^+(y)/w(y)$, $u^-(y) \leftrightarrow w^-(y)/w(y)$. Then we expect that $J^\varepsilon(s, x)$ converges to $J(s, x)$,

$$J^\varepsilon(s, x) = \inf E^\varepsilon_{s,x}\left[\int_s^{\tau \wedge T} k^u(x_t)\, dt + W \right],$$

$$J(s, x) = \inf \int_s^{\tau \wedge T} k^u(x_t)\, dt + W.$$

Here, in the stochastic control case $x(\cdot)$ is such that

$$x(t) - \int_s^t \{a^+(x_t)u^+(x_t) - a^-(x_t)u^-(x_t)\}\, dt$$

is a martingale with variation being $O(\varepsilon)$, whereas in the deterministic case

$$x(t) = x + \int_s^t \{a^+(x_t)u^+(x_t) - a^-(x_t)u^-(x_t)\}\, dt.$$

Remark 5. In the one-dimensional case, and also in the n-dimensional case with $\pi(x, \cdot)$ concentrated on some atoms, we can show that ϕ^ε is larger than zero. Then we do not need to use the device sketched at the beginning of this section.

We go one step further, by formal calculation, to look at a property of the deterministic control. We write

$$J(s,x)=\inf_{dx_t/dt=v(t)}\ \inf_{v(t)=v(u_t,x_t)}\left\{\int_s^{\tau\wedge T} k^{u_t}(x_t)\,dt+W\right\}. \tag{3.7}$$

This form suggests that we consider

$$\begin{aligned} &k(v,x)=\inf_{\substack{v=v(u,x),\\ u\geqq 0}} k^u(x), \qquad v\in R^n,\\ &v(u,x)=a(x)\int u(y)y\pi(x,dy). \end{aligned} \tag{3.8}$$

Fix $x\in R^n$, and call $a(x)\pi(x,\cdot)=\mu(\cdot)$. At this moment, we assume

$$\int e^{K|y|}\mu(dy)<\infty \quad \forall K>0. \tag{3.9}$$

Denote

$$k^u=\int (u(y)\log u(y)-u(y)+1)\mu(dy), \qquad u\geqq 0,$$

$$k(v)=\inf_{\substack{v=v(u),\\ u\geqq 0}} k^u, \qquad v\in R^n,$$

$$v(u)=\int u(y)y\mu(dy).$$

LEMMA 8. *If there is $c\in R^n$ such that $u^*=e^{c\cdot y}$ satisfies $v(u^*)=v$, then $k(v)=k^{u^*}$.*

Proof. Let u be such that $u\geqq 0$ and $v(u)=v$. Denote $h=u-u^*$, $u_\alpha=u^*+\alpha h$ for $0\leqq\alpha\leqq 1$. We have $v(u_\alpha)=v$, $u_\alpha\geqq 0$ and

$$\begin{aligned} \frac{dk^{u_\alpha}}{d\alpha}&=\int h(y)\log(e^{c\cdot y}+\alpha h(y))\mu(dy)\\ &\geqq\int h(y)c\cdot y\mu(dy)\\ &=c\cdot\int h(y)y\mu(dy)=0, \end{aligned}$$

$k^u\geqq k^{u^*}$. Here we use $r\log(e^\alpha+r)\geqq r\alpha$ if $e^\alpha+r\geqq 0$.

Remark 6. There is at most one u satisfying $u\geqq 0$, $v(u)=v$, $k(v)=k^u$ in the following sense. If there is a u^* as in Lemma 8 and another u satisfies the above condition, then $u^*=u$ a.s. μ. The proof is given as follows. Let $u_\alpha=u^*+\alpha h$, $h=u-u^*$, $0\leqq\alpha\leqq 1$. Since $k^u=k^{u^*}$ and k^u is convex in u, we must have $k^{u_\alpha}=k^u=k^{u^*}$ also. Therefore

$$0=\frac{dk^{u_\alpha}}{d\alpha}=\int h(y)\log(e^{c\cdot y}+\alpha h(u))\mu(dy)\geqq\int h(y)c\cdot y\mu(dy)=0.$$

This implies

$$\int h(y)\log(e^{c\cdot y}+\alpha h(y))\mu(dy)=\int h(y)c\cdot y\mu(dy).$$

Now we use the property $h(y)\log(e^{c\cdot y}+\alpha h(y))\geqq h(y)c\cdot y$; equality holds if and only if $h(y)=0$. Then from these two properties it is clear that we have

$$u-u^*=h=0 \quad \text{a.s. } \mu.$$

Remark 7. If $u=e^{c\cdot y}$ satisfies $v=v(u)$, then $k(v)=k^u$ takes the following familiar form which states that $k(v)$ is the dual to Laplace transform of $\mu(+\text{const})$

$$\begin{aligned} k(v)=k^u &= \int (e^{c\cdot y}\log e^{c\cdot y}-e^{c\cdot y}+1)\mu(dy) \\ &= c\cdot\int e^{c\cdot y}y\mu(dy)-\int e^{c\cdot y}\mu(dy)+a \\ &= c\cdot v-\int e^{c\cdot y}\mu(dy)+a \\ &= \sup_z\left(z\cdot v-\int e^{z\cdot y}\mu(dy)\right)+a, \end{aligned}$$

where $a=\mu(R^n)$.

4. Assumption and main results. We make the following assumptions.

A1. There are a probability measure π on R^n, a function $g(x, y)$ on $R^n\times R^n$ and a function $a(x)$ on R^n such that

$$\pi(x, dy)=g(x, y)\pi(dy).$$

A2. The convex hull of the support of π contains a neighborhood of the origin. There is an $\alpha>0$ such that $\int\exp(\alpha|y|^2)\pi(dy)<\infty$.

A3. There are c_1, c_2, $K>0$ such that

$$\begin{aligned} &c_1\leqq g(x, y)\leqq c_2, \qquad c_1\leqq a(x)\leqq c_2, \\ &|g(x_1, y)-g(x_2, y)|\leqq K|x_1, x_2|, \qquad |a(x_1)-a(x_2)|\leqq K|x_1-x_2|. \end{aligned}$$

Integrability of $\exp(\alpha|y|^2)$ with respect to π seems to be a very strong condition which is needed in proving (4.2) below. It seems that integrability of $\exp(K|y|)\,\forall\, K$ is enough. On the other hand, (4.2) enables us to control some processes.

We state first some easy consequence of the assumptions.

LEMMA 10.

$$\exp(c\cdot y)\in L^1(\pi), \tag{4.1}$$

$$\int g(y)y^2\mu(x, dy)\leqq c_1\int(g(y)\log g(y)-g(y)+1)\mu(x, dy)+c_2, \tag{4.2}$$

where $g\geqq 0$, c_1, c_2 are constants which are independent of g, $\mu(x, dy)=a(x)\pi(x, dy)$. Also

$$\left|\int g(y)\mu(x, dy)-g(y)\mu(\tilde{x}, dy)\right|\leqq c|x-\tilde{x}|\int g(y)\mu(x, dy), \tag{4.3}$$

where $g\geqq 0$.

Proof. (4.1) follows from A2; (4.3) follows from A3. As for (4.2), we note that

$$\int g(y)y^2\mu(x, dy) = \int_{\beta y^2 < \log g(y)} g(y)y^2\mu(x, dy) + \int_{\beta y^2 \geqq \log g(y)} g(y)y^2(x, dy)$$

$$\leqq \frac{1}{\beta}\int g(y)\log g(y)\mu(x, dy) + \int y^2 \exp(\beta y^2)\mu(x, dy)$$

$$\leqq c_1 \int (g(y)\log g(y) - g(y) + 1)\mu(x, dy) + c_2$$

if we take $\beta < \alpha$ in A2.

In the next paragraph we are going to show $\overline{\lim}\, J^\varepsilon(s, x) \leqq J(s, x)$. Under condition A2, (3.9) is satisfied. Then we have the following alternative formula for $J(s, x)$.

THEOREM 3. *$J(s, x) = \inf_{dx_t/dt = v(t), x_s = x} \{\int_s^{\tau \wedge T} k(v(t), x(t))\, dt + W\}$ where $v(t)$ is a bounded function on $[s, T]$,*

$$k(v, x) = \inf_{\substack{v = v(u,x),\\ u \geqq 0}} \int (u(y)\log u(y) - u(y) + 1)\mu(x, dy)$$

as in (3.8).

We first indicate the following lemma in convex analysis.

LEMMA 11. *Let $F: R^n \to R$ be strictly convex, differentiable and satisfy*

$$\lim_{\lambda \to \infty} \frac{1}{\lambda} F(\lambda c) = \infty \quad \forall c \neq 0.$$

Then $\nabla F: R^n \to R^n$ is one-one onto and the conjugate F^* has the same properties as F. F^* is given by

$$F^*(c^*) = \sup\{\langle c^*, c\rangle - F(c)\}.$$

Moreover $F^*(c^*) = \langle \nabla F(c), c\rangle - F(c)$ if $c^* = \nabla F(c)$. Thus we also have $\nabla F^*(c^*) = c$.

We refer to [6, p. 259] for the proof of this lemma.

Proof of the theorem. We now apply Lemma 11 to the function

$$F(c) = F(c, x) = \int e^{c\cdot y}\mu(x, dy).$$

Under our condition A2, F, as a function of c, satisfies the conditions in Lemma 11. In fact, strict convexity of F follows from the fact that $\pi(\{y; c \cdot y = 0\}) \neq \pi(R^n)\, \forall c \neq 0$, and the property $\lim((1/\lambda)F(\lambda c) = \infty$ follows from $\pi(\{y; c \cdot y > 0\}) > 0$. Lemma 11 tells us that for each $v \in R^n$, $x \in R^n$ there is a unique c such that

$$v = \nabla F(c) = \int e^{c\cdot y} y\mu(x, dy) = v(u, x), \qquad u = e^{c\cdot y}.$$

Following Lemma 8, we have

$$k(v, x) = \int (e^{c\cdot y}(c \cdot y) - e^{c\cdot y} + 1)\mu(x, dy) = F^*(v, x) + a(x) \tag{4.4}$$

by Remark 7. Clearly

$$J(s, x) \geqq \inf\left\{\int_s^{\tau \wedge T} k(\dot{x}(t), x(t))\, dt + W\right\}$$

by the construction of $k(v, x)$. On the other hand, if $dx_t/dt = v(t)$ with $v(t)$ bounded in $[s, T]$, then we can choose $u_t = e^{c_t \cdot y}$ with c_t satisfying $v(t) = \nabla c F(c_t, x_t) = v(u_t, x_t)$. Clearly

$$\frac{dx_t}{dt} = v(u_t, x_t), \qquad \int_s^{\tau \wedge T} k(\dot{x}(t), x(t))\, dt + W = \int_s^{\tau \wedge T} k^{u_t}(x_t)\, dt + W.$$

Thus the reverse inequality holds and this ends the proof of Theorem 3.

In the following, if we want to emphasize the dependence of J^ε, J on d we write $J^\varepsilon(s, x; d)$, $J(s, x, d)$.

THEOREM 4. $\overline{\lim}_{\varepsilon \to 0} J^\varepsilon(s, x, d) \leqq J(s, x, d)$ *for d small enough.*

Proof. Clearly $0 \leqq J(s, x, d) \leqq J(s, x, d') \leqq J(s, x, 0) < \infty$ if $d \geqq d' > 0$. Consider

$$J(s, x, d) = \inf\left(\int_s^{\tau \wedge T} k^{u_t}(x_t)\, dt + W(x_{\tau \wedge T}) \right)$$

with $W(x) = 0$ if $x \notin \Omega$ and $-\log d$ if $x \in \Omega$. We see if $\tau > T$, then $\int_s^{\tau \wedge T} k^{u_t}(x_t)\, dt + W(x_{\tau \wedge T}) \geqq -\log d$. This will be larger than $J(s, x, 0)$ if $d \leqq d^0$, where d^0 is a small number depending on $T - s$. From this it is not difficult to see that

$$\begin{aligned} J(s, x, d) &= \inf_{\tau \leqq T} \int_s^{\tau \wedge T} k^{u_t}(x_t)\, dt = \inf_{\tau < T} \int_s^{\tau \wedge T} k^{u_t}(x_t)\, dt, \qquad d \leqq d_0 \\ &= \inf_{\tau < T} \int_s^{\tau \wedge T} k(\dot{x}(t), x(t))\, dt = J(s, x, 0). \end{aligned}$$

For a small number $\delta > 0$, there is $v(t) = v(u_t, x_t)$, $u_t = e^{c_t \cdot y}$ with c_t bounded, such that

$$\begin{aligned} & J(s, x) \geqq \int_s^{\tau_0} k^u(x_t)\, dt - \delta, \\ & \frac{dx_t}{dt} = v(u_t, x_t) = \int e^{c_t \cdot y} y \mu(x_t, dy), \\ & x_s = x, \\ & \tau_0 < T, \quad \tau_0 \text{ the exit time of } x_t. \end{aligned} \tag{4.5}$$

Let $E^\varepsilon_{s,x}$ be the expectation with respect to the probability measure of the Markov process x^ε with generator L^u_ε, $u^\varepsilon_t(y) = u_t(y/\varepsilon)$. Then

$$\begin{aligned} d(x^\varepsilon_t - x_t)^2 &= d(x^\varepsilon_t)^2 - 2x_t\, dx^\varepsilon_t - 2x^\varepsilon_t\, dx_t - 2x_t\, dx_t \\ &= 2x^\varepsilon_t \cdot \int y u_t(y) \mu(x^\varepsilon_t, dy) - 2x_t \cdot \int y u_t(y) \mu(x^\varepsilon_t, dy) \\ &\quad - 2x^\varepsilon_t \cdot \int y u_t(y) \mu(x_t, dy) + 2x_t \cdot \int y u_t(y) \mu(x_t, dy) \\ &\quad + \varepsilon \int y^2 u_t(y) \mu(x^\varepsilon_t, dy) + dM(t). \end{aligned}$$

$M(t)$ is a martingale. Here we make use of Lemma 3 for calculating dx_t^ε and $d(x_t^\varepsilon)^2$.

$$E((x_t^\varepsilon - x_t)^2) \leqq 2E\int_s^t |x_t^\varepsilon - x_t| \left| \int yu_t(y)\mu(x_t^\varepsilon, dy) - \int yu_t(y)\mu(x_t, dy) \right| dt$$

$$+ \varepsilon E \int_s^t \int y^2 u_t(y)\mu(x_t^\varepsilon, dy)$$

$$\leqq c_1 \int_s^t E((x_t^\varepsilon - x_t)^2) \int |y| u_t(y)\mu(x_t, dy)$$

$$+ c_2\varepsilon \int_s^t \int y^2 u_t(y)\mu(x_t, dy).$$

Using then

$$\int |y| u_t(y)\mu(x_t, dy) \leqq c_3 k^{u_t}(x_t) + c_4,$$

$$\int |y|^2 u_t(y)\mu(x_t, dy) \leqq c_3 k^{u_t}(x_t) + c_4,$$

let $E((x_t^\varepsilon - x_t)^2) = f_t$, $c_3 k^{u_t}(x_t) + c_4 = g_t$. We get

$$f_t \leqq c_1 \int_s^t f_\theta g_\theta \, d\theta + c_2\varepsilon \int_s^t g_\theta \, d\theta.$$

Gronwall's inequality implies

$$f_t \leqq c_2\varepsilon \left(\int_s^t g_\theta \, d\theta \right) \exp\left(c_1 \int_s^t g_\theta \, d\theta \right) \leqq c_5\varepsilon \quad \text{for } t \leqq T,$$

(4.6)

$$E((x_t^\varepsilon - x_t)^2) \leqq c_5\varepsilon, \qquad t \leqq T.$$

By changing the value of u_t after τ_0, we may assume for some $\tau_0 < T_0 < T$, $d(x_{T_0}, \Omega) = r > 0$. Using the martingale property as in the proof above gives

$$J^\varepsilon(s, x) \leqq E\left[\int_s^{T_0} \tilde{k}_\varepsilon^u(Y_t^\varepsilon)\, dt + J^\varepsilon(T_0, Y_{T_0}^\varepsilon) \right],$$

$$Y_t^\varepsilon = \begin{cases} x_t^\varepsilon & \text{if } t \leqq \tau^\varepsilon, \\ x_\tau^\varepsilon & \text{if } t > \tau^\varepsilon, \end{cases}$$

$$\tilde{k}_\varepsilon^u(x) = \begin{cases} k_\varepsilon^u(x) & \text{if } x \in \Omega, \\ 0 & \text{if } x \notin \Omega. \end{cases}$$

Equation (4.6) implies $E_{s,x}^\varepsilon[\tau > T_0] \to 0$ as $\varepsilon \to 0$. Then for fixed d, J^ε being bounded by $-\log d$ and $J^\varepsilon(s, y) = 0$ if $y \notin \Omega$, implies $E_{s,x}^\varepsilon[J^\varepsilon(T_0, Y_{T_0}^\varepsilon)] \to 0$ as $\varepsilon \to 0$.

$$\overline{\lim}\, J^\varepsilon(s, x) \leqq \overline{\lim}\, E_{s,x}^\varepsilon \left[\int_s^{T_0} \tilde{k}^u(Y_t^\varepsilon)\, dt \right] = \overline{\lim} \int_s^{T_0} E[\tilde{k}^u(Y_t^\varepsilon)]\, dt \leqq \int_s^{T_0} k^u(x_t)\, dt.$$

The last inequality is due to $E[\tilde{k}^u(Y_t^\varepsilon)] \leqq E[k^u(x_t^\varepsilon)] \to k^u(x_t)$. Since T_0 can be chosen arbitrarily close to τ_0, we have

$$\overline{\lim}\, J^\varepsilon(s, x) \leqq \int_s^{\tau_0} k^u(x_t)\, dt \leqq J(s, x) + \delta \quad \forall \delta > 0,$$

i.e.

$$\overline{\lim}\, J^\varepsilon(s, x) \leqq J(s, x).$$

Remark 8. We can show $\overline{\lim}_{\varepsilon\to 0} J^\varepsilon(s, x, d) \leqq J(s, x, d)$ holds for all $d > 0$., In fact, if instead of (4.5) we have

$$J(s, x) \geqq \int_s^T k^u(x_t)\, dt - \log d - \delta, \qquad \tau_0 > T,$$

then using the same notation as in the proof, we have

$$J^\varepsilon(s, x) \leqq E\left[\int_s^T \tilde{k}_\varepsilon^u(Y_t^\varepsilon)\, dt + J^\varepsilon(T, Y_T^\varepsilon)\right] \leqq E\left[\int_s^T \tilde{k}_\varepsilon^u(Y_t^\varepsilon)\, dt - \log d\right],$$

$$\overline{\lim} J^\varepsilon(s, x) \leqq \overline{\lim} E\int_s^T \tilde{k}_\varepsilon^u(Y_t^\varepsilon)\, dt - \log d$$

$$\leqq \int_s^T k^u(x_t)\, dt - \log d \leqq J(s, x) + \delta \quad \forall \delta > 0.$$

This implies $\lim J^\varepsilon(s, x) \leqq J(s, x)$.

Next in order to prove $\underline{\lim} J^\varepsilon(s, x) \geqq J(s, x)$, we use the following procedure. Let $\Omega_n \subset \Omega_{n+1} \subset \cdots \subset \Omega$ be a family of domain with smooth boundary such that $d(\Omega_n, \partial\Omega) = \delta_n > 0$, $\cup\Omega_n = \Omega$ where $\partial\Omega$ = boundary of Ω. Define $J_n(s, x)$

$$J_n(s, x) = \inf_{dx_t/dt = V(u_t, x_t)} \left\{\int_s^{\tau_n \wedge T} k^{u_t}(x_t)\, dt + W_n\right\},$$

$$W_n(y) = \begin{cases} 0, & y \notin \Omega_n, \\ W(y), & y \in \Omega_n, \end{cases}$$

τ_n = exit time of x_t from Ω_n.

We complete the proof of $\underline{\lim} J^\varepsilon(s, x) \geqq J(s, x)$ by showing the following two results.

LEMMA 12. $\underline{\lim} J^\varepsilon(s, x) \geqq J_n(s, x) \forall n$.

LEMMA 13. $\underline{\lim}_{n\to\infty} J_n(s, x) = J(s, x)$.

Proof of Lemma 12. Taking $u_t^\varepsilon(x) = \exp(J^\varepsilon(t, x)/\varepsilon)$ (cf. proof of Lemma 7), then

$$J^\varepsilon(s, x) = E_{s,x}^\varepsilon\left[\int_s^{\tau\wedge T} k^u(x_t)\, dt + W(x_{\tau\wedge T})\right].$$

Here we omit ε in u^ε if we fix ε in the discussion. Define the stopping time θ_k

$$\theta_k = \inf\left\{t \leqq \tau \wedge T; \int_s^t k^u(x_t)\, dt \geqq k\right\},$$

(4.7)

$$M = \sup_{\varepsilon > 0} J^\varepsilon(s, x) < \infty;$$

then we have

$$P_{s,x}^\varepsilon[\tau \wedge T > \theta_k] \leqq \frac{M}{k}, \tag{4.8}$$

by using

$$kP_{s,x}^\varepsilon[\tau \wedge T > \theta_k] \leqq E_{s,x}^\varepsilon\left[\int_s^{\theta_k} k^u(x_t)\, dt;\ \tau \wedge T > \theta_k\right]$$

$$\leqq E_{s,x}^\varepsilon\left[\int_s^{\tau\wedge T} k^u(x_t)\, dt\right] \leqq J^\varepsilon(s, x) \leqq M.$$

Define two processes x_t, $\hat{x}_t$

$$\begin{aligned}\frac{dx_t}{dt} &= \int \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} y\mu(x_t, dy), \qquad x_s = x,\\ \frac{d\hat{x}_t}{dt} &= \int \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} y\mu(x_t^\varepsilon, dy), \qquad \hat{x}_s = x.\end{aligned} \tag{4.9}$$

Here we see $dx_t/dt = V(\tilde{u}_t, x_t)$, $u_t(y) = u_t(x_t^\varepsilon + \varepsilon y)/u_t(x_t^\varepsilon)$. We want to show that x_t, $\hat{x}_t$, x_t^ε are close to each other. First apply the martingale property to get

$$d(x_t^\varepsilon - \hat{x}_t)^2 = \varepsilon \int \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} y^2 \mu(x_t^\varepsilon, dy) + dm, g, \qquad t \leqq T.$$

We then use the martingale inequality

$$\begin{aligned}P\left[\sup_{t \leqq \theta_k} \left|(\hat{x}_t - x_t^\varepsilon)^2 - \varepsilon \int_s^t \int \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} y^2 \mu(x_t^\varepsilon, dy)\right| \geqq \delta\right]\\ \leqq \frac{1}{\delta} E\left[(\hat{x}_{\theta_k} - x_{\theta_k}^\varepsilon)^2 + \varepsilon \int_s^{\theta_k} \int \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} y^2 \mu(x_t^\varepsilon, dy)\right].\end{aligned} \tag{4.10}$$

As was explained in Lemma 10,

$$\int \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} y^2 \mu(x_t^\varepsilon, dy) \leqq c_1 k_\varepsilon^u(x_t^\varepsilon) + c_2.$$

Putting this into (4.10) gives

$$P[\sup_{t \leqq \theta_k} (\hat{x}_t - x_t^\varepsilon)^2 \geqq \delta + (c_1 k + c_2)\varepsilon] \leqq 2\frac{1}{\delta}(c_1 k + c_2)\varepsilon. \tag{4.11}$$

Apply Lemma 10 again to get

$$|\hat{x}_t - x_t| \leqq c \int_s^t \int \frac{u_\theta(x_\theta^\varepsilon + \varepsilon y)}{u_\theta(x_\theta^\varepsilon)} |y| \mu(x_\theta^\varepsilon, dy) |x_\theta^\varepsilon - x_\theta| \, d\theta.$$

Using the same argument as in proving Lemma 10 (4.2), gives

$$\int \frac{u_\theta(x_\theta^\varepsilon + \varepsilon y)}{u_\theta(x_\theta^\varepsilon)} |y| \mu(x_\theta^\varepsilon, dy) \leqq c_1 k_\varepsilon^u(x_\theta^\varepsilon) + c_2.$$

Using also $|\hat{x}_t - x_t| \leqq |x_t^\varepsilon - x_t| + |\hat{x}_t - x_t^\varepsilon|$, we have

$$|x_t^\varepsilon - x_t| \leqq c \int_s^t (c_1 k_\varepsilon^u(x_\theta^\varepsilon) + c_2) |x_\theta^\varepsilon - x_\theta| \, d\theta + |\hat{x}_t - x_t^\varepsilon|,$$

which takes the familiar form

$$\begin{aligned}f_t &\leqq c \int_s^t f_\theta g_\theta d_\theta + \rho, \qquad t \leqq \theta_k,\\ \rho &= \sup_{t \leqq \theta_k} |\hat{x}_t - x_t^\varepsilon|.\end{aligned}$$

We apply Gronwall's inequality to obtain

$$|x_t^\varepsilon - x_t| \leqq \rho \exp\left(c \int_s^{\theta_k} (c_1 k_\varepsilon^u(x_\theta^\varepsilon) + c_2)\, d\theta\right) \leqq \rho \exp(c_3 k + c_4). \tag{4.12}$$

For given $0 < r < 1$, if we choose k to be large enough such that $M/k < r$, then (4.8) implies

$$P_{s,x}[\tau \wedge T > \theta_k] \leqq \frac{M}{k} < r. \tag{4.13}$$

Fix n, and choose δ and ε_0 such that

$$(2\delta)^{1/2} \exp(c_3 k + c_4) = r\delta_n, \qquad \delta_n = d(\Omega_n, \partial\Omega),$$

$$(c_1 k + c_2)\varepsilon < \delta, \qquad 2\frac{1}{\delta}(c_1 k + c_2)\varepsilon < r \quad \text{for } \varepsilon \leqq \varepsilon_0.$$

Then (4.11), (4.12) imply

$$\begin{aligned} &P[\sup_{t \leqq \theta_k} |\hat{x}_t - x_t^\varepsilon| \geqq (2\delta)^{1/2}] \leqq r, \\ &\sup_{t \leqq \theta_k} |x_t^\varepsilon - x_t| \leqq r\delta_n \quad \text{if } \sup_{t \leqq \theta_k} |\hat{x}_t - x_t^\varepsilon| \leqq (2\delta)^{1/2}. \end{aligned} \tag{4.14}$$

Together with (4.8), for $\varepsilon \leqq \varepsilon_0$

$$P(\tau \wedge T = \theta_k, \sup_{t \leqq \theta_k} |\hat{x}_t - x_t^\varepsilon| < (2\delta)^{1/2}) \geqq 1 - 2r,$$

$$\tau \wedge T = \theta_k, \sup_{t \leqq \theta_k} |\hat{x}_t - x_t^\varepsilon| < (2\delta)^{1/2} \quad \text{implies } \sup_{t \leqq \theta_k} |x_t^\varepsilon - x_t| < r\delta_n.$$

We will say that condition (*) holds if $\tau \wedge T = \theta_k$, $\sup_{t \leqq \theta_k} |\hat{x}_t - x_t^\varepsilon| < 2\delta^{1/2}$. Then (*) will imply $\tau_n \wedge T \leqq \tau \wedge T$,

$$\begin{aligned} J_n(s, x) &\leqq \int_s^{\tau_n \wedge T} k^{\tilde{u}}(x_t)\, dt + W(x_{\tau_n \wedge T}) \\ &\leqq \begin{cases} \displaystyle\int_s^{\tau_n} k^{\tilde{u}}(x_t)\, dt & \text{if } \tau < T \text{ and (*)}, \\ \displaystyle\int_s^{T} k^{\tilde{u}}(x_t)\, dt + W(x_T^\varepsilon) & \text{if } \tau > T \text{ and (*)}. \end{cases} \end{aligned}$$

Note that $P(\tau = T) = 0$ and

$$\begin{aligned} k^{\tilde{u}}(x_t) &= \int \left(\frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} \log \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} - \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} + 1\right) \mu(x_t, dt) \\ &\leqq (1 + c|x_t - x_t^\varepsilon|) \int \left(\frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} \log \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} - \frac{u_t(x_t^\varepsilon + \varepsilon y)}{u_t(x_t^\varepsilon)} + 1\right) \mu(x_t^\varepsilon, dy) \\ &= (1 + c|x_t - x_t^\varepsilon|) k^u(x_t^\varepsilon). \end{aligned}$$

Therefore

$$J_n(s,x) \leqq \begin{cases} (1+cr\delta_n)\left\{\int_s^{\tau\wedge T} k^u(x_t^\varepsilon)\,dt + W(x_{\tau\wedge T}^\varepsilon)\right\} & \text{if (*) holds,} \quad \varepsilon \leqq \varepsilon_0. \\ M & \text{otherwise,} \end{cases}$$

$$J_n(s,x) \leqq (1+cr\delta_n)E_{s,x}\left[\int_s^{\tau\wedge T} k^u(x_t^\varepsilon)\,dt + W(x_{\tau\wedge T}^\varepsilon)\right] + MP_{s,x}\{(*) \text{ does not hold}\}$$

$$\leqq (1+cr\delta_n)J^\varepsilon(s,x) + 2rM, \qquad \varepsilon \leqq \varepsilon_0,$$

$$J_n(s,x) \leqq (1+cr\delta_n)\,\underline{\lim}\, J^\varepsilon(s,x) + 2rM \quad \forall 0 < r < 1.$$

Letting $r \to 0$, we get $\underline{\lim}_{\varepsilon\to 0} J^\varepsilon(s,x) \geqq J_n(s,x)$ which ends the proof of Lemma 12.

In order to prove Lemma 13, we need some properties of $F(c,x)$ introduced in the proof of Theorem 3.

$$F(c,x) = \int e^{c\cdot y}\mu(x,dy).$$

As was pointed out there,

$$k(v,x) = F^*(v,x) + a(x),$$

$$F^*(v,x) = \sup\{c\cdot v - F(c,x)\}.$$

In the following, ∇ denotes the gradient with respect to c.

LEMMA 14. *Fix $r>0$; there is $K>0$ and $M>0$ such that*

$$\langle c, \nabla F(c,x)\rangle \geqq K|c|F(c,x) \quad \text{if } |x| \leqq r, |c| \geqq M.$$

Proof. Suppose not. Then there are $c_n \varepsilon R^n$, $x_n \varepsilon R^n$, $\varepsilon_n > 0$ such that

$$|c_n| \to \infty, \quad \varepsilon_n \to 0, \quad |x_n| \leqq r,$$

$$\langle c_n, \nabla F(c_n, x_n)\rangle \leqq \varepsilon_n |c_n| F(c_n, x_n),$$

i.e.

$$\int c_n \cdot y \exp(c_n\cdot y)\mu(x_n,dy) \leqq \varepsilon_n|c_n| \int \exp(c_n\cdot y)\mu(x_n,dy). \tag{4.15}$$

By using the fact that

$$\inf_{|c|=1} \int_{c\cdot y \geqq 0} (c\cdot y)\mu(x,dy) > 0,$$

it is easy to see that $\int (c_n\cdot y)e^{c_n\cdot y}\mu(x_n,dy) \to \infty$. Then both sides of (4.15) tend to ∞ as $n\to\infty$. Therefore we can assume

$$\int_{c_n\cdot y>0} (c_n\cdot y)\exp(c_n\cdot y)\mu(x_n,dy) \leqq \varepsilon_n|c_n| \int_{c_n\cdot y>0} \exp(c_n\cdot y)\mu(x_n,dy).$$

From this, for any $k>0$

$$\int_{c_n\cdot y \geqq k\varepsilon_n|c_n|} ((c_n\cdot y) - \varepsilon_n|c_n|)\exp(c_n\cdot y)\mu(x_n,dy) \leqq \varepsilon_n|c_n| \int_{0<c_n\cdot y<k\varepsilon_n|c_n|} \exp(c_n\cdot y)\mu(x_n,dy).$$

Then

$$(k-1)\varepsilon_n|c_n| \exp(k\varepsilon_n|c_n|)\mu(x_n, \{y; c_n \cdot y > k\varepsilon_n|c_n|\})$$

$$\leqq \varepsilon_n|c_n| \exp(k\varepsilon_n|c_n|)\mu(x_n, \{y; 0 < c_n \cdot y < k\varepsilon_n|c_n|\})$$

or

$$(k-1)\mu\left(x_n, \left\{y; \frac{c_n}{|c_n|} \cdot y > k\varepsilon_n\right\}\right) \leqq \mu\left(x_n, \left\{y: 0 \leqq \frac{c_n}{|c_n|} \cdot y < k\varepsilon_n\right\}\right). \tag{4.16}$$

Now we can assume $c_n/|c_n| \to c_0$ with $|c_0| = 1$ and $x_n \to x_0$ with $|x_0| \leqq r$. Then (4.16) implies

$$(k-1)\mu(x_0; \{y; c_0 \cdot y > 0\}) \leqq \mu(x_0, \{y; 0 = c_0 \cdot y\}) \quad \forall k > 1,$$

i.e., $\mu(x_0: \{y; c_0 \cdot y > 0\}) = 0$, which is a contradiction to assumption A2.

LEMMA 15. Given $r > 0$, then for any $\varepsilon > 0$ there is $R > 0$ such that $|x| \leqq r$, $|v_1 - v_2| \leqq 1$ and $|v_1| \geqq R$ imply $F^*(v_1, x) \leqq (1+\varepsilon)F^*(v_2, x)$.

Proof. Take M, K as in Lemma 14, and write $F(c, x)$ as $F(c)$.

$$\begin{aligned} F^*(v_1) &= \sup(\langle v_1, c\rangle - F(c)) \\ &= \langle v_1, c_1\rangle - F(c_1), \qquad v_1 = \nabla F(c_1) \text{ as in Lemma 11} \\ &= \langle v_1 - v_2, c_1\rangle + \langle v_2, c_1\rangle - F(c_1) \\ &\leqq |c_1| + F^*(v_2) \quad \text{since } |v_1 - v_2| \leqq 1. \end{aligned}$$

By Lemma 14

$$\langle c, \nabla F(c)\rangle \geqq K|c|F(c), \qquad |c| \geqq M \tag{4.17}$$

since $v_1 = \nabla F(c_1) = \int e^{c_1 \cdot y} y\mu(dy)$, there is R such that

$$|v_1| > R \text{ implies } |c_1| > M. \tag{4.18}$$

Using (4.17), (4.18) and $F^*(v_1) = \langle c_1, \nabla F(c_1)\rangle - F(c_1)$,

$$\begin{aligned} F^*(v_1) &\geqq \langle c_1, \nabla F(c_1)\rangle - \frac{1}{K|c_1|}\langle c_1, \nabla F(c_1)\rangle \\ &= \left(1 - \frac{1}{K|c_1|}\right)\langle c_1, \nabla F(c_1)\rangle \geqq \left(1 - \frac{1}{K|c_1|}\right) KF(c_1)|c_1|. \end{aligned}$$

Since $F(c_1) \to \infty$ as $|c_1| \to \infty$, we can choose M to be large enough such that $(1 - 1/K|c_1|)KF(c_1) > 1/\delta$ for a fixed small number which will be chosen to be $\varepsilon/(1+\varepsilon)$. Then $|c_1| \leqq \delta F^*(v_1)$. After combining this with $F^*(v_1) \leqq |c_1| + F^*(v_2)$ we get $(1-\delta)F^*(v_1) \leqq F^*(v_2)$, i.e., $F^*(v_1) \leqq (1+\varepsilon)F^*(v_2)$ if we take $\delta = \varepsilon/(1+\varepsilon)$.

LEMMA 16. *Fix $r > 0$. Then for any $\varepsilon > 0$, there is $\delta > 0$ such that*

$$a(x) + F^*(v_1, x) \leqq (1+\varepsilon)(a(x) + F^*(v_2, x)) + \varepsilon \quad \text{if } |v_1 - v_2| < \delta, \quad |x| \leqq r.$$

Proof. First we choose R as in Lemma 15, $\delta < 1$. Then obviously the inequality holds for $|v_1| \geqq R$, $|x| \leqq r$. As for $|v_1| \leqq R$, the continuity of F^* implies $F^*(v_1, x) \leqq F^*(v_2, x) + \varepsilon$ if $|v_1 - v_2| \leqq \delta$. Again this gives the above inequality.

Remark 9. We did not indicate the continuity of $F^*(v, x)$ with respect to the joint variable. But it is obvious once we note that $F(c, x)/|c| \to \infty$ as $|c| \to \infty$ uniformly for x in a bounded set and $f^*(v, x) = \sup\{\langle v, c\rangle - F(c, x)\}$.

Let $\beta_n = \sup\{|x-y|;\ x\in\partial\Omega,\ y\in\partial\Omega_n\}$, $\delta_n = \inf\{|x-y|,\ x\in\partial\Omega,\ y\in\partial\Omega_n\}$. Then $\beta_n,\ \delta_n \to 0$.

LEMMA 17. *If $T-s > d(x,\partial\Omega)$, $x\in\Omega$, then $J(s,x) < cd(x,\partial\Omega)$ where c is independent of s, x.*

Proof. $x\in\Omega$, $x^*\in\partial\Omega$ such that $|x-x^*| = d(x,\partial\Omega)$. Let $v = (x-x^*)/(|x-x^*|)$, $dx(t)/dt = v$, $x(s) = x$. Then $x(t_0) = x^*$ where $t_0 - s = |x-x^*| = d(x,\partial\Omega) < T-s$, i.e., $t_0 < T$ and t_0 is the exit time of $x(t)$ from Ω. Now it is easy to see that

$$J(s,x) \leqq \int_s^{t_0} F^*(v, x_t) + a(x_t)\,dt \leqq c(t_0 - s) = c|x-x^*|.$$

Finally it is time to prove Lemma 13, i.e. $\lim_{n\to\infty} J_n(s,x) = J(s,x)$.

Proof of Lemma 13. Since $J_n(s,x)$ increases with respect to n and is smaller than $J(s,x)$, therefore there exists a limit $A = \lim J_n(s,x) \leqq J(s,x)$. It remains to show that $A \geqq J(s,x)$.

Let $\varepsilon > 0$; for each n there is a $v_n(t) = v(t)$, bounded in $[s,T]$, such that

$$\int^{\tau_n \wedge T} (F^*(\dot{x}_t, x_t) + a(x_t)\,dt + W(x_{\tau_n\wedge T}) < A + \varepsilon, \tag{4.19}$$

$$\dot{x}_t = \frac{dx_t}{dt} = v(t).$$

We consider the following cases of τ_n.

(i) Assume $\tau_n > T$ for some n. Then

$$J(s,x) \leqq \int_s^T (F^*(\dot{x}_n(t), x_n(t)) + a(x_n(t)))\,dt + W(x_n(T)) < A + \varepsilon.$$

(ii) Assume $\tau_n \leqq T$ for all n. Let $\alpha > 1$. With this α we define $\tilde{x}(t) = x(\alpha t)$. Then $\tilde{x}(t_n)\in\partial\Omega_n$ where $t_n = \tau_n/\alpha$ and hence $J(t_n, \tilde{x}(t_n)) \leqq c\beta_n$ by Lemma 17. Therefore, we obtain

$$J(t_n, \tilde{x}(t_n)) \to 0 \quad \text{as } n\to\infty. \tag{4.20}$$

Now considering

$$J(s,x) \leqq \int_s^{t_n} (F^*(\dot{x}(t), \tilde{x}(t)) + a(\tilde{x}(t)))\,dt + J(t_n, \tilde{x}(t_n))$$

$$= \frac{1}{\alpha}\int_s^{\tau_n} (F^*(\alpha\dot{x}(t), x(t)) + a(x(t)))\,dt + J(t_n, \tilde{x}(t_n)),$$

and applying (4.19) and the following inequality:

$$F^*(\alpha v, x) + a(x) \leqq c(\alpha) + (1 + c(\alpha))(F^*(v,x) + a(x)), \tag{4.21}$$

we can get $J(s,x) \leqq A + 2\varepsilon$ by first choosing $\alpha > 1$ but close enough to 1, and then by letting $n\to\infty$. Here we only note that the proof of (4.21) is similar to that of Lemma 15. Finally, we conclude from (i) and (ii) $J(s,x) \leqq A$, i.e., $J(s,x) = A$.

We can now combine Theorem 4, Lemma 12, and Lemma 13 to get

THEOREM 5. $\lim_{\varepsilon\to\infty} J^\varepsilon(s,x,d) = J(s,x,d)$ *for every* $s < T$, $x\in\Omega$, $d > 0$.

THEOREM 6. *For any* $t > 0$, $x\in\Omega$, *there is* $\varepsilon_0 > 0$ *such that*

$$\phi^\varepsilon(t,x) > 0 \quad \textit{for } \varepsilon < \varepsilon_0.$$

Moreover $\lim\{-\varepsilon \log \phi^{\varepsilon}(t, x)\}$ *exists and is equal to*

$$\inf_{\tau<t} \int_0^{\tau} k^{u_t}(x_t)\, dt = \inf_{\tau<t} \int_0^{\tau} k(\dot{x}_t, x_t)\, dt.$$

Proof. As was indicated at the beginning of § 3, Theorem 5 and Lemma 6 are sufficient for getting the result if we note that $J(T-t, x, d) = \inf_{\tau<t} \int_0^{\tau} k(\dot{x}_t, x_t)\, dt$ for d small enough.

Acknowledgments. Most of the material is taken from the second chapter of my thesis *Stochastic control and its application* (Department of Mathematics, Brown University, 1982). It is a pleasure to have this opportunity to thank my advisor, Wendell H. Fleming, for his constant support, encouragement and guidance in this interesting field. Without these, it would not be possible for my work to be successful.

REFERENCES

[1] W. H. FLEMING AND R. W. RISHEL (1975), *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York.

[2] W. H. FLEMING (1977), *Inclusion probability and optimal stochastic control*, IRIA Seminars Review.

[3] ——— (1978), *Exit probabilities and stochastic control*, Appl. Math. Optim., 4, pp. 329-346.

[4] W. H. FLEMING AND CHUN-PING TSAI (1981), *Optimal exit probabilities and differential games*, Appl. Math. Optim., 7, pp. 253-282.

[5] I. I. GIKHMAN AND A. V. SKOROKHOD (1970), *The Theory of Stochastic Processes* II. Springer-Verlag, New York.

[6] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ.

[7] A. D. VENTSEL' (1976), *Rough limit theorems on large deviations for Markov stochastic processes* I, II, in Theory Prob. Appl., 31 (1976), pp. 227-242, 499-512.

[8] ONESIMO HERNANDEZ-LERMA (1977), *Stability of differential equations with Markov parameter and exit problem*, Thesis, Applied Mathematics Dept., Brown Univ., Providence, RI.

[9] R. AZENCOTT, *Grandes déviations et applications*, Lecture Notes in Mathematics 774, Springer-Verlag, New York, 1978.

[10] W. H. FLEMING, *Logarithmic Transformations and Stochastic Control*, Lecture Notes in Control and Information Sciences 42, Springer-Verlag, New York, 1982.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

A SURVEY OF SOME RESULTS IN STOCHASTIC ADAPTIVE CONTROL*

P. R. KUMAR†

Abstract. Some results in discrete-time stochastic adaptive control are surveyed. The survey divides itself into two parts—Bayesian and non-Bayesian adaptive control. In the former area, the problems of converting an incompletely observed system into a completely observed one, multi-armed bandit processes, Bayesian adaptive control of Markov chains and Bayesian adaptive control of linear systems are exposed and surveyed. In the latter area, non-Bayesian adaptive control of Markov chains and the self-tuning regulator are dealt with. Proofs are given, where appropriate, to illustrate the methods involved.

Key words. adaptive control, stochastic adaptive control, Bayesian adaptive control, bandit problems, Bayesian control of Markov chains, non-Bayesian adaptive control, adaptive control of linear systems, self-tuning regulators, self-optimizing systems

CONTENTS

1. Introduction.

We shall be concerned with the control of discrete time stochastic systems. The distinguishing feature of the class of problems we address is that the system under control is unknown. In spite of this, we desire to design control laws which will result in adequate behaviour of the system. The adequacy of the system behaviour will, in turn, be mainly measured by a given cost criterion.

We shall *not* deal with the problem of adaptive control of deterministic systems. Rather, random or noisy behaviour will be an essential feature of the systems we study. We shall also not deal with *continuous* time stochastic systems.

* Received by the editors January 16, 1984, and in revised form June 18, 1984. This is a special expository paper written at the invitation of the editors. This research was supported in part by the National Science Foundation under grant ECS-8304435, and in part by the Army Research Office under contract DAAG29-84-K0005.

† Department of Electrical Engineering and Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801.

The survey divides itself naturally into two parts—Bayesian adaptive control problems (BACP's) and non-Bayesian adaptive control problems (NACP's). We shall suppose, throughout, that the system behaviour depends on a parameter θ, and it is the fact that the *value* of θ is unknown to us which makes the *system unknown.* Now we can distinguish the essential difference between the Bayesian and non-Bayesian formulations. In the former, we are given a probability distribution $q_0(d\theta)$ for the value of the unknown parameter θ. In the latter, we are *not* given any such initial prior distribution. Rather, we are only given a *set* Θ and a *guarantee* that the unknown parameter θ is some element of Θ.

The *Bayesian* N-armed bandit problem and the "dual" control problem are examples of BACP's. The self-tuning regulator and the non-Bayesian adaptive control problem for Markov chains are examples of NACP's. All these topics and others are surveyed in what follows. In §§2–5, we survey BACP's and in §§6–8, we survey NACP's. Each section is addressed to one type of problem.

Some comments on the goal and style of this paper are in order. It has been a primary goal in writing this paper to produce an exposé of the field which will cover, in an understandable way, some of the problems, ideas and mathematical techniques of this field. To achieve this, *proofs* of several results are provided throughout the paper. (However a theorem or proof attributed to an author is not necessarily an exact replica of the original; some modifications have sometimes been made in the interests of brevity, clarity, etc.). No attempt has been made to provide an exhaustive list of *all* papers in the field. Such an approach, it was felt, would only tend to make the narrative very disjointed. The emphasis, instead, is on the coverage of *ideas.* Apologies are thus owed to several authors for such omissions. Lastly, we have made no real attempt at attribution of results to authors.

2. Formulation and reduction of the Bayesian adaptive control problem. The Bayesian approach to adaptive control is this. There is a stochastic dynamic system which depends on a parameter θ. We are given an initial probability distribution (a *prior* distribution) $q_0(d\theta)$ for the unknown parameter. At each time instant $t = 1, 2, 3, \cdots$ we obtain a noisy (or, as a special case, perfect) observation y_t of the state x_t of the system. Our goal is to minimize some given cost criterion, such as $E\sum_0^\infty \beta^t c(x_t, u_t)$. Here u_t is the control input that we apply to the system on the basis of the observations $(u_0, y_1, u_1, y_2, \cdots, y_{t-1}, u_{t-1}, y_t)$ made on the system. β with $0 < \beta \leqq 1$ is the discount factor. The heart of the problem is that the expectation in the cost criterion $E\sum_0^\infty \beta^t c(x_t, u_t)$ is taken not only with respect to the random behaviour of the stochastic system, but *also* with respect to the random choice of θ according to $q_0(d\theta)$.

The standard approach to solving this problem is to transform the BACP into an equivalent *dynamic programming* problem and then to bring to bear the well-developed theory of dynamic programming. The "*state*" of this new dynamic programming problem, which we shall refer to as the *hyperstate*, will be the conditional probability distribution of the *old* state *and* the parameter value given the observations made.

However, in achieving this transformation, some delicate measurability questions must be resolved in order to ensure that one obtains a mathematically well formulated dynamic programming problem. It is one of the accomplishments of the past two decades that this problem of converting a partial or imperfect observations stochastic control problem (for this is what a BACP is) into an equivalent dynamic programming problem has been more or less satisfactorily resolved.

The problem to be examined below is, in some sense, purely technical, and it is somewhat unfortunate that in order to preserve logical continuity this purely technical problem is the first issue examined in detail in this survey. The reader without an appetite for technical issues may wish to gloss over this section and proceed to § 3 and the rest of this paper where, in contrast to this section, purely *structural* issues are addressed.

The development we follow is due to Bertsekas and Shreve [1] and consists of the following:

i) X, Y, U, Θ are Borel spaces (i.e. homeomorphic to a Borel subset of some complete separable metric space) which are, respectively, the state space, observation space, control set and parameter set.

ii) $q_0(dx_0, d\theta)$ is a given probability distribution for the initial state x_0 and the parameter θ.

iii) $p(dx_{t+1}|x_t, u_t, \theta)$ is a Borel measurable stochastic kernel (i.e. $p(B|\cdot)$ is Borel measurable for every Borel set $B \subseteq X$) which specifies the probability distribution of the new state x_{t+1} given the previous state x_t, applied control u_t and the parameter value θ.

iv) $r(dy_t, x_t, u_{t-1}, \theta)$ is a Borel measurable stochastic kernel which specifies the probability distribution of the observation y_t given the state x_t and control u_{t-1}.

v) $c(x_t, u_t)$ is a lower semi-analytic function which is the one-step cost function. (This means $\{(x, u): c(x, u) < a\}$ is an analytic set for every a. A subset of A is analytic if it is the projection on A of a Borel subset of $A \times B$ where B is some uncountable Borel space.) $\beta \in (0, 1]$ is a discount factor. If $\beta = 1$, we assume that either $c \geqq 0$ always or $c \leqq 0$ always. c will always be assumed to be bounded.

vi) A policy π is a sequence $\pi = (\pi_0, \pi_1, \cdots)$ where each $\pi_t(du_t|q_0, u_0, y_1, u_1, \cdots, y_t)$ is a universally measurable stochastic kernel (i.e. $\pi(B|\cdot)$ is universally measurable for every Borel $B \subseteq U$. A function f is universally measurable if the inverse image of every Borel set is measurable with respect to the completion of every Borel measure). Let Π be the set of all such policies.

vii) For every $\pi \in \Pi$ and q_0, one can define the associated cost function $J(\pi, q_0) = E \sum_0^\infty \beta^t c(x_t, u_t)$. Let $J(q_0) = \inf_{\pi \in \Pi} J(\pi, q_0)$ be the *optimal* cost function.

The interpretation of this model is as follows. At time 0, the initial state x_0 and the parameter value θ are distributed according to $q_0(dx_0, d\theta)$. Based on q_0, the controller chooses a $u_0 \in U$ according to the distribution $\pi_0(du_0|q_0)$. Based on (x_0, u_0, θ) the new state x_1 is distributed according to $p(dx_1|x_0, u_0, \theta)$. Based on (x_1, u_0, θ) the controller receives an observation y_1 distributed according to $r(dy_1|x_1, u_0, \theta)$. Based on (q_0, u_0, y_1) the controller chooses u_1 according to $\pi_1(du_1|q_0, u_0, y_1)$ etc.

THEOREM 2.1 (Bertsekas and Shreve).

i) *There exists a Borel measurable stochastic kernel* $q_n(A|q_0, u_0, \cdots, y_n) = E_\pi(1_A(x_n, \theta)|q_0, u_0, \cdots, y_n)$ *for all* π, q_0 *and a.e.* $(u_0, \cdots, y_n)$. *Here* $E_\pi(\cdot)$ *is the conditional expectation under the probability measure induced by the policy* π *and* $1_A(\cdot)$ *is the indicator function of the set* A.

ii) *There exists a Borel measurable stochastic kernel* $\hat{p}(dq_{k+1}|q_k, u_k)$ *such that* $\hat{p}(Q|q_k, u_k) = E_\pi(1_Q(q_{k+1}(\cdot|q_0, \cdots, y_{k+1}))|q_0, q_k(\cdot|q_0, \cdots, y_k), u_k)$ *for every* π, q_0 *and a.e.* $(u_0, \cdots, y_k)$.

iii) *There exists a lower semi-analytic function* $\hat{c}(q_k, u_k)$ *such that* $\hat{c}(q_k, u_k) = E_\pi(c(x_k, u_k)|q_0, q_k(\cdot|q_0, \cdots, y_k), u_k)$ *for every* π, q_0.

iv) *Let* $\hat{\Pi}$ *be the set of all policies* $\hat{\pi}_k = (\hat{\pi}_0, \hat{\pi}_1, \cdots)$ *such that* $\hat{\pi}_k(du_k|q_k)$ *is a universally measurable kernel. For every policy* $\pi \in \hat{\Pi}$, *there is a policy* $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$

such that $\pi_k(du_k|q_0, u_0, \cdots, y_k) = \hat{\pi}_k(du_k|q_k(\cdot|q_0, u_0, \cdots, y_k))$. *Thus* $\hat{\Pi}$ *can be identified with a subset of* Π.

v) *If* $\hat{\pi} \in \hat{\Pi}$ *is nonrandomized, then the corresponding element* $\pi \in \Pi$ *with which it is identified is also nonrandomized.* (*A policy* $\pi = (\pi_0, \pi_1, \cdots)$ *is nonrandomized if each* π_k *is a degenerate probability distribution concentrated on just one point.*)

vi) *Let* $\hat{J}(\hat{\pi}, q_0) = E_{\hat{\pi}}(\sum_0^\infty \hat{c}(q_k, u_k)\beta^k|q_0 = q)$ *be the cost function of a dynamic programming problem with transition kernel* $\hat{p}$, *policy set* $\hat{\Pi}$, *and cost function* $\hat{c}$. *Then* $J(\hat{\pi}, q) = \hat{J}(\hat{\pi}, q)$ *for every* $\hat{\pi} \in \hat{\Pi}$ *and* q. (*In the left-hand side, by* $\hat{\pi}$ *we mean the element of* Π *with which it is identified.*)

vii) *For every* q *and* $\pi \in \Pi$, *there exists a* $\hat{\pi} \in \hat{\Pi}$ *such that* $J(\pi, q) = \hat{J}(\hat{\pi}, q)$.

viii) *Let* $\hat{J}(q) = \inf_{\hat{\pi} \in \hat{\Pi}} \hat{J}(\hat{\pi}, q)$. *Then* $\hat{J}(q) = J(q)$ *for every* q.

ix) *If* $\hat{\pi}^*$ *satisfies* $\hat{J}(\hat{\pi}^*, q) = \hat{J}(q)$ (*or* $\leqq \hat{J}(q) + \varepsilon$) *for every* q, *then* $J(\hat{\pi}^*, q) = J(q)$ (*or* $\leqq J(q) + \varepsilon$) *for every* q.

x) *For every* $\varepsilon > 0$, *there exists a* $\hat{\pi} \in \hat{\Pi}$ *which is nonrandomized, such that* $\hat{J}(\hat{\pi}, q) \leqq \hat{J}(q) + \varepsilon$ *for every* q.

Proof. See [1, Lemma 10.2 and Propositions 10.2, 10.3, 10.5].

The interpretation is as follows. $q_k(dx_k, d\theta|q_0, u_0, \cdots, y_k)$ is the conditional probability distribution of (x_k, θ) given the initial probability distribution $q_0(dx_0, d\theta)$ and the observation history $(u_0, y_1, \cdots, y_k)$. This will be the *hyperstate* of the new dynamic programming problem. $\hat{p}(dq_{k+1}|q_k, u_k)$ specifies the transition probability function for this new problem while $\hat{c}(q_k, u_k)$ is the new one-step cost function. Together, (i)-(iii) show that q_k, $\hat{p}$, $\hat{c}$ satisfy the assumptions to provide us a well-formulated dynamic programming problem which is of the type studied in [1]. The main point is that the hyperstate is completely observed.

The relationship between this new dynamic programming problem and the original BACP is provided through (iv)-(x). (iv) shows that any policy for the new dynamic programming problem can also be implemented on the BACP. This is clear in the sense that policies in $\hat{\Pi}$ are *only* allowed to depend on $(q_0, q_1, \cdots, u_0, u_1, \cdots)$ while policies in Π are allowed to depend on $(q_0, y_1, y_2, y_3, \cdots, u_0, u_1, \cdots)$ and each q_n is itself calculated on the basis of $(q_0, y_1, \cdots, y_n, u_0, \cdots, u_{n-1})$. Thus $\hat{\Pi}$ can be identified as a subset of Π. (vi) shows that the cost of using $\hat{\pi}$ in the new dynamic programming problem is the same as using it in the BACP. (vii) is a deep result which shows that for the BACP, every policy $\pi \in \Pi$ can be replaced by a policy $\hat{\pi} \in \hat{\Pi}$ which has the same cost. The advantage of this is that one may then restrict attention to policies $\hat{\pi} \in \hat{\Pi}$ which depend only on the hyperstate, thus rendering the hyperstate a "sufficient statistic" in some sense. (viii) and (ix) complete the process of identifying the BACP with the new dynamic programming problem. (viii) shows that both have the same optimal cost functions. (ix) shows that a policy $\hat{\pi} \in \hat{\Pi}$ optimal (or ε-optimal) for the new dynamic programming problem is also optimal (or ε-optimal) for the BACP. (x) is a consequence of our allowing universally measurable policies and shows that there exists an ε-optimal nonrandomized policy for the new dynamic programming problem (and therefore also for the BACP). A development similar to the above can also be given for the finite horizon case, see [1].

For other treatments of the conversion of an imperfectly observed problem into a completely observed dynamic programming problem, the reader is referred to Bellman [2], Dynkin [3], Aoki [4], Åström [5], Shiryaev [6], Striebel [7], [8], Hinderer [9], Sawaragi and Yoshikawa [10], Rhenius [11], Martin [12], Rieder [13] and van Hee [14]. [2]-[5] are early references featuring examples. [7] examines the concept(s) of a "sufficient statistic". [10] deals with countable state spaces and shows versions of (vi) and (vii) above. [11] considers the same issues for general Borel spaces. [12]-[14] also

achieve the conversion to a completely observed dynamic programming problem, making specific reference to the BACP. We have chosen here to follow the approach of [1] because its allowance of universally measurable policies gives the useful and reassuring property (x) above.

3. Bandit processes. As we have seen in § 2, a BACP can, like other imperfectly, observed problems, be replaced by an equivalent dynamic programming problem. However, since a BACP is a *special* type of an imperfect observations problem, the question naturally arises as to whether and to what extent we can take advantage of the special structure.

There is one class of problems, the multi-armed bandit problems, which is very special even *within* the class of BACP's and for this class of problems we can exploit this highly special structure to provide a rather deep theory.

Suppose that there are N (slot) machines. For machine i, $1 \leqq i \leqq N$, θ_i is the probability that, if it is played, it yields a reward of one unit, while $(1-\theta_i)$ is the probability that it yields no reward. The parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_N)$ is unknown. At every time $t = 0, 1, 2, \cdots$ *one* of the machines *has* to be played, and suppose that the reward accrued at time t is c_t. The goal is to choose which machine to play at each time $t = 0, 1, 2, \cdots$ so that $E \sum_0^\infty \beta^t c_t$ is maximized. $\beta \in (0, 1)$ is a discount factor.

We are not told of the value of $\boldsymbol{\theta}$, but instead, we are given a prior probability distribution of $q_0(d\boldsymbol{\theta}) = q_0^1(d\theta_1) q_0^2(d\theta_2) \cdots q_0^N(d\theta_N)$ for the value of $\boldsymbol{\theta}$. This immediately renders this a BACP.

As in the previous section, we therefore define the *hyperstate* $\mathbf{q} = (q^1, q^2, \cdots, q^N)$ where each $q^i(d\theta_i)$ is the conditional distribution of the value of θ_i. This gives a dynamic programming problem characterized by the following dynamic programming equation:

$$J(\mathbf{q}) = \max_{i \leqq i \leqq N} \left\{ \int_0^1 \theta_i q^i(d\theta_i) + \beta \left[\int_0^1 \theta_i q^i(d\theta_i) J(q^1, q^2, \cdots, q^{i-1}, S(q^i), q^{i+1}, \cdots, q^N) \right.\right.$$
$$\left.\left. + \int_0^1 (1-\theta_i) q^i(d\theta_i) J(q^1, q^2, \cdots, q^{i-1}, F(q^i), q^{i+1}, \cdots, q^N) \right] \right\},$$

where

$$S(q^i)(d\theta_i) = \theta_i q^i(d\theta_i) \left[\int_0^1 \eta q^i(d\eta) \right]^{-1}$$

and

$$F(q^i)(d\theta_i) := (1-\theta_i) q^i(d\theta_i) \left[\int_0^1 (1-\eta) q^i(d\eta) \right]^{-1}$$

The interpretation of this equation is straightforward. If machine i is played, then the probability that it yields a reward of one unit is $\int_0^1 \theta_i q^i(d\theta_i)$, and such an occurrence will cause us to revise the *posterior* distribution of the probability of success on machine i to $S(q^i)$ by Bayes' rule. Thus the state $\mathbf{q} = (q^1, \cdots, q^i, \cdots, q^N)$ changes to $(q^1, \cdots, S(q^i), \cdots, q^N)$. A similar analysis holds when a failure results from our playing machine i. $J(\mathbf{q})$, the optimal reward for the problem starting in the hyperstate $\mathbf{q}$, thus satisfies the given dynamic programming equation.

However, this dynamic programming equation, as written, is a rather formidable functional equation to solve. Through the work of Gittins and Jones [15], there is now a deep theory for this problem.

THEOREM 3.1.

i) (Gittins and Jones). *There is a real valued index function* $\gamma(\cdot)$ *with the property that if at any time* t, *the hyperstate is* $\mathbf{q}(t) = (q^1(t), \cdots, q^N(t))$, *then it is optimal to play any machine* i (*there may be more than one*) *for which* $\gamma(q^i(t))$ *is largest.*

ii) (Gittins and Glazebrook). *Suppose the state of a machine* (*it does not matter which one*) *is* q. *Let* $\{d_0, d_1, d_2, \cdots\}$ *be the stochastic process representing the successive random rewards obtainable by* <u>continuously</u> *playing this machine. Let* $\mathcal{F}_t := \sigma(d_0, d_1, \cdots, d_t)$ *be the* σ*-algebra generated by the reward sequence up to time* t. *Then*

$$\gamma(q) = \max_{\tau \geqq 1} \frac{E \sum_0^{\tau-1} \beta^t d_t}{E \sum_0^{\tau-1} \beta^t}$$

will suffice in (i), *where the maximum is taken over all stopping times* τ *of* $\{\mathcal{F}_t\}$.

iii) (Nash). *In* (ii), *let* $q(t)$ *be the hyperstate of the machine at time* t (*after* t *successive plays of the machine*). *Then a maximizing stopping time* τ *in* (ii) *is given by*

$$\tau = \inf\{t \geqq 1 \colon \gamma(q(t)) \leqq \gamma(q)\}.$$

iv) *Consider a problem for which there is* only *one machine which initially, as in* (ii), *has a hyperstate* q. *At any given time, one may either continue to play the machine, or collect* M *units of reward and quit forever. Let* $V(q, M)$ *be the optimal expected reward for this optimal stopping problem. Then*

$$V(q, M) = \max\left\{M; \int_0^1 \theta q(d\theta) + \beta\left[\int_0^1 \theta q(d\theta) V(S(q), M) + \int_0^1 (1-\theta) q(d\theta) V(F(q), M)\right]\right\}.$$

v) (Gittins and Jones). $\gamma(\cdot)$, *the index function of* (i) *and* (ii) *is given by*

$$\gamma(q) = \inf\{M(1-\beta) \colon V(q, M) = M\} = \sup\{M(1-\beta) \colon V(q, M) > M\}.$$

vi) (Whittle). *The optimal reward function* J *is related to* V *by*

$$J(\mathbf{q}) = (1-\beta)^{-1} - \int_0^{(1-\beta)^{-1}} \prod_{j=1}^{N} \frac{\partial}{\partial M} V(q^j, M)\, dM.$$

Proof. It is clear that $\gamma(q)$ in (ii) is well defined. It can be shown that this optimal stopping problem of (ii) actually has a *maximizing* stopping time, which demonstration we omit. Now we will show that the τ defined in (iii) attains the maximum in (ii). Let σ be *any* optimal stopping time (we have already assumed that there exists at least one such). Then consider the new stopping time $\mu := \tau \wedge \sigma$. Elementary calculations show that

$$\begin{aligned}
&\left(E \sum_0^{\mu-1} \beta^t d_t\right)\left(E \sum_0^{\mu-1} \beta^t\right)^{-1} - E\left(\sum_0^{\sigma-1} \beta^t d_t\right)\left(E \sum_0^{\sigma-1} \beta^t\right)^{-1} \\
&= \left\{E\left[-1(\tau < \sigma) \sum_\tau^{\sigma-1} d_t \beta^t\right] E \sum_0^{\sigma-1} \beta^t - E\left[-1(\tau < \sigma) \sum_\tau^{\sigma-1} \beta^t\right] E \sum_0^{\sigma-1} \beta^t d_t\right\} \\
&\quad \cdot \left\{E \sum_0^{\mu-1} \beta^t E \sum_0^{\sigma-1} \beta^t\right\}^{-1} \\
&= \left\{E\left[-1(\tau < \sigma) \sum_\tau^{\sigma-1} d_t \beta^t\right] - \gamma(q) E\left[-1(\tau < \sigma) \sum_\tau^{\sigma-1} \beta^t\right]\right\}\left\{E \sum_0^{\mu-1} \beta^t\right\}^{-1}
\end{aligned}$$

$$
\begin{aligned}
&= \left\{ \gamma(q) E\left[1(\tau<\sigma) \sum_{\tau}^{\sigma-1} \beta^t \right] - EE\left[1(\tau<\sigma) \sum_{\tau}^{\sigma-1} d_t \beta^t | \mathscr{F}_\tau \right] \right\} \left\{ E \sum_{0}^{\mu-1} \beta^t \right\}^{-1} \\
&\geqq \left\{ \gamma(q) E\left[1(\tau<\sigma) \sum_{\tau}^{\sigma-1} \beta^t \right] - E\left[1(\tau<\sigma)\gamma(q(\tau)) \sum_{\tau}^{\sigma-1} \beta^t \right] \right\} \left\{ E \sum_{0}^{\mu-1} \beta^t \right\}^{-1} \\
&\geqq \left\{ \gamma(q) E\left[1(\tau<\sigma) \sum_{\tau}^{\sigma-1} \beta^t \right] - E\left[1(\tau<\sigma)\gamma(q) \sum_{\tau}^{\sigma-1} \beta^t \right] \right\} \left\{ E \sum_{0}^{\mu-1} \beta^t \right\}^{-1} \\
&= 0.
\end{aligned}
$$

This shows that if σ is optimal, then $\sigma \wedge \tau = \mu$ is also optimal. We now show that $\mu = \tau$ a.s. Suppose not, then $P(\mu < \tau) > 0$, i.e. $P(\gamma(q(u)) > \gamma(q)) > 0$. For some $\varepsilon > 0$, therefore $P(\gamma(q(\mu)) > \gamma(q) + 2\varepsilon) > 0$. Define $\tilde{\Omega} := \{\omega\colon \gamma(q(\mu)) > \gamma(q) + 2\varepsilon\}$ and a new stopping time ξ by $\xi := \mu$ on $\tilde{\Omega}^c$, and $\xi := \mu + \alpha(q(\mu))$ on $\tilde{\Omega}$, where $\alpha(q(\mu))$ is ε-optimal starting from the state $q(\mu)$, i.e.

$$
E\left[\sum_{\mu}^{\mu+\alpha(q(\mu))-1} \beta^t d_t | \mathscr{F}_\mu \right] \left\{ E\left[\sum_{\mu}^{\mu+\alpha(q(\mu))-1} \beta^t | \mathscr{F}_\mu \right] \right\}^{-1} \geqq \gamma(q(\mu)) - \varepsilon \geqq \gamma(q) + \varepsilon \quad \text{on } \tilde{\Omega}.
$$

Another calculation shows that

$$
\begin{aligned}
\gamma(q) &\geqq E \sum_{0}^{\xi-1} \beta^t d_t \left\{ E \sum_{0}^{\xi-1} \beta^t \right\}^{-1} \\
&= \left\{ E \sum_{0}^{\mu-1} \beta^t d_t + E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t d_t \right] \right\} \left\{ E \sum_{0}^{\mu-1} \beta^t + E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t \right] \right\}^{-1} \\
&\geqq \left\{ \gamma(q) E \sum_{0}^{\mu-1} \beta^t + E\left[1(\xi>\mu)(\gamma(q)+\varepsilon) \sum_{\mu}^{\xi-1} \beta^t \right] \right\} \\
&\quad \cdot \left\{ E \sum_{0}^{\mu-1} \beta^t + E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t \right] \right\}^{-1} \\
&= \gamma(q) + \varepsilon E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t \right] \left\{ E \sum_{0}^{\mu-1} \beta^t + E\left[1(\xi>\mu) \sum_{\mu}^{\xi-1} \beta^t \right] \right\}^{-1} \\
&> \gamma(q),
\end{aligned}
$$

which is impossible. This shows that $\mu = \tau$ a.s., showing that τ is optimal. This proves (iii).

Now we follow essentially the interchange argument of Glazebrook [19], which in turn is essentially the argument in [15], in showing (i) and (ii). Let π be a particular policy satisfying (i) with γ defined as in (ii) and which breaks ties between competing maximizers by choosing the lexicographically smallest machine. Suppose that π chooses machine i at time 0. Then $\gamma(q^i(0)) = \max_j \gamma(q^j(0))$.

Consider $k \neq i$, and let $\hat{\pi}$ be the (nonstationary) policy which chooses machine k at time $t = 0$, and thereafter proceeds according to π. Note that under $\hat{\pi}$, machine k will be chosen at times $t = 0, \cdots, \tau_k - 1$ where $\tau_k := \min\{t \geqq 1\colon \gamma(q^k(\tau_k)) \leqq \gamma(q^i(0))\}$ if $k > i$ or $\tau_k := \min\{t \geqq 1\colon \gamma(q^k(\tau_k)) < \gamma(q^i(0))\}$ if $k < i$. Thereafter, under $\hat{\pi}$, machine i will be chosen at times $t = \tau_k, \cdots, \tau_k + \tau_i - 1$ where $\tau_k + \tau_i = \min\{t > \tau_k\colon \gamma(q^i(t)) \leqq \gamma(q^i(\tau_k)) = \gamma(q^i(0))\}$.

On the other hand, consider the policy $\hat{\hat{\pi}}$ which chooses machine i at times $t = 0, \cdots, \tau_i - 1$ and thereafter chooses machine k at times $t = \tau_i, \cdots, \tau_i + \tau_k - 1$. After

this, $\hat{\hat{\pi}}$ coincides with $\hat{\pi}$. Hence

$$
\begin{aligned}
&E_{\hat{\hat{\pi}}} \sum_0^\infty \beta^t d_t - E_{\hat{\pi}} \sum_0^\infty \beta^t d_t \\
&= E_{\hat{\hat{\pi}}}\left[\sum_0^{\tau_i - 1} \beta^t d_t + \sum_{\tau_i}^{\tau_i + \tau_k - 1} \beta^t d_t\right] - E_{\hat{\pi}}\left[\sum_0^{\tau_k - 1} \beta^t d_t + \sum_{\tau_k}^{\tau_i - 1} \beta^t d_t\right] \\
&= E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t d_t + E_{\hat{\hat{\pi}}} \beta^{\tau_i} E_{\hat{\pi}} \sum_0^{\tau_k - 1} \beta^t d_t - E_{\hat{\pi}} \sum_0^{\tau_k - 1} \beta^t d_t - E_{\hat{\pi}} \beta^{\tau_k} E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t d_t \\
&= E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t d_t [1 - E_{\hat{\pi}} \beta^{\tau_k}] - E_{\hat{\pi}} \sum_0^{\tau_k - 1} \beta^t d_t [1 - E_{\hat{\hat{\pi}}} \beta^{\tau_i}] \\
&\geqq \gamma(q^i(0)) E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t E_{\hat{\pi}} \sum_0^{\tau_k} \beta^t (1-\beta)^{-1} - \gamma(q^k(0)) E_{\hat{\pi}} \sum_0^{\tau_k - 1} \beta^t E_{\hat{\hat{\pi}}} \sum_0^{\tau_i - 1} \beta^t (1-\beta)^{-1} \\
&\geqq 0.
\end{aligned}
$$

Thus $\hat{\hat{\pi}}$ is an improvement of $\hat{\pi}$.

Our main goal is to show that π is an improvement of $\hat{\pi}$. Now, at time τ_i, $\hat{\hat{\pi}}$ *might not* use the machine that π does. In this case, by shifting the time origin to τ_i and by repeating the above argument, one can obtain an improvement $\hat{\hat{\hat{\pi}}}$ of $\hat{\hat{\pi}}$. This policy $\hat{\hat{\hat{\pi}}}$ will coincide with π over a strictly longer initial time segment than $\hat{\hat{\pi}}$ does. Continuing in this way, we can obtain policies which coincide with π over arbitrarily large initial time segments and which are improvements of $\hat{\pi}$. Since this is a discounted cost problem with discount factor $\beta < 1$, it follows that π itself is an improvement of $\hat{\pi}$.

Since $\hat{\pi} = (\hat{\pi}_0, \tilde{\pi}, \tilde{\pi}, \tilde{\pi}, \cdots)$ where $\pi = (\tilde{\pi}, \tilde{\pi}, \tilde{\pi}, \cdots)$, it follows by standard results on the discounted cost problem that π is optimal.

π was special in that it used the natural ordering of $\{1, 2, \cdots, N\}$ to break ties. Clearly *any* ordering of $\{1, 2, \cdots, N\}$ could have been used. Standard results again show that at any given time t, one can use *any* machine i with the largest value $\gamma(q^i(t))$ and still achieve optimality. This proves (i) and (ii).

For (v), consider the problem equivalent to (iv) where the option of retiring and collecting M is replaced by a machine which has a *known* probability $M(1-\beta)$ of success (for $0 \leqq M \leqq 1/(1-\beta)$). By (ii) the index of the *known* machine is $M(1-\beta)$. By (i) the index of the unknown machine is therefore that value of $M(1-\beta)$ when one is exactly indifferent to playing the unknown machine once or the known machine once.

The result (vi) is from Whittle [18] and for a proof we refer the reader to Whittle [20] or Ross [21]. (In the attribution of (iii) we have followed [16]). □

The above results possess some very special features which we now discuss. First, (i) shows that to each machine one may assign a (desirability) index $\gamma(q)$ which *just* depends on the *state* of the machine and *nothing else*. The optimal rule then, according to (i), has the intuitively appealing interpretation that at any given time one should merely *compare* the indices of the available machines, and then play the machine with the largest index. Thus, we have the very useful property that the problem of dealing with N machines simultaneously, with attendant state $\mathbf{q} = (q^1, \cdots, q^N)$ *simplifies* (or *decouples*) into N *separate* machines each with state q^i, $1 \leqq i \leqq N$.

Second, the index of a machine which from now on we refer to as its *Gittins* index, can be interpreted according to (ii) as the maximum reward per unit time (where both rewards and time are progressively discounted by a factor β) given that one may

stop anytime. Thus the problem of computation of the Gittins index is an optimal stopping problem.

Third, (v) affords yet another interpretation of the Gittins index. Consider the problem of (iv) where there is a machine and a retirement option of M units. Then, the Gittins index is that value of $M(1-\beta)$ at which one is exactly indifferent to retiring or playing the machine once and preserving the retirement option.

Last, (vi) shows that one can actually obtain an expression for the optimal reward function $J(\mathbf{q})$ of the N-armed bandit problem in terms of the optimal reward functions of *single*-armed bandits with retirement options. This is quite remarkable in view of the formidable functional equation defining $J(\cdot)$.

From the characterization of τ in (iii) it follows that for $\sigma < \tau$, $\gamma(q(\sigma)) > \gamma(q(0))$. Thus at any time *prior* to τ, the Gittins index is larger than the Gittins indices of the other machines if it were so at time 0. Hence, one should play the machine with the largest index continuously for a period of at least τ units. Then only need one consider changing machines.

This result can be generalized in several ways. The first is as follows. Consider the situation where there are N Markov processes with states $q^1(t), q^2(t), \cdots, q^N(t)$. At each time t one can choose to let *one* of the processes evolve while freezing *all* the rest. If process i is chosen for evolution, then the next state $q^i(t+1)$ is chosen according to a transition probability function $P_i(dq^i(t+1)|q^i(t))$ and a reward (or a cost) $c^i(q^i(t))$ is obtained. All other processes have $q^j(t+1) = q^j(t)$ for $j \neq i$. The generalization here is that the process $\{q^i(t)\}$ need not represent the "hyperstates" or conditional probability distributions of some other process and the transition probability function $P_i(\cdot|\cdot)$ can be quite general. Further, the reward function $c^i(\cdot)$ can also be general. This general version is the one treated in [15] and it is quite clear that the proof above works without modifications. One can also consider an adaptive version of this problem where the transition probabilities are unknown, see [16].

The next generalization is to allow each of the above Markov processes to be a Markov decision process, i.e., after choosing one process for evolution, one also has to choose a *control* action to apply to this process, see Gittins [15] where each process is now referred to as a "superprocess". Now, however, it is not generally true that the index result generalizes; some restrictions have to be imposed. Whittle [18] shows that if *each* individual superprocess i with a retirement option of M units has an optimal stationary policy (in the sense of Markov decision processes) with the property that it is optimal for *all* values of M, then an index rule is optimal. (Note that in the previously considered processes the control set is a singleton and so clearly this condition is satisfied.) Glazebrook [23] shows that this condition is both weakly and strongly necessary in a certain sense.

Another generalization of the bandit problem is to the case where new (slot) machines arrive in the course of time; this is treated by Whittle [24].

Varaiya, Walrand and Buyukkoc [25] consider a different sort of extension and show that the process involved need not even be Markov. They also exhibit certain problems which are equivalent to bandit problems, but which are nevertheless more useful from the viewpoint of certain applications.

There are many applications of these results in the areas of stochastic scheduling (of jobs by a server), searching (for an object hidden in one of several boxes), planning of research (which option should be investigated next), design of experiments etc. Just as an illustration; if there are many customers and only one server, then the problem of deciding which customer to serve is analogous to deciding which of several machines to play in a bandit problem. We refer the reader to [19] for references.

Let us revert to the N-armed bandit formulation. For each $\beta \in (0, 1)$, the implementation of each β-optimal policy converges a.s. onto *some* machine, i.e., after some time only one machine is exclusively played and the others are ignored (under some technical conditions), see Rothschild [26] and Kelly [27]. Thus every β-optimal policy *experiments* with the machines for only a certain amount of time after which it settles down to playing some machine exclusively, i.e., *stops* experimenting. It is of course of interest to obtain the probability P_β that the β-optimal policy will settle down ultimately to playing the machine with the *largest* probability of success. As $\beta \to 1$ it is shown in [27] that $P_\beta \to 1$. However $P_\beta < 1$ for all β, see Rothschild [26].

Now we consider another aspect. First we define a policy, the "least failures policy", which operates as follows. At each instant of time we play the machine with the least number of previous failures, except that if there is more than one such machine, then we select the one with the greatest number of successes, and again if there is more than one such, then we choose in a uniformly randomized way from among all the contenders. Kelly [27] shows that as $\beta \to 1$, the β-optimal policies evaluated at each state converge (but *not* uniformly over all states) to the least failures policy (we assume all machines start alike).

For β close to 1 each β-optimal policy therefore plays the least failures rule for a certain amount of time (until it leaves a certain set of states over which the β-optimal policy coincides with the least failures rule). Thus we now have a very nice interpretation of the behaviour of the β-optimal policies. There are three phases of behaviour. During the first phase, there is *experimentation* with the various machines (because the least failures rule constantly switches between machines). Then there is an intermediate period and finally, in the last phase, *learning* completely stops and the policy only plays one machine to the exclusion of all others. This explanation of Kelly [27] is as good and rigorous as any we have seen regarding the so-called "dual" effect in Bayesian adaptive control, see § 5.

There is a discontinuity in the behaviour at $\beta = 1$. As $\beta \to 1$, the limiting policy, the least failures rule is *not* optimal with respect to maximizing the average reward per unit time. However for β close to 1, the β-optimal policies do perform well with respect to the average cost criterion in the sense that $P_\beta \approx 1$.

This discussion has shown that we can obtain policies which are optimal with respect to the discounted cost criterion and *close* to optimal with respect to the average cost criterion. The converse can also be done. By considering randomized policies, Glazebrook [28] shows that one can obtain policies which are optimal with respect to the average cost criterion and *close* to optimal with respect to the discounted cost criterion; see also Bather [29].

As we have seen in the above theorem, we can reduce the N-armed bandit problem to N separate one-armed bandit problems of the type featured in (iv) of the theorem, which, of course, leaves us with the task of dealing with these one-armed bandit problems. Gittins [15] shows how one can develop recursive approximation schemes. Glazebrook [20] shows how the index results can be used to obtain bounds on the quality of *any* stationary strategy.

Explicit analytic results on these one-armed bandit problems are available in the case of "improving" or "deteriorating" arms, see [15]. In [31] by using careful bounds for the dynamic programming equation it is shown that if the known arm has a probability $m/(n+m)$ of success where m and n are relatively prime integers, then if the unknown machine has a probability of success which is Beta distributed with integer parameters x and y, then the optimal policy is to play the unknown machine or the known machine, respectively, according as $x/(x+y) \geqq m/(m+n)$ or $x/(x+y) <$

$m/(m+n)$ for all $0<\beta\leqq 1/(n+1)$. Another result in the same vein is described in [30] where it is indicated that by computation it was determined that for all $0<\beta<0.801$ and $m=n=1$, the optimal policy is to play the unknown machine until it has more failures (x) than successes (y).

4. Bayesian control of Markov chains. As we have seen in the previous § 3, the bandit family of problems possesses a highly special structure which can be usefully exploited to provide a deep theory. What results are available in the general BACP where such special structure does not exist? If one considers either the total discounted or total undiscounted cost criteria, deep results do not appear to be available, other than in highly degenerate situations, some examples of which are given in van Hee [14].

Attention has therefore been turned to the problem of obtaining accurate approximations to the solution of the dynamic programming equation. We consider below the case where one obtains perfect (as opposed to noisy or incomplete) observations of the state of the system. In such situations one can take advantage of the special structure with which BACP's are endowed.

Let X, U, Θ, all finite, be the state, control and parameter spaces, and suppose that the transition probabilities are given by the function $p(i,j,u;\theta)=$ Prob $(x_{t+1}=j|x_t=i, u_t=u, \theta)$. The cost function is $E\sum_0^\infty \beta^t c(x_t, u_t, x_{t+1})$ where $0<\beta<1$.

First consider the problem of (nonadaptive) control when the parameter value is *known* to be θ. Let $\Pi:=\{\pi|\pi: X\to U\}$ be the set of stationary policies, $J^\pi(i,\theta)$ the expected cost when π is employed and the initial state is i, and $J(i,\theta)$ the optimal cost function.

Turning now to the BACP, let $P(\Theta)$ be the set of probability measures on Θ and let $J(i,q)$ be the optimal cost function where q is the prior distribution. (We have used the same notation as for the nonadaptive problem since one can identify a known parameter θ with a degenerate distribution concentrated on θ.) Define T, the standard dynamic programming operator by

$$(TV)(i,q)=\min_{u\in U}\sum_{\theta\in\Theta}\sum_{j\in X} q(\theta)p(i,j,u;\theta)[c(i,u,j)+\beta V(j,S(i,u,j;q))].$$

Here $S(i,u,j;q)\in P(\Theta)$ is the posterior distribution of θ, if q is the prior distribution and a transition of the state from i to j under u is observed. For a standard reference on discounted dynamic programming, the reader is referred to Blackwell [32].

THEOREM 4.1 (van Hee). i) *Let*

$$L(i,q):=\sum_{\theta\in\Theta} q(\theta)J(i,\theta),$$

$$U(i,q):=\min_{\pi\in\Pi}\sum_{\theta\in\Theta} q(\theta)J^\pi(i,\theta),$$

$$T^n := \textit{nth iterate of the map } T.$$

Then

$$L(i,q)\leqq TL(i,q)\leqq\cdots\leqq T^nL(i,q)\leqq\lim_N T^NL(i,q)=J(i,q)$$

$$=\lim_N T^NU(i,q)\leqq T^nU(i,q)\leqq\cdots\leqq TU(i,q)\leqq U(i,q).$$

ii)

$$\sup_{i,q}|T^nU(i,q)-T^nL(i,q)|\leqq\beta\sup_{i,q}|T^{n-1}U(i,q)-T^{n-1}L(i,q)|.$$

Proof. The lower and upper bounds $L(i, q) \leqq J(i, q) \leqq U(i, q)$ are obvious. It is well known that $TJ = J$. Since T is monotone, it follows that $T^n L \leqq T^n J = J \leqq T^n U$. Now

$$\begin{aligned} TL(i, q) &= \min_{u \in U} \sum_{\theta \in \Theta} \sum_{j \in X} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta L(j, S(i, u, j; q))] \\ &= \min_{u} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta \sum_{\theta' \in \Theta} S(i, u, j; q)(\theta') J(j, \theta'). \end{aligned}$$

By Bayes' rule we can obtain

$$\sum_{\theta} p(i, j, u; \theta) q(\theta) S(i, j, u; q)(\theta') = q(\theta') p(i, j, u; \theta').$$

So

$$\begin{aligned} TL(i, q) &= \min_{u} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta J(j, \theta)] \\ &\geqq \sum_{\theta} q(\theta) \left\{ \min_{u} \sum_{j} p(i, j, u; \theta)[c(i, u, j) + \beta J(j, \theta)] \right\} \\ &= \sum_{\theta} q(\theta) J(i, \theta) = L(i, q). \end{aligned}$$

By the monotonicity of T, it follows that $L \leqq TL \leqq T^2 L \leqq \cdots \leqq T^n L$. Also

$$\begin{aligned} TU(i, q) &= \min_{u \in U} \sum_{\theta \in \Theta} \sum_{j \in X} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta U(j, S(i, u, j; q))] \\ &= \min_{u} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta) \left[c(i, u, j) + \beta \min_{\pi \in \Pi} \sum_{\theta' \in \Theta} S(i, u, j; q)(\theta') J^{\pi}(j, \theta') \right] \\ &\leqq \min_{u} \min_{\pi \in \Pi} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta) \left[c(i, u, j) + \beta \sum_{\theta'} S(i, u, j; q)(\theta') J^{\pi}(j, \theta') \right] \\ &= \min_{u} \min_{\pi \in \Pi} \sum_{\theta} \sum_{j} q(\theta) p(i, j, u; \theta)[c(i, u, j) + \beta J^{\pi}(j, \theta)] \\ &\leqq \min_{\pi \in \Pi} \sum_{\theta} \sum_{j} q(\theta) p(i, j, \pi(i); \theta)[c(i, \pi(i), j) + \beta J^{\pi}(j, \theta)] \\ &= \min_{\pi \in \Pi} \sum_{\theta} q(\theta) J^{\pi}(i, \theta) = U(i, q). \end{aligned}$$

Again by monotonicity and contractivity of T, the result follows. □

The usefulness of the above theorem lies in the fact that $T^n L$ and $T^n U$ are lower and upper bounds for J for every n, which moreover converge monotonically to J as $n \to \infty$. How does one use these results? First, $J^{\pi}(i, \theta)$ and $J(i, \theta)$ are obtainable by standard algorithms such as the policy iteration algorithm, see Howard [33]. Thus $L(i, q)$ and $U(i, q)$ are obtainable by standard methods. The determination of $T^n L(i, q)$ (or similarly of $T^n U(i, q)$) for a fixed (i, q) clearly requires the values $T^{n-1} L(j, S(i, u, j; q))$ for every (i, u, j), of which there are finitely many. To see exactly what is involved in the computation of $T^n L(i, q)$, define $R(q) := \cup_{i,u,j} \{S(i, u, j; q\}$ and $R^j(q) = R(R^{j-1}(q))$ (= image of $R^{j-1}(q)$ under R). Thus $R^j(q)$ is the set of posterior distributions which could possibly result after j stages. It is clear that the data needed for the computation of $T^n L(i, q)$ for fixed (i, q) are $\cup_{j=1}^{n-1} \{T^j L(\cdot, \bar{q}): \bar{q} \in R^{n-j}(q)\}$. In theory, therefore, one can calculate $T^n L(i, q)$ after a finite set of computations. Since, by (ii) of the theorem, or by any other way, one can choose n so large that $T^n U(i, q) - T^n L(i, q) \leqq \varepsilon$, it follows that one can approximate $J(i, q)$ to within an error of ε. However, one should note that the computation of $T^{n+1} L(i, q)$ requires the data

$\cup_{j=1}^{n}\{T^jL(\cdot,\bar{q})\colon \bar{q}\in R^{n+1-j}(q)\}$. Since $R^{n-j}(q)$ and $R^{n+1-j}(q)$ are different sets, and quite possibly disjoint sets, it follows that one cannot proceed from $T^nL(i,q)$ to $T^{n+1}L(i,q)$ in a recursive way.

Thus, the above process of approximation can be quite cumbersome. Some special situations are identified in [14] where one may use simpler procedures. [14] also considers a discretization approach when Θ is not finite.

Martin [12] and Satia and Lave [34] have provided other upper and lower bounds. The bounds in [34] are obtained through an intermediate process (involving a worst case choice of θ and a best case choice of θ, i.e., by solving a min-max and a max-max situation) and are generally poorer than the bounds L and U given above.

A branch and bound algorithm. An interesting "branch and bound" algorithm to obtain an ε-optimal control for the state (i,q) is given by Satia and Lave [34]. This uses, in an essential way, upper and lower prior bounds for the optimal cost function, and we may suppose here that these bounds are U and L. Basically, one examines the possible branches (trajectories) leading out of the state (i,q). At any node, say state $(j,\bar{q})$, one has, at each stage, bounds for the costs of the states resulting from the application of, say, u. One can clearly eliminate from consideration those u's for which the lower bound on the cost-to-go is *larger* than the upper bound on the cost-to-go available through some other control. Thus, at each stage of the algorithm some decisions at nodes are eliminated from consideration and, in addition, the upper and lower bounds at all nodes are refined. Then the decision tree is extended by considering one more time unit of horizon. The prior bounds for these newly introduced nodes are U and L, and the whole process of eliminating decisions and refining the bounds is carried out all over again.

The algorithm is guaranteed to yield an ε-optimal decision in a finite number of iterations. Systems with four states and two decisions are described as being of moderate size, and for such systems the algorithm is regarded as being efficient [34]. However, convergence is reported to be slow when β is close to 1.

In contrast with the discounted cost problem considered above, optimal policies for the BACP are obtainable when the cost criterion is of the average cost type: $E\lim 1/N\sum_1^N c(x_t,u_t)$. This sort of a cost criterion is however more properly examined within a non-Bayesian context, as we see in §§ 6 and 7.

5. Bayesian adaptive control of linear systems with quadratic costs. An example will clarify the situation vis-à-vis BACP's in this category. Consider the Auto Regressive Moving Average System with EXogeneous Inputs (ARMAX) system,

$$y(t+1)=a_0y(t)+\cdots+a_ny(t-n)+b_0u(t)+\cdots+b_nu(t-n)+w(t+1)$$

where $(a_0,a_1,\cdots,a_n,b_0,b_1,\cdots,b_n)^T:=\theta$. θ, the unknown parameter, will be regarded as having a prior distribution which is normal $N(\bar{\theta},\bar{\Sigma})$. $\{w(t)\}$ is a sequence of independent, identically distributed $N(0,\sigma^2)$ random variables, i.e., white Gaussian noise. The goal is to minimize $E\sum_1^N y^2(t)$ (say). (We assume that $(y(0),y(-1),\cdots,y(-n-1),u(0),\cdots,u(-n-1)$ are known initial conditions.)

We have seen in § 2 that a central part of the BACP is to obtain the posterior distribution of the unknown parameter θ, given the observations $y(0)$, $u(0)$, $y(1),u(1),\cdots,y(t),u(t)$ made up to time y. To solve this problem of obtaining the posterior distribution, we rewrite the above system as

$$\theta(t+1)=\theta(t),\qquad \theta(0)=\theta\sim N(\bar{\theta},\bar{\Sigma}),$$

$$y(t+1)=\phi^T(t)\theta(t)+w(t+1)$$

where $\phi(t) := (y(t), y(t-1), \cdots, y(t-n), u(t), u(t-1), \cdots, u(t-n))^T$. Now it is clear that the posterior distribution is a normal distribution, the mean $\hat{\theta}(t)$ and covariance $\Sigma(t)$ of which are obtained through the Kalman filtering equations.

(If θ initially is *not* normally distributed, then some recent results for the continuous time problem, Makowski [35], may prove useful.)

The hyperstate for the BACP is thus $(\hat{\theta}(t), \Sigma(t), y(t), \cdots, y(t-n), u(t), \cdots, u(t-n))$. Thus it is perfectly clear that in view of the highly nonlinear manner in which the variables $\hat{\theta}(t)$ and $\Sigma(t)$ depend on the past $(y(t), u(t), y(t-1), u(t-1), \cdots)$, we really have a highly nonlinear system with a quadratic cost criterion. Indeed one can solve the problem of minimizing the finite horizon cost criterion $\sum_1^N y^2(t)$ when $N=1$, but no analytical solutions are available when $N \geqq 2$, see Åström and Wittenmark [36].

This lack of solutions has spurred many attempts at qualitatively understanding the nature of the optimal input sequence. The two qualitative features which are most popular are "caution" and "probing", which we shall briefly explain. Consider the simplest system of the type considered:

$$y(t+1) = y(t) + bu(t) + w(t+1)$$

where the prior distribution for b is $N(\bar{b}, \Sigma)$ with $\bar{b} > 0$. $\{w(t)\}$ is, as before, an i.i.d. $N(0, \sigma^2)$ sequence. $y(0)$ is known.

First consider the cost criterion $E(y^2(1))$. The optimal value of $u(0)$ is easily calculated to be

$$u(0) = \left(\frac{\bar{b}}{\bar{b}+\Sigma}\right) y(0).$$

Thus we note that as Σ increases, $|u(0)|$ decreases, i.e., as the uncertainty (here, variance) of b increases, the control (in absolute value) decreases. The controller is said to be *cautious.*

On the other hand, if the cost criterion is $E(y^2(2))$, then we know that

$$\min_{u(1)} E[y^2(2)|y(0), u(0), y(1)] = \frac{\Sigma(1)}{\hat{b}^2(1)+\Sigma(1)} y^2(1) + \sigma^2$$

where $\hat{b}(1) = E(b|u(0), y(1))$ and $\Sigma(1) = E[(b-\hat{b}(1))^2|u(0), y(1)]$. Thus at time $t=1$, it is preferable to have a smaller value of $\Sigma(1)$—all other considerations remaining the same. Since $\Sigma(1)$, the conditional variance of the estimation error, is given by

$$\Sigma(1) = \frac{\sigma^2 \Sigma}{\sigma^2 + u^2(0)\Sigma}$$

it would appear at first sight that choosing a large value of $|u(0)|$ is helpful. The choice of large values of the control input to enhance identification is called "*probing*". However, increasing $|u(0)|$ could also increase $|y(1)| (= |y(0) + bu(0) + w(1)|)$ and hence also $y^2(1)$. This of course makes the cost larger. Thus there is a tradeoff between increasing the control to probe the system and decreasing the control to reduce the uncurred cost. This sort of a problem is therefore referred to as a "dual control" problem, in view of the possibly dual purposes in applying a control action.

For another discussion of the "dual" control effect, the reader is referred to that portion of § 3 where the work of Kelly [27] is discussed.

Problems of the type described here have been considered by Feldbaum [37], Jacobs and Patchell [38], Tse and Bar-Shalom [39]-[41], Bar-Shalom and Tse [42],

Wittenmark [43], Wenk and Bar-Shalom [44], Bar-Shalom [45], Deshpande, Upadhyay and Lainiotis [46], Lainiotis [47], Dersin, Athans and Kendrick [48]. [37] is an early reference. [38] treats an incompletely observed problem which allows computation of the optimal solution and examines the nature of resulting optimal control law. [39] proposes a control law based on replacing the nonlinear system by a version linearized (in some sense) around a trajectory resulting from a nominal control law. [41] and [42] show when one may expect a certainty-equivalence control law to be optimal. In [45] a decomposition of an approximation of the cost-to-go function into three components is made, which supposedly reflect caution, probing and the residual cost. In [44] the performance of an approximation of [46], [47], which consists of using a control which is the weighted average (given by the posterior distribution) of the optimal controls for the various parameters, is examined as well as another algorithm which first averages over the parameter values and then chooses the control based on it. [48] examines a particular problem and evaluates the results of [45].

In the continuous time case, Rishel [140] has recently shown that the optimal control can be written in terms of the solution of a certain stochastic two-point boundary value problem. Also, Hijab has shown that for a particular measurement equation, and a particular cost criterion including an entropy term, the optimal control is the conditional mean of the optimal controls in the various parameter values, see [141].

6. Non-Bayesian adaptive control. In the *Bayesian* adaptive control problem (BACP), the specification of the problem is fairly rigid. Given an a priori probability distribution for the unknown parameter, one has to obtain a control law which minimizes the *expected* value of a certain cost criterion.

In the *non-Bayesian* adaptive control problem (NACP), more flexibility is allowed in the *design* of control laws. However, the designed control law must meet certain other (asymptotic) criteria in order to be deemed acceptable. To illustrate some of the central concepts, we start with a very simple example due to Robbins [49]—the non-Bayesian two-armed bandit adaptive control problem.

6.1. The non-Bayesian two-armed bandit problem. There are two slot machines: A and B. When machine A (or B) is used, one obtains one unit of reward with probability p_A (or p_B) and zero units of reward with probability $1-p_A$ (or $1-p_B$). Without loss of generality, we assume $p_A > p_B$.

Consider first the case when p_A and p_B are known. If at *each* time $t = 1, 2, 3, \cdots$ we play machine A exclusively, then $\lim (1/N) \sum_1^N r_t = p_A$ a.s., where $r_t :=$ random reward earned at time t. This is, almost surely, the maximum long-term average reward that one could possibly gain even by switching between machines, and obviously this is achievable by playing machine A exclusively.

Suppose now that we do *not* know the values of p_A and p_B. All we know is that $p_A \in (0, 1)$ and $p_B \in (0, 1)$. (Note that we are *not* provided with an initial probability distribution for $\theta := (p_A, p_B)$. This is what distinguishes this from a BACP.) However, we are as ambitious as before. We would still like to have a policy of playing the machines which ensures that $\lim (1/N) \sum_1^N r_t = \max (p_A, p_B)$ a.s.

We now exhibit a policy which achieves this goal. Let $u_t \in \{A, B\}$ be the control input chosen at time t, where $u_t = A$ or B according to whether A or B is played. Let $y_t \in \{0, 1\}$ denote the observation made at time t, where $y_t = 0$ or 1 according to whether a reward was not earned or was earned at time t. Here $r_t = y_t$.

To define the policy, we first choose any increasing sequence $\{a_n\}$ of positive integers, such that $\lim (1/N) \sum_{t=1}^N 1(t = a_i \text{ for some } i) = 0$. At each time t, we make

"estimates" $\hat{p}_A(t)$ and $\hat{p}_B(t)$ of p_A and p_B respectively, by

$$\hat{p}_A(t) := \frac{\sum_0^{t-1} 1(u_n = A, y_n = 1)}{\sum_0^{t-1} 1(u_n = A)} \quad \text{and} \quad \hat{p}_B(t) := \frac{\sum_0^{t-1} 1(u_n = B, y_n = 1)}{\sum_0^{t-1} 1(u_n = B)}.$$

We note, in passing, that these are the maximum likelihood estimates of p_A and p_B. We now choose u_t according to:

$$u_t = \begin{cases} A & \text{if } (t = a_{2n} \text{ for some } n) \text{ or } (t \neq a_n \text{ for all } n \text{ and } \hat{p}_A(t) \geqq \hat{p}_B(t)), \\ B & \text{if } (t = a_{2n+1} \text{ for some } n) \text{ or } (t \neq a_n \text{ for all } n \text{ and } \hat{p}_A(t) < \hat{p}_B(t)). \end{cases}$$

Basically, except for the times $t = a_1, a_2, a_3, \cdots$ we play whichever of A or B has the larger "estimated probability" of a win. However, the times $t = a_1, a_2, a_3, \cdots$ are *reserved* for experimentation.

Now we show that the above policy attains our goal $\lim (1/N) \sum_1^N r_t = \max (p_A, p_B)$ a.s., and does so *without* knowing the values of (p_A, p_B). To see this, first note that by the reservation of the experimentation times, each of A and B is played infinitely often (i.o.) a.s. By the law of large numbers therefore, $\hat{p}_A(t) \to p_A$ and $\hat{p}_B(t) \to \hat{p}_B$ a.s., and so for all $t \geqq$ some $T(\omega)$, $\hat{p}_A(t) > \hat{p}_B(t)$. So, A is exclusively played after time T, except at some of the *reserved* times. But these reserved times are so sparse, that they make no contribution to the *average* cost. More precisely:

$$\begin{aligned} \lim \frac{1}{N} \sum_1^N r_t &\geqq \lim \frac{1}{N} \sum_1^N 1(u_t = A, y_t = 1) \\ &= \lim \frac{1}{N} \sum_1^N 1(u_t = A) \lim \frac{\sum_1^N 1(u_t = A, y_t = 1)}{\sum_1^N 1(u_t = A)} \\ &= p_A \lim \frac{1}{N} \sum_1^N 1(u_t = A) \\ &\geqq \lim \frac{p_A}{N} \left(N - T - \sum_{t=1}^N 1(t = a_i \text{ for some } i) \right) = p_A \quad \text{a.s.} \end{aligned}$$

Three properties of this non-Bayesian adaptive control scheme should be noted.

i) $\lim (1/N) \sum_1^N r_t = \max (p_A, p_B)$ a.s. Thus, the cost of this scheme is *optimal*, i.e., it could not be bettered even if we knew the values of p_A and p_B.

ii) $\lim (\hat{p}_A(t), \hat{p}_B(t)) = (p_A, p_B)$. Thus the parameter estimates are consistent, i.e., the true parameters are identified.

iii) $\lim u_t$ does not exist. Hence the control scheme does not converge, and therefore, of course, it does not converge to an optimal control scheme. (However, it does converge in a Cesaro sense, $\lim (1/N) \sum_1^N 1(u_t = A) = 1$ a.s.)

6.2. Non-Bayesian adaptive control versus Bayesian adaptive control. We shall discuss some of the differences between the Bayesian and non-Bayesian approaches to adaptive control.

In the previous section, we have obtained a policy π for which

$$\lim \frac{1}{N} \sum_1^N r_t = \max (p_A, p_B) \quad \text{a.s.}$$

A more precise way of stating the above fact is as follows. There exists a policy π such that

$$lim \frac{1}{N} \sum_1^N r_t = \max J(\theta^0), \qquad P_{\theta^0}^{\pi}\text{-a.s. for every } \theta^0 \in \Theta.$$

(Here $\theta^0 := (p_A, p_B)$, $J(\theta^0) := \max(p_A, p_B)$, $\Theta = (0,1)\times(0,1)$ and $P_{\theta^0}^{\pi}$ is the probability measure induced on the trajectories of the system by the policy π when the parameter value is θ^0.) This clearly shows that no matter *what* the value of θ^0 is, the policy π attains the maximum reward attainable for *that* value of θ^0. In non-Bayesian adaptive control with respect to an *average* cost criterion, we shall frequently impose such a requirement on a policy, viz. it should be optimal *uniformly* for all $\theta^0 \in \Theta$ a.s.

If the requirement above can be met by some policy π, then it is clear that *all* Bayesian problems with the above cost criterion are also immediately solved. The reason is that if $q(\theta)$ is the prior distribution of θ, then the policy π, when implemented, attains the expected cost $\sum_\theta q(\theta)J(\theta)$, and clearly no policy can do better. (Thus π attains the obvious lower bound in Theorem 4.1 of § 4.) Hence π is optimal even in a Bayesian framework *irrespective* of the prior distribution q.

We will show in the sequel that one can often obtain a policy π meeting the requirements above, and frequently there will be several π's which do so. Thus, insofar as just the long-term average cost criterion is concerned, the non-Bayesian formulation is unquestionably superior to the Bayesian formulation.

Optimal policies for the long term average cost criterion are nonunique in an essential way, since what happens in the initial period does not alter the cost. Since one is often (practically) interested in attaining a fast rate of convergence to optimal control laws (or a fast rate of convergence of the parameter estimates etc.), one may choose between several π's meeting the above requirement by imposing some other criterion, such as rate of convergence, to judge them. Alternatively, one could pose, as in § 3 in the discussion of [27]–[29], the problem of obtaining a policy which is optimal for the average cost criterion and nearly optimal for the discounted cost problem or vice-versa. Or, one could study the rate at which the difference between the finite horizon cost of adaptive control and its optimal value grows with the horizon.

However, if the cost criterion is of the discounted type $\sum_0^\infty \beta^t c(x_t, u_t)$, then the BACP and NACP are rather different. For the NACP, it is clear that one cannot expect that $\sum_0^\infty \beta^t c(x_t, u_t)$ will converge a.s. to a constant, as it did in the average cost case. If we replace this by the requirement that $E_{\theta^0}^{\pi} \sum_0^\infty \beta^t c(x_t, u_t) = J(x_0, \theta^0)$ for all $\theta^0 \in \Theta$, then this is clearly too strong to be met by a single π. So finally, we arrive at a requirement of the type,

$$\lim_N \left\{ E_{\theta^0}^{\pi} \sum_N^\infty \beta^{t-N} c(x_t, u_t) - E_{\theta^0}^{\pi} J(x_N, \theta^0) \right\} = 0 \quad \text{for all } \theta^0 \in \Theta$$

which is of a reasonable nature, see Schäl [63].

So far, however, we have only paid attention to the convergence of the costs, and *not* the controls. So, we now address the problem of convergence of control laws in a NACP. For each $\theta \in \Theta$, let π^θ be an optimal stationary policy for the discounted cost problem (and for simplicity of discussion we assume that it is unique). Then one can ask the following question: Is there a policy $\pi = (\pi_0, \pi_1, \cdots)$ for which

$$\lim \{\pi_N(x_0, u_0, x_1, u_1, \cdots, x_N) - \pi^{\theta^0}(x_N)\} = 0, \qquad P_{\theta^0}^{\pi}\text{-a.s. for all } \theta^0 \in \Theta \quad ?$$

Thus we are requiring that the controls generated by the adaptive scheme π should converge asymptotically to the optimal controls, uniformly for all $\theta^0 \in \Theta$. This, again, is a reasonable requirement in many instances.

Since the asymptotic requirements on the costs and the controls are closely related, we shall refer to either of these requirements (perhaps with slight modifications), in an NACP, as a *self-optimizing* requirement. One of the main goals of non-Bayesian adaptive control is to obtain a self-optimizing policy.

This situation is in contrast to the BACP's for which, frequently, the optimal policy is not self-optimizing. An example is given in [26], and the reader is referred to the discussion in § 3 on [26], [27].

From a practical point of view, in a BACP, one is typically faced with problems where the computational burden is very high. For an NACP, one is typically interested in the rate of convergence of the self-optimizing policy, if indeed one can obtain one.

7. Non-Bayesian adaptive control of Markov chains. To begin, we consider the case where the state, control and parameter spaces, X, U and Θ are all finite. The transition probabilities are given, for each $\theta \in \Theta$, by $\{p(i, j, u; \theta): i, j \in X, u \in U\}$. The value of the true parameter is θ^0. All we know is that θ^0 is *some* element of Θ. Our goal is to design a policy π which a.s. attains the minimum of $\lim (1/N) \sum_1^N c(x_t, u_t)$, where $c(i, u)$ is a one-stage cost function.

We start by considering a scheme that, *at first sight*, looks very reasonable. At each time t we will have accumulated a history $(x_0, u_0, x_1, u_1, \cdots, x_t)$ and we can use this history to form (say) a maximum-likelihood estimate (MLE) $\hat{\theta}_t$ of the unknown parameter. Thus, we choose $\hat{\theta}_t \in \Theta$ so that

$$\prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \hat{\theta}_t) \geqq \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \theta) \quad \text{for all } \theta \in \Theta.$$

(In the event that there is more than one maximizer of the likelihood function, one can choose a particular maximizer according to some prespecified priority order on elements of Θ.) Then we choose a control input u_t which is optimal if the parameter value was $\hat{\theta}_t$, i.e., we choose

$$u_t = \phi(x_t, \hat{\theta}_t)$$

where, for each $\theta \in \Theta$, $\phi(\cdot, \theta): X \to U$ is an optimal stationary control law (policy).

Many questions arise.

i) Does $\hat{\theta}_t$ converge a.s.?

ii) Does $\hat{\theta}_t$ converge to θ^0 a.s.?

iii) Does $\phi(\cdot, \hat{\theta}_t)$ converge a.s.?

iv) Does $\phi(\cdot, \hat{\theta}_t)$ converge to $\phi(\cdot, \theta^0)$ a.s.?

v) Does $\lim (1/N) \sum_1^N c(x_t, u_t)$ converge a.s.?

vi) Does $\lim (1/N) \sum_1^N c(x_t, u_t)$ converge to $J(\theta^0)$ a.s.? Here $J(\theta^0)$ is the optimal cost achievable for the parameter θ^0, and we assume that it does not depend on the initial state x_0.

vii) At what rate do these quantities converge, if they do so?

In a nice counterexample, Borkar and Varaiya [50] demonstrate that (ii) need not hold. We provide below a counterexample in a similar vein, from [51], to illustrate that (ii), (iv) and (vi) do not hold.

Counterexample. Let $X = \{1, 2\}$, $U = \{1, 2\}$, $\Theta = \{1, 2, 3\}$, $\theta^0 = 1$. The transition probabilities are $p(1, 1, 2; 1) = p(1, 1, 2; 2) = 0.8$, $p(1, 1, 2; 3) = 0.2$, $p(2, 1, u; \theta) = 1$ for all u, θ. The cost function $c(x_t, u_t, x_{t+1})$ is $c(i, u, j) = 3 + (2 - i)\ (7.8 - 0.3u - b_j)$. It is easily calculated, see [33], that the optimal policies are $\phi(i, 1) = 1$, $\phi(i, 2) = \phi(i, 3) = 2$ for all $i \in X$.

To see that $\hat{\theta}_t$ need not converge to 1, consider the starting values $x_0 = 1$, $u_0 = 1$. With probability 0.5, the next state is $x_1 = 1$. But then $\hat{\theta}_1 = 2$. Hence $u_1 = \phi(1, 2) = 2$. It can then be checked easily that $\hat{\theta}_t = 2$ for all $t \geqq 1$. Thus, there is a probability of at least 0.5 that $\lim \hat{\theta}_t \neq \theta^0$. This shows that the answer to each of the questions (ii), (iv) and (vi) is a no. □

The basic problem here is that one cannot fully identify a process in closed loop. Borkar and Varaiya [50] show the best that one may expect.

THEOREM 7.1 (Borkar and Varaiya). *If* $p(i,j,u,\theta) \geqq \varepsilon$ *for all* i,j,u,θ *then* $\lim \hat{\theta}_t = \hat{\theta}_\infty$ *a.s. and* $p(i,j,\phi(i,\hat{\theta}_\infty),\theta^0) = p(i,j,\phi(i,\hat{\theta}_\infty),\hat{\theta}_\infty)$ *a.s.*

Proof. Define $L_t(\theta) := \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \theta) p^{-1}(x_s, x_{s+1}, u_s; \theta^0)$, the likelihood ratio. For each $\theta \in \Theta$, $\{L_t(\theta), \mathcal{F}_t\}$ is a positive martingale (where $\mathcal{F}_t := \sigma\{x_0, u_0, x_1, u_1, \cdots, x_t\}$ is the σ-algebra generated by the observed past). Hence $\{L_t(\theta)\}$ converges a.s. for every $\theta \in \Theta$. Fix $\omega \in \Omega$, the sample space. If $\{\hat{\theta}_t(\omega)\}$ has $\bar{\theta} \in \Theta$ as a limit point, then $\hat{\theta}_t(\omega) = \bar{\theta}$ i.o. Since $L_t(\hat{\theta}_t) \geqq L_t(\theta^0) = 1$, it follows that $L_t(\bar{\theta}, \omega) \geqq 1$ i.o., and so $\lim L_t(\bar{\theta}, \omega) \geqq 1$ (as long as ω is not in a certain null set). This shows that $p(x_t(\omega), x_{t+1}(\omega), u_t(\omega); \bar{\theta}) p^{-1}(x_t(\omega), x_{t+1}(\omega), u_t(\omega); \theta^0) = L_{t+1}(\bar{\theta}, \omega) L_t^{-1}(\bar{\theta}, \omega) \to 1$. Hence, for all $t \geqq T(\omega)$, $p(x_t(\omega), x_{t+1}(\omega), u_t(\omega), \bar{\theta}) = p(x_t(\omega), x_{t+1}(\omega), u_t(\omega), \theta^0)$. If $\bar{\bar{\theta}}$ is any other limit point of $\{\hat{\theta}_t(\omega)\}$, then a similar result holds for all $t \geqq \bar{T}(\omega)$. Hence $L_t(\bar{\theta}, \omega) = L_t(\bar{\bar{\theta}}, \omega)$ for all $t \geqq \max(T(\omega), \bar{T}(\omega)) =: \bar{\bar{T}}(\omega)$. Since one always breaks ties by picking the particular maximizer which is highest in the priority ordering, it follows that $\bar{\theta} = \bar{\bar{\theta}}$, showing that $\hat{\theta}_t \to \hat{\theta}_\infty$ a.s. Hence we obtain $p(x_t, x_{t+1}, u_t; \theta_\infty) = p(x_t, x_{t+1}, u_t; \theta^0)$ a.s. By the Martingale Stability Theorem, $\lim (1/N) \sum_1^N 1(x_t = i, x_{t+1} = j) - E(1(x_t = i, x_{t+1} = j) | \mathcal{F}_{t-1}) = 0$ a.s. and since $E(1(x_t = i, x_{t+1} = j) | \mathcal{F}_{t-1}) \geqq \varepsilon^2$ it follows that $(x_t = i, x_{t+1} = j)$ i.o. a.s. Hence, the result follows. □

The above result is fundamental and needs elaboration. Why does $\hat{\theta}_t$ not converge to θ^0? The answer is this. If one could guarantee that $(x_t = i, u_t = u)$ i.o. for every (i, u), then one could hope to identify $p(i,j,u;\theta^0)$. But in closed loop one uses only those u_t's for which $u_t = \phi(x_t, \hat{\theta}_t)$. Thus if $\hat{\theta}_t \to \bar{\theta}$, then $(x_t = i, u_t = u)$ i.o. *only* if $u = \phi(i, \bar{\theta})$, and so one can only hope to identify $p(i,j,\phi(i,\bar{\theta});\theta^0)$, and this is the content of the above theorem.

Sagalovsky [52] treats the situation where the unknown probabilities depend in an affine way on a real valued unknown parameter, and shows how this structure can be exploited. [53] shows what further difficulties are encountered in generalizing Theorem 7.1 to the case where Θ is compact, instead of finite as above. Borkar and Varaiya [54] consider the situation where the state space is countable.

Earlier, Mandl [55]–[57] had considered the problem where U, Θ are compact, $p(i,j,u;\theta)$ and $c(i,u)$ continuous and, for simplicity, $p(i,j,u;\theta) > 0$. [55] considers the class of estimators based on general "contrast" functions. This class includes the maximum likelihood estimator considered above, if the following assumption is made.

Identifiability condition. For every $\theta, \theta' \in \Theta$ and $\theta \neq \theta'$, there exists an $i = i(\theta, \theta') \in X$ such that for every $u \in U$, there is a $j = j(i, u, \theta, \theta') \in X$ with $p(i,j,u;\theta) \neq p(i,j,u;\theta')$.

THEOREM 7.2 (Mandl). *If* U *and* Θ *are compact,* $p(\cdot)$, $c(\cdot)$, $\phi(\cdot)$ *and* $g(\cdot)$ *(see below) are continuous,* $p(i,j,u;\theta) > 0$ *and the Identifiability Condition is satisfied, then*

i) *For any policy* $\pi = (\pi_0, \pi_1, \cdots)$, $\lim \hat{\theta}_t = \theta^0$ *a.s.*

ii) *If* $\pi_t(x_0, u_0, x_1, \cdots, x_t) := \phi(x_t, \hat{\theta}_t)$, *and* ϕ *is continuous, then* $\lim (1/N) \sum_1^N c(x_t, u_t) = J(\theta^0)$ *a.s.*

Proof. The proof of part (i) is omitted, because the reader can deduce that at least for the policy of (ii), and when Θ is finite, Theorem 7.1 and the Identifiability Condition give the required result. For (ii), we use the theory of the long-term average cost criterion [33] to note that there exist $\{v_i(\theta): i \in X\}$ satisfying

$$J(\theta) + v_i(\theta) = c(i, \phi(i,\theta)) + \sum_j p(i,j,\phi(i,\theta);\theta) v_j(\theta)$$

$$\leqq c(i,u) + \sum_j p(i,j,u;\theta) v_j(\theta) \quad \text{for all } u, i, \theta.$$

Defining $g(i, u, \theta) := c(i, u) + \sum_j p(i, j, u; \theta) v_j(\theta) - J(\theta) - v_i(\theta)$, we see that $g(i, u, \theta) \geqq 0$ and $g(i, u, \phi(i, \theta)) = 0$. Now note that if $y(t+1) := c(x_t, u_t) - J(\theta^0) + v(x_{t+1}, \theta^0) - v(x_t, \theta^0) - g(x_t, u_t, \theta^0)$, then $E(y_{t+1}|\mathcal{F}_t) = 0$, and since $\{y_{t+1}\}$ is bounded, it follows by the Martingale Stability Theorem that $\lim (1/N) \sum_1^N y_{t+1} = \lim (1/N) \sum_1^N E(y_{t+1}|\mathcal{F}_t) = 0$ a.s. Hence, substituting for y_{t+1} and noting that $\{v_i(\theta^0)\}$ is bounded in i, gives $\lim (1/N) \sum_1^N c(x_t, u_t) = J(\theta^0) + \lim (1/N) \sum_1^N g(x_t, u_t, \theta^0)$. By part (i) however, $\lim (u_t - \phi(x_t, \theta^0)) = \lim (\phi(x_t, \hat{\theta}_t) - \phi(x_t, \theta^0)) = 0$, and so, by continuity of ϕ and g, $\lim g(x_t, u_t, \theta^0) = 0$. □

Note that a result such as (i) is very strong, since it guarantees that $\hat{\theta}_t \to \theta^0$ a.s. for *all* policies π. This of course points to the restrictive nature of the Identifiability Condition. However, in some problems, for example, in the control of queueing systems, such a condition is satisfied.

Kurano [58] considers a similar problem. Kolonko [59], [60] determines the appropriate generalization of the Identifiability Condition and other regularity assumptions which are sufficient to ensure that results such as (i) and (ii) of the above theorem hold in the adaptive control of Markov renewal processes. Applications to the adaptive control of queueing systems are also studied. Georgin [61] also considers the generalization of the above theorem to more general state spaces. Baranov [62] considers a different scheme where even the control laws are obtained by a recursive process.

(i) of the above theorem also has implications for cost criteria other than of the long-term average type. For example, consider the case of a discounted cost criterion, and suppose that the policy $\pi = (\pi_0, \pi_1, \pi_2, \cdots)$ is such that $\pi_t(x_0, \cdots, x_t) = \phi(x_t, \hat{\theta}_t)$ where $\phi(\cdot, \theta)$ is a stationary control law which is optimal for the discounted cost problem. Then, under reasonable continuity conditions, $\phi(\cdot, \hat{\theta}_t) \to \phi(\cdot, \theta^0)$. Under appropriate conditions, it then also follows that $E \sum_t^\infty \beta^{n-t} c(x_n, u_n) - EJ(x_t, \theta^0) \to 0$ a.s. Such a generalization is carried out in Schal [63].

How does one deal with the general situation where an Identifiability Condition may not hold? This is a difficult problem and essentially requires some procedure for overcoming the fundamental closed loop identification problem. There are several ways of doing this, which we now take up for consideration.

7.1. Forced choice schemes. Here, just as in the non-Bayesian bandit problem of § 6.1, some time instants $t = a_1, a_2, a_3, \cdots$ are set aside for experimentation. At these times, one must use forced choices of all the elements of U in (say) cyclic order. Thus, since $(x_t = i, u_t = u)$ occurs i.o. a.s., it follows that $\hat{\theta}_t \to \theta^0$ a.s. However if $\lim (1/N) \sum_{t=1}^N 1(t = a_i \text{ for some } i) = 0$, then the control actions taken at times $t = a_1, a_2, \cdots$ do not make any contribution to the average cost. At other times one just uses the control inputs $u_t = \phi(x_t, \hat{\theta}_t)$. Since one has $\hat{\theta}_t \to \theta^0$ a.s., it follows that optimal cost can be obtained. An approach of this type is followed in Fox and Rolph [64] for Markov renewal processes. Van Hee [14] considers a comparable scheme albeit in a Bayesian formulation.

7.2. Randomization schemes. These are schemes for which each $\pi_t(\cdot | x_0, u_0, x_1, \cdots, x_t)$ is allowed to be a probability measure on U according to which u_t is picked. Since u_t is random, every $u \in U$ has (in some sense) a positive probability of being applied at each time.

Doshi and Shreve [65] impose a less restrictive condition than the identifiability condition given earlier. At each time instant t, a parameter $\hat{\theta}_t$ is randomly chosen from among all those which very nearly maximize the likelihood function. This is then shown to overcome the identifiability problem. Borkar and Varaiya [54] also consider similar randomization schemes for countable Markov chains.

Sato, Ake and Takeda [66] have proposed quite a different scheme for the *discounted* cost problem, which we now describe. Let us assume that:

i) $p(i, j, u; \theta^0) > 0$ for all i, j, u.

ii) There is an *unique* optimal stationary control law $\phi(\cdot, \theta^0)$ for the parameter value θ^0.

The algorithm to generate the controls proceeds as follows.

SATO, ABE, TAKEDA ALGORITHM

Step 1. Set $t = 0$ and choose $\{\hat{p}_0(i, j, u)\}$, "estimates" of the transition probabilities, so that there is an unique stationary optimal control law ϕ_0 for these estimates.

Step 2. Choose $\{n_0(i, j, u)\}$ and $\{n_0(i, u)\}$, all positive, so that $\hat{p}_0(i, j, u) = n_0(i, j, u) n_0^{-1}(i, u)$ for all i, j, u.

Step 3. Choose $\gamma_0 > 0$ so small that if one defines

$$S_0(i, j, u) := \frac{n_0(i, j, u)}{n_0(i, u) + g(\gamma_0)} \quad \text{and} \quad B_0(i, j, u) := \frac{n_0(i, j, u) + g(\gamma_0)}{n_0(i, u) + g(\gamma_0)}$$

where $g(\gamma) := \gamma/(1-\gamma)$, then "for all models M satisfying $S_0 \leqq M \leqq B_0$", the policy ϕ_0 is still optimal. By a model M, we shall mean a set of transition probabilities $\{p(i, j, u)\}$ satisfying $p(i, j, u) \geqq 0$ and $\sum_j p(i, j, u) = 1$ for all i, u. By $S_0 \leqq M \leqq B_0$ we shall mean $S_0(i, j, u) \leqq p(i, j, u) \leqq B_0(i, j, u)$ for all i, j, u.

Step 4. Apply $u_t = \phi_t(x_t)$ with probability γ_t and all other elements of U with equal probability.

Step 5. Set $t = t + 1$.

Step 6. Set $n_t(i, j, u) = n_{t-1}(i, j, u) + 1(x_{t-1} = i, x_t = j, u_{t-1} = u)$ for all i, j, u and $n_t(i, u) = n_{t-1}(i, u) + 1(x_{t-1} = i, u_{t-1} = u)$ for all i, u.

Step 7. Set

$$S_t(i, j, u) := \frac{n_t(i, j, u)}{n_t(i, u) + g(\gamma_{t-1})} \quad \text{for all } i, j, u,$$

$$B_t(i, j, u) = \frac{n_t(i, j, u) + g(\gamma_{t-1})}{n_t(i, u) + g(\gamma_{t-1})} \quad \text{for all } i, j, u.$$

Step 8. If there is a single control law ϕ_t which is optimal for all models M with $S_t \leqq M \leqq B_t$, then set $\gamma_t = h(\gamma_{t-1})$ where $h(\gamma) := (1+\gamma)/2$ and go to Step 4. If a ϕ_t as above cannot be chosen, then set $\gamma_t := \gamma_{t-1}$, $\phi_t := \phi_{t-1}$ and go to Step 4.

THEOREM 7.3 (Sato, Abe, Takeda). *Assume*

i) $p(i, j, u; \theta^0) > 0$ *for all* i, j, u;

ii) *for* θ^0 *there is an* unique *stationary optimal control law* $\phi(\cdot, \theta^0)$ *for the discounted cost problem.*

Then for the above algorithm, $\lim \phi_t = \phi(\cdot, \theta^0)$ *a.s.*

Proof. First note that $\{\gamma_t\}$ is an increasing sequence. Assume that on a subset $\tilde{\Omega} \subseteq \Omega$ of the sample space of positive measure, $\gamma_t \to \gamma_\infty < 1$. By Step 4, $u_t = u$ i.o. for every u on $\tilde{\Omega}$, and by our assumptions $n_t(i, u) \to \infty$. By the law of large numbers, therefore, $n_t(i, j, u)/n_t(i, u) =: \hat{p}_t(i, j, u) \to p(i, j, u; \theta^0)$ on $\tilde{\Omega}$. This means for every $\omega \in \tilde{\Omega}$, for every $\varepsilon > 0$, there is a $T(\omega)$ large enough so that $[S_t(i, j, u), B_t(i, j, u)] \subseteq [p(i, j, u; \theta^0) - \varepsilon, p(i, j, u; \theta^0) + \varepsilon]$ for all $t \geqq T$. Now note that by (ii), *all* control laws other than $\phi(\cdot, \theta^0)$ produce a cost strictly larger than $\phi(\cdot, \theta^0)$ does. By continuity, therefore, this must also hold for all models M with $p(\cdot, \theta^0) - \varepsilon \leqq M \leqq p(\cdot, \theta^0) + \varepsilon$ for $\varepsilon > 0$ sufficiently small. Hence, for all $t \geqq$ some T', $\phi(\cdot, \theta^0)$ is optimal for all models M with $S_t \leqq M \leqq B_t$ on $\tilde{\Omega}$. Hence by Step 8, it must be the case that $\gamma_t = h(\gamma_{t-1})$ for all $t \geqq T'$ on $\tilde{\Omega}$. But this shows that $\gamma_t \to 1$ on $\tilde{\Omega}$, contradicting our original assumption. Hence we can deduce that $\gamma_t \to 1$ a.s.

For every (i, j, u), either $\lim n_t(i, j, u) = +\infty$ or not. If the former, then we have already seen a demonstration that $[S_t(i, j, u), B_t(i, j, u)] \subseteq [p(i, j, u; \theta^0) - \varepsilon, p(i, j, u; \theta^0) + \varepsilon]$ for all $t \geqq$ some T. If the latter, then since $\gamma_t \to 1$, it follows that $g(\gamma_t) \to +\infty$ and so by Step 7, $S_t(i, j, u) \to 0$ and $B_t(i, j, u) \to 1$. In either case therefore we see that $[p(i, j, u; \theta^0) - \varepsilon, p(i, j, u; \theta^0) + \varepsilon] \cap [S_t(i, j, u), B_t(i, j, u)]$ is nonempty for all $t \geqq$ some T. Choose $\varepsilon > 0$ so small that $\phi(\cdot, \theta^0)$ is the *only* stationary control law which is optimal for all models M with $p(\cdot, \theta^0) - \varepsilon \leqq M \leqq p(\cdot, \theta^0) + \varepsilon$. Then, for all $t \geqq$ some $\bar{T}$, the only stationary control law which is possibly optimal for all models M with $S_t \leqq M \leqq B_t$, is $\phi(\cdot, \theta^0)$, provided there is one such control law.

Now let ϕ^* be any limit point of $\{\phi_t\}$. Since the set of stationary control laws is finite, it follows that $\phi^* = \phi_t$ i.o., and since $\gamma(t) \to 1$, it follows that ϕ^* is optimal for all models M with $S_t \leqq M \leqq B_t$ for some $t \geqq \bar{T}$. Hence $\phi^* = \phi(\cdot, \theta^0)$, showing that $\phi_t \to \phi(\cdot, \theta^0)$ a.s. □

Another approach to the problem of generating an adaptive control policy is by using methods common in the theory of learning automata. Lyubchik and Poznyak [67] have proposed several such schemes. No proofs are provided and the identifiability issue is not alluded to. El-Fattah [68], [69] has analyzed one recursive identification and control scheme, which we now examine.

Assume that $\Theta \subseteq \mathbb{R}^n$ is a Cartesian product of closed intervals. Consider the following stochastic approximation type scheme, see Tsypkin [70], [71], for the generation of the parameter estimates:

$$\hat{\theta}_{t+1} = \hat{\theta}_t + \frac{\gamma}{t+1} \frac{\nabla(\sum_u p(x_t, x_{t+1}, u; \hat{\theta}_t)\pi_t(u|x_t))}{\sum_u p(x_t, x_{t+1}, u; \hat{\theta}_t)\pi_t(u|x_t)}.$$

Here $\pi_t(u|x_t)$ is the probability of using the control $u_t = u$ at time t, given x_t. This is a reasonable updating scheme since, at each time t, we take a step $(\hat{\theta}_{t+1} - \hat{\theta}_t)$ in the direction of the gradient, i.e., in the direction in the parameter space in which the probability of the transition from (x_t, u_t) to x_{t+1} increases most rapidly. The probabilities $\{\pi_t(u|x); u \in U, x \in X\}$ which specify the randomization scheme, are updated as follows:

$$\pi_{t+1}(u|x) = \begin{cases} \pi_t(u|x) & \text{if } x \neq x_t, \\ \pi_t(u|x) + \dfrac{a\Delta}{t+1} & \text{if } u = u_t \text{ and } x = x_t, \\ \pi_t(u|x) - \dfrac{a\Delta}{t+1} \cdot \dfrac{1}{|U|-1} & \text{if } u \neq u_t \text{ and } x = x_t \quad (|U| = \text{cardinality of } U). \end{cases}$$

Here $\Delta := f_{x_t, u_t}(\pi_t, \hat{\theta}_{t+1})$ where we have assumed that there are functions $F_x(\pi, \theta)$ and $f_{x,u}(\pi, \theta)$ satisfying $F_x(\pi, \theta) > 0$ and such that $\partial J(\pi, \theta)/\partial \pi(u|x) = F_x(\pi, \theta) f_{x,u}(\pi, \theta)$. Here $J(\pi, \theta)$ is the cost of using the stationary policy π when the parameter value is θ. Ignoring this assumption on the representation of $\partial J(\pi, \theta)/\partial \pi(u|x)$, the updating scheme for $\pi_{t+1}(u|x)$ is reasonable because it merely increases the probability of using u in state x, if this tends (infinitesimally) to reduce the cost. The term $1/(t+1)$ tends to reduce the "step-sizes" at a rate appropriate for convergence, and is standard in stochastic approximation theory.

THEOREM 7.4 (El-Fattah). *Consider the following conditions*:

i) $\Theta \subseteq \mathbb{R}^n$ *is a Cartesian product of closed intervals.*

ii) $p(i, j, u; \theta) > 0$ *for all* i, j, u, θ.

iii) *For each state* $x \in X$, *there is a possibly empty set* $S_x \subseteq \{1, 2, \cdots, n\}$ *such that*

$\bigcup_x S_x = \{1, 2, \cdots, n\}$ *and there is a* $c > 0$ *such that*

$$\sum_{k \in A} R_k^{\pi}(x, \theta)(\theta^k - \theta^{0k}) \leqq -c \sum_{k \in A} (\theta^k - \theta^{0k})^2 \quad \textit{for all } \pi, \textit{ if } A = S_x$$

$$= 0 \quad \textit{for all } \pi, \textit{ if } A = S_x^c.$$

Here

$$R^{\pi}(x, \theta) := E\left[\frac{\nabla \sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t)}{\sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t)}\middle| x_t = x, \theta^0\right].$$

iv) *There exist* $\lambda_1 > 0$, $\lambda_2 > 0$ *so that, for every* $k = 1, 2, \cdots, n$; $\operatorname{tr} K^{\pi}(x, \theta) \leqq \lambda_1 + \lambda_2 \|\theta - \theta^0\|^2$ *for all* π, *where*

$$K^{\pi}(x, \theta) := E\left[\frac{(\nabla \sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t))(\nabla^T \sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t))}{(\sum_u p(x_t, x_{t+1}, u; \theta)\pi(u|x_t))^2}\middle| x_t = x, \theta^0\right].$$

v) *There exist functions* $F_x(\pi, \theta) > 0$ *and* $f_{x,u}(\pi, \theta)$ *such that* $\partial J(\pi, \theta)/\partial \pi(u|x) = F_x(\pi, \theta) f_{x,u}(\pi, \theta)$ *for all* π, θ, u, x.

vi) *There exists* $c_1 > 0$ *such that*

$$\sum_u (f_{x,u}(\pi, \theta) - f_{x,u}(\pi, \theta^0))^2 \leqq c_1 \sum_{k \in S_x} (\theta^k - \theta^{0k})^2 \quad \textit{for all } x \in X.$$

vii)

$$\sum_u (f_{x,u}(\pi, \theta))^2 \pi(u|x) \geqq \lambda > 0 \quad \textit{for all } x \in X.$$

Then:

viii) *If* (i)-(iv) *hold, then for* <u>*any*</u> *nonanticipative randomized policy,*

$$\lim \hat{\theta}_t = \theta^0 \quad a.s. \quad \textit{and} \quad \sum_t \frac{1}{t+1} \sum_{k \in S_{x_t}} (\hat{\theta}_t^k - \theta^{0k})^2 < \infty \quad a.s.$$

ix) *If* (i)-(vii) *hold, then the adaptive scheme above is such that* $\lim (1/N) \sum_1^N c(x_t, u_t) = J(\theta^0)$ *a.s.*

Proof. Define the "stochastic Lyapunov function", $V_t := \|\hat{\theta}_t - \theta^0\|^2$. By calculation, we obtain (with $\mathcal{F}_t := \sigma(x_0, u_0, x_1, \cdots, x_t)$)

$$E(V_{t+1}|\mathcal{F}_t) = V_t + \frac{2\gamma}{t+1} \sum_{k \in S_{x_t}} (\hat{\theta}_t^k - \theta^{0k}) R_k^{\pi_t}(x_t, \hat{\theta}_t) + \frac{\gamma^2}{(t+1)^2} \operatorname{tr} K^{\pi_t}(x_t, \hat{\theta}_t)$$

$$\leqq V_t\left(1 + \frac{\gamma^2 \lambda_2}{(t+1)^2}\right) + \frac{\lambda_1 \gamma^2}{(t+1)^2} - \frac{2\gamma c}{(t+1)} \sum_{k \in S_{x_t}} (\hat{\theta}_t^k - \theta^{0k})^2.$$

Hence, $\{V_t, \mathcal{F}_t\}$ is "nearly a positive supermartingale" in the sense of Neveu [72] or Robbins and Siegmund [73]. Applying the convergence theorem for such, we deduce that $\{V_t\}$ converges a.s. and $\sum_t 1/(t+1) \sum_{k \in S_{x_t}} (\hat{\theta}_t^k - \theta^{0k})^2 < \infty$ a.s. By positive recurrence of $\{x_t = x\}$ for every x, [68] deduces that $\liminf \sum_{k \in S_x} (\hat{\theta}_t^k - \theta^{0k})^2 = 0$ a.s. for every $x \in X$. Then $\liminf \|\hat{\theta}_t - \theta^0\|^2 = 0$ a.s. also follows, and since $\{V_t\}$ converges, shows that $V_t \to 0$ a.s. This completes the proof of part (viii). The proof of part (ix), being algebraically tedious is omitted; the reader being referred to [68]. □

It is clear that (viii) is a strong result which shows that under the conditions (i)-(iv) one can identify θ^0 under *any* policy. Thus, it is clear that even though the assumptions (iii) and (iv) are not identical to Mandl's condition [55], they are nevertheless *restrictive.* (iii) is the "pseudo-gradient" condition of Polyak and Tsypkin [74] and

guarantees that the *expected value* of the step $\hat{\theta}_{t+1} - \hat{\theta}_t$ forms an "acute angle" with the *desired* direction $\theta^0 - \hat{\theta}_t$. The strict negativity for $A = S_x$ in (iii) appears therefore to play a role analogous to that of an Identifiability Condition.

The proof used here is remarkably similar to the proof of Goodwin, Ramadge and Caines [75] in their treatment of their version of the self-tuning regulator. This only serves to illustrate how closely connected all these problems are.

7.3. The cost biased maximum likelihood method. This approach requires no restrictive identifiability condition of any sort and is motivated by the following argument. Without the imposition of identifiability conditions of any sort, many reasonable parameter identification schemes will provide a limit θ^* of $\{\hat{\theta}_t\}$ which satisfies $p(i, j, \phi(i, \theta^*); \theta^*) = p(i, j, \phi(i, \theta^*); \theta^0)$ (Theorem 7.1 for example). This has an immediate consequence, to see which we define

$\pi(i, \phi, \theta) :=$ steady state probability of i when control law ϕ is used in θ (for simplicity we assume $p(i, j, u; \theta) > 0$);

$J(\phi, \theta) :=$ long-term average cost of using ϕ in θ;

$\phi_\theta :=$ optimal control law for θ;

$J(\theta) :=$ optimal long-term average cost for θ.

Now,

$$\begin{aligned} J(\theta^*) &= J(\phi_{\theta^*}, \theta^*) \\ &= \sum_{i,j} \pi(i, \phi_{\theta^*}, \theta^*) p(i, j, \phi(i, \theta^*), \theta^*) c(i, \phi(i, \theta^*)) \\ &= \sum_{i,j} \pi(i, \phi_{\theta^*}, \theta^0) p(i, j, \phi(i, \theta^*), \theta^0) c(i, \phi(i, \theta^*)) \\ &= J(\phi_{\theta^*}, \theta^0) \\ &\geqq J(\theta^0). \end{aligned}$$

Here we have used the well-known, see [33], representation of the long-term average cost in terms of the steady state probabilities *and* the implication $\{p(i, j, \phi(i, \theta^*); \theta^*) = p(i, j, \phi(i, \theta^*); \theta^0)$ for all $i, j\} \Rightarrow \{\pi(i, \phi_{\theta^*}, \theta^*) = \pi(i, \phi_{\theta^*}, \theta^0)$ for all $i\}$. So $\hat{\theta}_t \to \{\theta: J(\theta) \geqq J(\theta^0)\}$. To capitalize on this, one can "bias" the parameter identification scheme in "favour" of parameters θ for which $J(\theta)$ is small, knowing of course that θ^0 has the smallest value $J(\theta^0)$ in $\{\theta: J(\theta) \geqq J(\theta^0)\}$. However, this biasing has to be done delicately, so that we do *not* destroy the parameter identification scheme itself. This is done below.

THEOREM 7.5 (Kumar and Becker). *Let*

i) $p(i, j, u; \theta) > 0$ *for all* i, j, u, θ;

ii) $c(i, u) > 0$ *for all* i, u;

iii) Θ *be finite*;

iv) $o(t)$ *be such that* $\lim o(t) = +\infty$ *and* $\lim t^{-1} o(t) = 0$. *Choose* $\hat{\theta}_t$ *and* u_t *so that*

$$\hat{\theta}_t = \begin{cases} \arg\max_\theta J(\theta)^{-o(t)} \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \theta) & \text{for } t = 0, 2, 4, 6, \cdots, \\ \hat{\theta}_{t-1} & \text{for } t = 1, 3, 5, \cdots \end{cases}$$

and $u_t = \phi(x_t, \hat{\theta}_t)$. (*If there is more than one maximizer above, choose one according to some prespecified priority order on* Θ.) *Then*:

v) $\lim (1/N) \sum_1^N c(x_t, u_t) = J(\theta^0)$ *a.s.*

vi) $\lim (1/N)\sum_1^N 1(\hat{\theta}_t = \theta^*) = 1$ *for some* θ^* *a.s.*

vii) $p(i, j, \phi(i, \theta^*), \theta^*) = p(i, j, \phi(i, \theta^*), \theta^0)$ *a.s.*

viii) ϕ_{θ^*} *is optimal for* θ^0 *a.s.*

Proof. Since $\hat{\theta}_t$ is a maximizer of the criterion according to which it is chosen,

$$J(\hat{\theta}_t)^{-o(t)} \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \hat{\theta}_t) \geqq J(\theta^0)^{-o(t)} \prod_{s=0}^{t-1} p(x_s, x_{s+1}, u_s; \theta^0).$$

Taking logarithms,

$$\frac{o(t)}{t} \log\left(\frac{J(\theta^0)}{J(\hat{\theta}_t)}\right) + \frac{1}{t}\sum_{s=0}^{t-1} \log \frac{p(x_s, x_{s+1}, u_s; \hat{\theta}_t)}{p(x_s, x_{s+1}, u_s; \theta^0)} \geqq 0.$$

Hence, if $\hat{\theta}_t = \theta^*$ i.o., for some $\omega \in \Omega$ the sample space, it follows that

$$\liminf \frac{1}{N}\sum_0^{N-1} \log \frac{p(x_s, x_{s+1}, u_s; \theta^*)}{p(x_s, x_{s+1}, u_s; \theta^0)} \geqq 0.$$

By the Martingale Stability Theorem, it follows that for every $\theta \in \Theta$,

$$\lim \frac{1}{t}\sum_0^{t-1} \frac{p(x_s, x_{s+1}, u_s; \theta)}{p(x_s, x_{s+1}, u_s; \theta^0)} = 1 \quad \text{a.s.}$$

Putting these two facts together and making use of the positive recurrence of each event $(x_t = i)$, it follows that a.s. $p(i, j, \phi(i, \theta^*); \theta^*) = p(i, j, \phi(i, \theta^*); \theta^0)$ whenever $\limsup (1/N)\sum_0^{N-1} 1(\hat{\theta}_t = \theta^*) > 0$. On the other hand,

$$\left(\frac{J(\theta^0)}{J(\hat{\theta}_t)}\right)^{o(t)} \prod_{s=0}^{t-1} \frac{p(x_s, x_{s+1}, u_s; \hat{\theta}_t)}{p(x_s, x_{s+1}, u_s; \theta^0)} \geqq 1.$$

Since we know that

$$\left\{\prod_0^{t-1} \frac{p(x_s, x_{s+1}, u_s; \theta)}{p(x_s, x_{s+1}, u_s; \theta^0)}\right\}$$

is a positive martingale and therefore converges a.s., we see that if $\hat{\theta}_t = \theta^*$ i.o., then $J(\theta^0) \geqq J(\theta^*)$. Together with the argument preceding this theorem, this shows that $\limsup (1/N)\sum_0^{N-1} 1(\hat{\theta}_t = \theta^*) > 0$ implies $\phi(\cdot, \theta^*)$ is optimal for θ^0. The argument in Theorem 7.2 can now be used to show (v). (vi), (vii) and (viii) will follow if we can show that a.s. θ^* is unique. This is a consequence of the priority ordering of elements of Θ. The full details are in [51]. □

The unique feature of the adaptive control scheme given here is that without resorting to either forced choices or randomization, it can attain optimal performance. Condition (ii) of the theorem is unnecessary and condition (i) can be replaced by a positive recurrence condition, see [51].

In [76] the above result has been generalized to the situation where Θ is not finite, but is the class of *all* models. In [77], the generalization to the case where X and U are Borel spaces is developed. The case of linear systems with quadratic cost criteria is analyzed in [78] when the parameter set Θ is finite. In this situation, an additional complexity is to prove that the system is stable as well, and this is also done in [78].

If Θ is not finite, then implementation of schemes of this type will depend on the availability of methods to perform the requisite on-line computations. In the general case such methods are not yet available. In some specific situations however, see [79], explicit solutions can be obtained.

The above approach can also be used for discounted (or other, for example, finite horizon) cost problems to obtain adaptive policies which converge in a Cesaro sense to the optimal policy. This is done in [77].

Before ending the topic of this section, we note the early work of Riordan [80] who has also simulated the control of a heat treatment process by modeling it as a problem in the adaptive control of Markov chains.

8. Self-tuning regulators. This class of non-Bayesian adaptive control problems (NACP's) has recently enjoyed much practical success. See [81]–[85] for a sampling of some of the applications culled from the last two years' issues of one journal.

The basic problem can be posed as follows. There is a system

$$y(t+1)=\sum_{i=0}^{n} a_i y(t-i)+\sum_{i=0}^{p} b_i u(t-i-d+1)+\sum_{i=0}^{m} c_i w(t-i)+w(t+1).$$

Here u is the input, y the output and $\{w(t)\}$ is a "white" noise sequence (defined more precisely later on). $d \geqq 1$ is called the *delay*, and is assumed to be known. The problem is that the coefficients $\{a_i, b_i, c_i\}$ of the linear system are unknown. In spite of this, the goal is to choose $u(t)$ based on $(y(0), u(0), y(1), u(1), \cdots, y(t))$, for each t, so that $\lim (1/N)\sum_1^N y^2(t)$—the sample path variance of the output—is almost surely a minimum. As is clear, this is a standard NACP of the sort defined in § 6.

To obtain a full appreciation of this problem, however, it is necessary first to understand the complexities involved in a) *identifying* the coefficients of an Auto Regressive Moving Average System with EXogeneous Inputs (ARMAX system) and b) the problem of *controlling* such processes. Accordingly, we first take up these two issues.

8.1. Least squares estimation of coefficients. We shall *very* briefly explain the various issues involved in identifying the coefficients of an ARMAX model.

8.1.1. Suppose that we have available a finite sequence $\{y(-n), y(-n+1), \cdots, y(-1)\} \cup \{y(0), y(1), \cdots, y(N)\}$ and we wish to model this sequence by a relationship of the form $y(t+1) \approx \alpha_0 y(t)+\alpha_1 y(t-1)+\cdots+\alpha_n y(t-n)$ for $t=0, 1, 2, \cdots, N-1$. The question is: What are the "best" coefficients $(\alpha_0, \alpha_1, \cdots, \alpha_n)$ to choose? One way, of course, is to select them so that they minimize $\sum_1^N (y(t)-\alpha_0 y(t-1)-\alpha_1 y(t-2)-\cdots-\alpha_n y(t-n-1))^2$. This is a standard "curve fitting" sort of procedure which seeks to minimize the sum of the squares of the errors of "fit".

Note that this method of fitting a model to data is *totally nonprobabilistic.*

One way of solving the minimization problem is to define $\phi_{t-1} := (y_{t-1}, y_{t-2}, \cdots, y_{t-n-1})^T$ and $\theta := (\alpha_0, \alpha_1, \cdots, \alpha_n)^T$. (We use both $y(t)$ and y_t interchangeably.) Then the goal is to choose $\theta \in \mathbb{R}^{n+1}$ to minimize $\|(y_1, y_2, \cdots, y_N)-\theta^T(\phi_0, \phi_1, \cdots, \phi_{N-1})\|^2$. A standard application of the projection theorem shows that the minimizing θ, denoted by $\hat{\theta}_N$, is given by

$$\begin{aligned}\hat{\theta}_N &= ((\phi_0, \phi_1, \cdots, \phi_{N-1})(\phi_0, \phi_1, \cdots, \phi_{N-1})^T)^{-1} \\ &\quad \cdot (\phi_0, \phi_1, \cdots, \phi_{N-1})(y_1, y_2, \cdots, y_N)^T \\ &= \left(\sum_0^{N-1} \phi_i \phi_i^T\right)^{-1} \sum_0^{N-1} \phi_i y_{i+1} \\ &= \left(\frac{1}{N}\sum_0^{N-1} \phi_i \phi_i^T\right)^{-1} \left(\frac{1}{N}\sum_0^{N-1} \phi_i y_{i+1}\right).\end{aligned}$$

This will be called the "least squares estimate".

8.1.2. Suppose one more piece of data y_{N+1} becomes available; then one can obtain another estimate $\hat{\theta}_{N+1}$ which minimizes $\sum_1^{N+1} (y_t-\alpha_0 y_{t-1}-\cdots-\alpha_n y_{t-n-1})^2$. The

relationship between $\hat{\theta}_{N+1}$ and $\hat{\theta}_N$ can be written in a recursive manner as

$$\hat{\theta}_{N+1} = \hat{\theta}_N + R_N^{-1}\phi_N(y_{N+1} - \hat{\theta}_N^T\phi_N)$$

where $R_N := \sum_{i=0}^{N} \phi_i\phi_i^T$.

One can also obtain a recursive expression for R_N^{-1} as

$$R_{N+1}^{-1} = R_N^{-1} - \frac{R_N^{-1}\phi_{N+1}\phi_{N+1}^T R_N^{-1}}{1 + \phi_{N+1}^T R_N^{-1}\phi_{N+1}}.$$

This is done by using the well-known Matrix Inversion Lemma.

8.1.3. So far we have only examined the question of fitting a model to data, without mentioning in *any way whatsoever*, how the data were generated in the first place. Let us suppose that the sequence $\{y_t\}$ is actually generated by the *system*

$$y_{t+1} = a_0 y_t + a_1 y_{t-1} + \cdots + a_n y_{t-n} + v_t$$

where $\{v_t\}$ is "white" noise.

By employing the ergodic theorem (assuming all roots of the polynomial $1 - a_0 z - a_1 z^2 - \cdots - a_n z^{n+1}$ are strictly outside the unit circle) it can be seen that

$$\left(\lim \frac{1}{N} \sum_{i=0}^{N-1} \phi_i\phi_i^T\right)_{jk} = r_{|j-k|} \quad \text{where } r_j = E(y_t y_{t-j}) \text{ in steady state.}$$

Also

$$\lim \frac{1}{N} \sum_{0}^{N-1} \phi_i y_{i+1} = (r_1, r_2, \cdots, r_{n+1})^T.$$

Now we can examine the *probabilistic* behaviour of the sequence $\{\hat{\theta}_N\}$. From the above it follows that

$$\lim \hat{\theta}_N = [r_{|j-k|}]^{-1}(r_1, r_2, \cdots, r_{n+1})^T.$$

Since $y_{t-j}y_{t+1} = \sum_{i=0}^{n} a_i y_{t-j}y_{t-i} + y_{t-j}v_{t+1}$, taking expectations, gives $r_{j+1} = \sum_{i=0}^{n} a_i r_{j-i}$ for $j \geqq 0$. Hence $(r_1, \cdots, r_{n+1})^T = [r_{|j-k|}](a_0, \cdots, a_n)^T$ and so $(a_0, \cdots, a_n)^T = [r_{|j-k|}]^{-1}(r_1, \cdots, r_{n+1})^T$. This shows that $\lim \hat{\theta}_N = (a_0, a_1, \cdots, a_n)^T$ a.s.

To summarize, if the noise entering into the system (here v_t) is "white", then least squares estimates of the parameters $\{\hat{\theta}_N\}$ are consistent.

For a general treatment of this, the reader is referred to Lai and Wei [86].

8.1.4. Crucial use was made in the above of the fact that $\{v_t\}$ was a "white" noise. Suppose now that it is *not* white; then, in general, $Ey_{t-j}v_{t+1} \neq 0$ for $j \geqq 0$. It then follows, that $\hat{\theta}_N \to (a_0, a_1, \cdots, a_n)^T + \Delta$ where $\Delta \neq 0$. See Goodwin and Payne [87] for an example.

Now if $v_{t+1} = w_{t+1} + \sum_{i=0}^{m} c_i w_{t-i}$ where $\{w_t\}$ is "white", then $\{v_t\}$ is a moving average of white noise and is *not* itself white. Thus if one uses the least squares procedure to obtain estimates $\{\hat{\theta}_N\}$ of the coefficients $(a_0, a_1, \cdots, a_n)$ in the system

$$y(t+1) = \sum_{i=0}^{n} a_i y(t-i) + \sum_{i=0}^{m} c_i w(t-i) + w(t+1),$$

then the least squares estimates are asymptotically *biased.*

8.1.5. For a system of the type just mentioned one might also be interested in estimating the coefficients $(c_0, c_1, \cdots, c_m)$ in *addition* to the coefficients $(a_0, a_1, \cdots, a_n)$.

If one had access to $(y_0, y_1, \cdots, y_N)$ *and* $(w_0, w_1, \cdots, w_N)$, then one could form asymptotically (as $N \to \infty$) consistent estimates $(\alpha_0, \cdots, \alpha_n, \gamma_0, \cdots, \gamma_m)$ by minimizing

$$\sum_{t=1}^{N} (y_t - \alpha_0 y_{t-1} - \cdots - \alpha_n y_{t-n-1} - \gamma_0 w_{t-1} - \gamma_1 w_{t-2} - \cdots - \gamma_m w_{t-m-1})^2.$$

However $\{w_t\}$ is *not* generally available to observe. It is merely an innovations *representation* of the coloured noise $v_t = w_t + c_0 w_{t-1} + \cdots + c_m w_{t-m-1}$ in the system. But one could *estimate* w_t by $y_t - \phi_{t-1}\hat{\theta}_{t-1}$ or $y_t - \phi_{t-1}^T \hat{\theta}_t$, and use these in place of the true values of w_t.

A scheme of this *general* sort is examined by Solo [88]. It is shown that a sufficient condition to obtain strongly consistent estimates is that the polynomial $C(z) = 1 + c_0 z + \cdots + c_m z^{m+1}$ satisfies the condition

$$\operatorname{Re}\left[\frac{1}{C(e^{i\omega})} - \frac{1}{2}\right] > 0 \quad \text{for all } \omega.$$

(Here $i = \sqrt{-1}$.) Note that since $v_{t+1} = w_{t+1} + c_0 w_t + \cdots + c_m w_{t-m}$ is just a *representation* of $\{v_t\}$, we can assume without loss of generality that all roots of $C(z)$ are outside the unit circle, see Åström [89].

Conditions of the type $\operatorname{Re}(C e^{i\omega}) > \varepsilon$, called Positive Real Conditions, occur in other analyses of identification algorithms also, see Solo [148]. Proofs, both in identification and adaptive control, as we shall see in §§ 8.4 and 8.5, have been built around such conditions. To this author, however, the full extent of their role is not completely clear: some insights, however, are offered by Ljung [94]. It should be noted that the Positive Real Condition also plays a role in deterministic adaptive control, where it is used via the concept of "hyperstability".

8.1.6. Suppose that we also have control inputs $\{u_t\}$ in the system, i.e.,

$$y(t+1) = \sum_{i=0}^{n} a_i y(t-i) + \sum_{i=0}^{p} b_i u(t-d-i+1) + \sum_{i=0}^{m} c_i w(t-i) + w(t+1).$$

In such a case, besides estimating $\{a_i, c_i\}$ we may also wish to estimate $\{b_i\}$.

Clearly, if $u_t = 0$ for all t, then one could not possibly identify $\{b_0, \cdots, b_p\}$. Thus, some "excitation" conditions have to be imposed on $\{u_t\}$ in order to guarantee asymptotic consistency of the estimates of $\{a_0, \cdots, a_n, b_0, \cdots, b_p, c_0, \cdots, c_m\}$. These *sufficiency* conditions are usually of the form

$$\lim \frac{1}{N} \sum_{1}^{N} \phi(t)\phi^T(t) \quad \text{is positive define}$$

where $\phi(t) := (y(t), \cdots, y(t-n), u(t-d+1), \cdots, u(t-d-p+1), w(t), \cdots, w(t-m))^T$ and are called "persistency of excitation" conditions, see Åström [90] and Solo [88].

8.2. Minimum variance control of an ARMAX model. We now examine the problem of *control* of an ARMAX model, when the coefficients $\{a_i, b_i, c_i\}$ are *known*. We shall proceed by examining a sequence of special cases until we obtain full generality. Only the fully general case is proved; for the special cases we provide informal arguments which are quick and easy to understand.

8.2.1. For the system

$$y_{t+1} = a_0 y_t + \cdots + a_n y_{t-n} + b_0 u_t + \cdots + b_p u_{t-p} + w_{t+1}$$

where $\{w_t\}$ is "white" noise, it is clear that the control law

$$u_t = \frac{-1}{b_0}(a_0 y_t + \cdots + a_n y_{t-n} + b_1 u_{t-1} + \cdots + b_p u_{t-p})$$

minimizes the sample path variance $\lim (1/N) \sum_1^N y^2(t)$. For then, $y_{t+1} = w_{t+1}$ and so

$$\lim \frac{1}{N} \sum_1^N y_t^2 = \lim \frac{1}{N} \sum_1^N w_t^2 = \sigma^2 \quad \text{a.s.}$$

the best achievable, where $\sigma^2 := E(w_t^2)$.

8.2.2. One special feature of the above is that the noise is "white". Suppose now that the system is

$$y(t+1) = a_0 y_t + \cdots + a_n y_{t-n} + b_0 u_t + \cdots + b_p u_{t-p} + w_{t+1} + c_0 w_t + \cdots + c_n w_{t-n}.$$

If we *could* use $u_t = (-1/b_0)(a_0 y_t + \cdots + a_n y_{t-n} + b_1 u_{t-1} + \cdots + b_p u_{t-p} + c_0 w_t + \cdots + c_n w_{t-n})$, then this is clearly the best achievable, since then $y_{t+1} = w_{t+1}$ and so $\lim (1/N) \sum_1^N y_t^2 = \sigma^2$ a.s. However $\{w_t\}$ is *not* accessible, and so we replace it by $\{y_t\}$, which is what it *would* be if the above control law could be implemented. This gives

$$u_t = \frac{-1}{b_0}[(a_0 + c_0) y_t + \cdots + (a_n + c_n) y_{t-n} + b_1 u_{t-1} + \cdots + b_p u_{t-p}]$$

the optimal control law.

It is important to note that in order to implement this control law, one needs knowledge only of $\{a_i + c_i, b_i\}$ and *not* $\{a_i, c_i, b_i\}$ separately.

8.2.3. The special feature of the above model is that the delay d is *exactly* one unit. Consider now the general case $d \geqq 1$. The system is

$$y_{t+1} = a_0 y_t + \cdots + a_n y_{t-n} + b_0 u_{t-d+1} + \cdots + b_p u_{t-d-p+1} + w_{t+1} + c_0 w_t + \cdots + c_n w_{t-n}.$$

This is more conveniently represented in the "polynomial" format

$$A(z) y_{t+1} = z^d B(z) u_{t+1} + C(z) w_{t+1}$$

where $A(z) := 1 - a_0 z - \cdots - a_n z^{n+1}$; $B(z) = b_0 + b_1 z + \cdots + b_p z^p$; $C(z) = 1 + c_0 z + \cdots + c_n z^{n+1}$, and z is the *backward shift operator* $z y_t := y_{t-1}$. (Note that often in the literature z^{-1} is the backward shift operator; but we find this more convenient.) Let us *divide* the polynomial C by the polynomial A, carrying out d steps of the long division process to get

$$C = AF + z^d G$$

where $F(z) = 1 + f_1 z + \cdots + f_{d-1} z^{d-1}$ is the "quotient" and $G(z) = g_0 + g_1 z + \cdots + g_n z^n$. Now we can formally represent the system as

$$y_{t+1} = z^d \frac{B}{A} u_{t+1} + \frac{C}{A} w_{t+1} = z^d \left(\frac{B}{A} u_{t+1} + \frac{G}{A} w_{t+1} \right) + F w_{t+1}.$$

Clearly u_{t+1-d} cannot be based on the "future" $(w_{t+1}, \cdots, w_{t+2-d})$. Thus one chooses $u_{t+1} = (-G/B) w_{t+1}$. This results in $y_{t+1} = F(z) w_{t+1}$, thus giving us the optimal control

law

$$u_t = \frac{-1}{BF} y_t.$$

Further, the minimum output variance is $\lim (1/N) \sum_1^N y_t^2 = (1 + f_1^2 + \cdots + f_{d-1}^2)\sigma^2$. For more details, the reader is referred to Åström [89].

8.2.4. There is one potential "practical" flaw with this procedure. To see the nature of this, consider the very simple example

$$y_{t+1} = -y_t + u_t - 2u_{t-1} + w_{t+1}.$$

Applying the procedure of § 8.2.1, which is but a special case of § 8.2.3, shows that the optimal control law is

$$u_t = y_t + 2u_{t-1}.$$

Under this optimal control law, $y_{t+1} = w_{t+1}$ and so the sequence of applied controls is $u_t = 2u_{t-1} + w_t$. This is clearly an *unstable* difference equation, in view of the coefficient 2. Thus, although theoretically this is an optimal control law, from a practical point of view this is *unacceptable.* One reason is that if the true system is actually $y_{t+1} = -y_t + u_t - (2+\varepsilon)u_{t-1} + w_{t+1}$ where ε is very small, then after the control law $u_t = 2u_{t-1} + w_t$ is applied, the closed loop system is $y_{t+1} = \varepsilon u_{t-1} + w_{t+1}$ which with $u_t = 2u_{t-1} + y_t$ is an *unstable* system. In any case, u_t is an "exploding" sequence.

The source of this instability is that the polynomial $B(z) = 1 - 2z$ does *not* have all its roots strictly outside the unit circle. This added requirement that the polynomial $B(z)$ have all its roots strictly outside the unit circle is called a *minimum phase condition,* and should necessarily be imposed if the *optimal* control law of § 8.2.3 is to be useful *practically.*

8.2.5. Thus, we see that in general the true practical problem at hand is *not* to minimize $\lim (1/N) \sum_1^N y_t^2$ *unconditionally,* but to minimize it *conditionally* subject to the *constraint* that the closed loop system is stable.

THEOREM 8.1 (Peterka). *Let* i) $C(z)$ *have no roots on or inside the unit circle*;
ii) $b_p \neq 0$.

Then the control law which minimizes the output variance subject to the stability requirement above is

$$u_t = -\frac{S(z)}{R(z)} y_t$$

where $S(z)$ *and* $R(z)$ *are polynomials determined by the polynomial equation*

$$B^* C = RA + z^d BS$$

with $\deg R = p + d - 1$. *Here*

$B^+ :=$ *factor of* B *containing all roots of* B *lying outside or on the unit circle normalized so that* $B^+(0) = 1$;

$B^- :=$ *factor of* B *containing all roots of* B *lying inside the unit circle and satisfying* $B^+ B^- = B$;

$$\tilde{B}^- := \frac{1}{b_p} z^p B^-(z^{-1});$$

$$B^* := B^+ \tilde{B}^-.$$

Proof. For any polynomial P, define P^+, P^-, P^* as above and also $\tilde{P} := (1/p_k)z^k P(z^{-1})$, $\bar{P} := P(z^{-1})$. Note that a general linear system can be represented as $y_t = z^d(b(z)/a(z))u_t + (\beta(z)/\alpha(z))w_t$ where $\beta = \beta^+$, $\alpha = \alpha^+$. Now this can be rewritten as $a\alpha y_t = z^d b\alpha u_t + a\beta w_t$, which in turn is the same as $a\alpha y_t = z^d b\alpha u_t + a^*\beta w_t$. So we shall assume for our purposes that $A = a\alpha$, $B = b\alpha$ and $C = a^*\beta$ where $\alpha = \alpha^+$, $\beta = \beta^+$, $\alpha(0) = \beta(0) = a(0) = 1$. Now if a control law $u_t = -(s/r)y_t$ is applied, then $\operatorname{var}(y_t) = \sigma^2(1/2\pi i)\oint W(z)W(z^{-1})\,dz/z$ where $W = (\beta/\alpha)ar/(ar + z^d bs)$. One can rewrite W as

$$W = \frac{z^d a^- b^-}{\tilde{a}^-\tilde{b}^-}\left(\frac{\tilde{a}^-\tilde{b}^-\beta}{z^d a^- b^-\alpha} - \frac{\tilde{a}^-\beta b^* s}{a^-\alpha(ar + z^d bs)}\right).$$

A partial fraction expansion gives,

$$\frac{\tilde{a}^-\tilde{b}^-\beta}{z^d a^- b^-\alpha} = \frac{p}{z^d b^-} + \frac{q}{a^-\alpha}$$

where $\deg p = (\deg b^-) + d - 1$, $p(0) = 1$. This gives the polynomial equation $\tilde{a}^-\tilde{b}^-\beta = a^-\alpha p + z^d b^- q$. Now W can be rewritten as

$$W = \frac{z^d a^- b^-}{\tilde{a}^-\tilde{b}^-}\left(\frac{p}{z^d b^-} + \psi\right)$$

which defines ψ. ψ can be rewritten as

$$\psi = \frac{a^+ qr - \alpha p b^+ s}{\alpha(ar + z^d bs)}.$$

Now we see that

$$\operatorname{var}(y_t) = \text{const}\left[\frac{1}{2\pi i}\oint \frac{p\bar{p}}{b^-\bar{b}^-}\frac{dz}{z} + \frac{1}{2\pi i}\oint \psi\bar{\psi}\frac{dz}{z} + \frac{2}{2\pi i}\oint \frac{\bar{p}}{z^{-d}\bar{b}^-}\psi\frac{dz}{z}\right]$$

where const $:= (a^- b^-/\tilde{a}^-\tilde{b}^-)(\bar{a}^-\bar{b}^-/\bar{\tilde{a}}^-\bar{\tilde{b}}^-)$. The point of all this algebraic manipulation is that, for stability, the denominator of ψ must have its roots strictly outside the unit circle, i.e., ψ must be holomorphic inside and on the unit circle. This makes the last integral in the above expression zero. The first integral is unaffected by the choice of the control polynomials r and s. Thus, to minimize $\operatorname{var}(y_t)$ subject to the stability constraint, the best one can do is to make $\psi = 0$, which happens when $s/r = a^+ q/\alpha p b^+$. This can be further simplified if one notes that by multiplying the polynomial equation determining p and q by $a^+ b^+$, we get $a^* b^*\beta = a\alpha p b^+ + z^d b a^+ q$. After defining $v = pb^+$, this simplifies to $a^* p^*\beta = a\alpha v + z^d bs$ and $r = \alpha v$. Multiplying by α, and using the original notation gives $CB^* = Ar + z^d Bs$, and the optimal control law is as claimed. □

Recently, [138] has also obtained the stable control laws which minimize the variance for general multivariable ARMAX systems.

8.3. The self-tuning regulator. We now take up for consideration the self-tuning regulator. Given a finite sequence of data $\{y_0, u_0, y_1, u_1, \cdots, y_N\}$ we shall first fit a model of the form

$$y_{t+1} \approx \alpha_0 y_{t-d+1} + \cdots + \alpha_l y_{t-l} + \beta_0 u_{t-d+1} + \cdots + \beta_m u_{t-d-m+1} \quad \text{for some } d \geqq 1$$

by minimizing $\sum_{t=0}^{N-1}(y_{t+1} - \alpha_0 y_{t-d+1} - \cdots - \alpha_l y_{t-l-d+1} - \beta_0 u_{t-d+1} - \cdots - \beta_m u_{t-d-m+1})^2$ over $(\alpha_0, \cdots, \alpha_l, \beta_0, \cdots, \beta_m)^T \in \mathbb{R}^{l+m+2}$. Denoting the minimizer by $\hat{\theta}_N$, we know that it can be written in the recursive form

$$\hat{\theta}_N = \hat{\theta}_{N-1} + R_{N-1}^{-1}\phi_{N-1}(y_N - \hat{\theta}_{N-1}^T\phi_{N-d})$$

where $R_N = R_{N-1} + \phi_N \phi_N^T$ and $\phi_N := (-y_N, -y_{N-1}, \cdots, -y_{N-l}, u_N, \cdots, u_{N-m})^T$. (The scheme of Åström and Wittenmark [92] fixes β_0 arbitrarily and minimizes the least-squares criterion subject to this constraint.) Then the control input u_N is chosen so that

$$u_N = \frac{1}{\hat{\beta}_0}[\hat{\alpha}_0 y_N + \cdots + \hat{\alpha}_l y_{N-l} + \hat{\beta}_1 u_{N-1} + \cdots + \hat{\beta}_m u_{N-m}]$$

where $\hat{\theta}_N =: (\hat{\alpha}_0, \cdots, \hat{\alpha}_l, \hat{\beta}_0, \cdots, \hat{\beta}_m)$. Equivalently u_N is chosen so as to make $\phi_N^T \hat{\theta}_N = 0$.

This is a recursive scheme which operates strictly off incoming data, and in "real time". From data $\{y_0, u_0, \cdots, y_N\}$ a control u_N is calculated. This control input u_N is then applied to some system; it does not matter for the present what it is. The main feature is that some output y_{N+1} is obtained from the system. This gives the enlarged data set $\{y_0, \cdots, y_N, u_N, y_{N+1}\}$ from which u_{N+1} is calculated and then applied to the system etc.

8.3.1. The above self-tuning regulator (STR) scheme can be used to control any system. The question is: For what classes of systems is the scheme asymptotically optimal? Clearly, the control law we are applying is of the type $u_t = \sum_{i=1}^{m} h_i u_{t-i} + \sum_{i=0}^{l} g_i y_{t-i}$ and so it is clear that one must only consider systems for which control laws of the above type (and with the given orders) are optimal. A natural candidate class of systems is those of the type surveyed in § 8.2.3 with the appropriate orders. Thus we consider systems which are of the form

$$y_{t+1} = a_0 y_t + \cdots + a_n y_{t-n} + b_0 u_{t-d+1} + \cdots + b_p u_{t-d-p+1} + w_{t+1} + c_0 w_t + \cdots + c_n w_{t-n}$$

where $p \leqq m - d + 1$, $n \leqq l$.

8.3.2. Åström and Wittenmark [92] were, apparently, the first to attempt an analysis of the STR scheme when it is applied to the above system.

The basic contention of [92] is that the vectors $\{\hat{\theta}_N\}$ *cannot* converge to *arbitrary* values. If $\{\hat{\theta}_N\}$ converges, it can *only* converge to values which result in a feedback control law which is *optimal* for the true system.

We now show the arguments used in [92] on an example—where the arguments used are most transparent. The general case proceeds along *exactly* the same lines.

Generic example 8.2 (Åström and Wittenmark). The true system is $y(t) - a_0 y(t-1) = b_0 u(t-2) + b_1 u(t-3) + w(t) + c_0 w(t-1)$. We choose estimates $(\hat{\alpha}(N), \hat{\beta}_1(N), \hat{\beta}_2(N))$ which minimize

$$\sum_{t=1}^{N} (y(t) - \alpha y(t-2) - \beta_0 u(t-2) - \beta_0 \beta_1 u(t-3) - \beta_0 \beta_2 u(t-4))^2$$

over all $(\alpha, \beta_1, \beta_2)$. Here β_0 is prechosen and fixed (this makes it slightly different from the scheme above). Then u_N is chosen as

$$u(N) = -\frac{\hat{\alpha}(N)}{\beta_0} y(N) - \hat{\beta}_1(N) u(N-1) - \hat{\beta}_2(N) u(N-2).$$

Now we examine the possible limits of $\{(\hat{\alpha}(N), \hat{\beta}_1(N), \hat{\beta}_2(N))\}$. If $\{(\hat{\alpha}(N), \hat{\beta}_1(N), \hat{\beta}_2(N))\}$ does converge to $(\alpha, \beta_1, \beta_2)$ (say), then asymptotically we

will have, approximately,

$$\lim \frac{1}{N}\sum_1^N \begin{pmatrix} y_i^2 & \beta_0 y_i u_{i-1} & \beta_0 y_i u_{i-2} \\ \beta_0 y_i u_{i-1} & \beta_0^2 u_{i-i}^2 & \beta_0^2 u_{i-1}u_{i-2} \\ \beta_0 y_i u_{i-2} & \beta_0^2 u_{i-1}u_{i-2} & \beta_0^2 u_{i-2}^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\approx \lim \frac{1}{N}\sum_1^N \begin{pmatrix} y_{i+2}y_i - \beta_0 u_i y_i \\ y_{i+2}u_{i-1} - \beta_0 u_i u_{i-1} \\ y_{i+2}u_{i-2} - \beta_0 u_i u_{i-2} \end{pmatrix}$$

and $u_N \approx -(\alpha/\beta_0)y_N - \beta_1 u_{N-1} - \beta_2 u_{N-2}$. Substituting in the above gives

$$\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \approx \lim \frac{1}{N}\sum_1^N \begin{pmatrix} y_{i+2}y_i \\ y_{i+2}u_{i-1} \\ y_{i+2}u_{i-2} \end{pmatrix}.$$

From the relationship for u_N, we additionally obtain $\lim (1/N)\sum_1^N y_{i+2}u_i \approx 0$. Note that, so far, we have *not* made use of the fact that the true system is modeled as an ARMAX process.

Proceeding, we see that asymptotically the closed loop system is $Ay_t \approx z^2 B(-\mathscr{A}/\mathscr{B})y_t + Cw_t$ where $A(z) = 1 - a_0 z$, $B(z) = b_0 + b_1 z$, $C(z) = 1 + c_0 z$, $\mathscr{A}(z) = \alpha/\beta_0$, $\mathscr{B}(z) = 1 + \beta_1 z + \beta_2 z^2$. This is equivalent to $(A\mathscr{B} + z^2 B\mathscr{A})y_t \approx C\mathscr{B}w_t$. Define the process $v_t := C/(A\mathscr{B} + z^2 B\mathscr{A})w_t$. Then $y_t \approx \mathscr{B}v_t$ and $u_t \approx -(\mathscr{A}/\mathscr{B})y_t \approx -\mathscr{A}v_t$. Since $\lim (1/N)\sum_1^N u_i y_{i+j} \approx 0$ for $j = 2$, 3 and 4 it follows, through $u_t \approx (-\alpha/\beta_0)v_t$, that $\lim (1/N)\sum_1^N v_i y_{i+j} \approx 0$ for $j = 2$, 3 and 4. Now y_{i+5} is a linear combination of (y_{i+4}, y_{i+3}, y_{i+2}, w_{i+5}, w_{i+4}, w_{i+3}, w_{i+2}). Hence $\lim (1/N)\sum_1^N v_i y_{i+j} \approx 0$ for $j \geqq 5$ also, and so for all $j \geqq 2$. Since $y_i \approx v_i + \beta_1 v_{i-1} + \beta_2 v_{i-2}$, it follows that $\lim (1/N)\sum_1^N y_i y_{i+j} \approx 0$ for all $j \geqq 2$. Hence $\{y_i\}$ is a pure *moving average process* of order 2, i.e., $y_t \approx f_1 w_t + f_2 w_{t-1}$ Hence $C\mathscr{B}/(A\mathscr{B} + z^2 B\mathscr{A}) \approx F$ where $F(z) = f_1 + f_2 z$, i.e., $C\mathscr{B} = FA\mathscr{B} + z^2 BF\mathscr{A}$ or $C = FA + z^2 BF\mathscr{A}/\mathscr{B}$. Since C and FA are polynomials, so is $z^2 BF\mathscr{A}/\mathscr{B} =: \hat{G}$. Now $z^2 BF\mathscr{A} = \hat{G}\mathscr{B}$ implies that $G := (1/z^2)\hat{G}$ is also a polynomial. Hence $C = FA + z^2 G$, but this shows, see § 2.3, that $u_t = -(G/BF)y_t$ is an optimal control law. But $G/BF = \mathscr{A}/\mathscr{B}$ and so the limiting control law $u_t = -(\mathscr{A}/\mathscr{B})y_t$ is also optimal. □

To the author's knowledge, there is as yet no self-contained proof of convergence of the above original self-tuning regulator. Much progress, though, has been made, as we shall see in the sequel.

In § 8.5.3, we give an explanation of the above result for the case $d = 1$.

8.4. The ordinary differential equation method of analysis. Åström and Wittenmark [92] do not answer the question of *when* the parameter estimates will converge, but indicate, by simulations, that in many cases they do.

Ljung [93], [94] addresses this convergence question. In [93], [94] an ordinary differential equation (ODE) is associated with the pair of recursions

$$\hat{\theta}_{N+1} = \hat{\theta}_N + R_N^{-1}\phi_N(y_{N+1} - \hat{\theta}_N^T\phi_N),$$

$$R_{N+1} = R_N + \phi_{N+1}\phi_{N+1}^T.$$

The justification for replacing the above stochastic difference equations by deterministic ODE's is based on the following idea. Since $R_N = \sum_1^N \phi_i\phi_i^T$, we suspect that elements of R_N are $O(N)$. Thus let us consider the pair, $\Delta\hat{\theta}(N) := \hat{\theta}(N+1) - \hat{\theta}(N)$ and $\Delta(R(N)/N) := R(N+1)/(N+1) - R(N)/N$.

Simple calculations show that

$$\Delta\hat{\theta}(N) := \frac{1}{N}\bar{R}(N)^{-1}\phi_N(y_{N+1} - \hat{\theta}_N^T\phi_N),$$

$$\Delta\bar{R}(N) = \frac{1}{N+1}(\phi_{N+1}\phi_{N+1}^T - \bar{R}(N))$$

where $\bar{R}(N) := R(N)/N$ and $\Delta\bar{R}(N) := \bar{R}(N+1) - \bar{R}(N)$. Note that because of the presence of the multiplier $1/N$, $\hat{\theta}(N)$ changes very slowly for large N. Thus changes in the *control law* take place very slowly. Over long periods of time, it may be reasonable to expect that the control law applied is approximately constant. In that case, perhaps $\phi_N y_{N+1}$ and $\phi_{N+1}\phi_{N+1}^T$ can be replaced by their ergodic limits. Thus, the following ODE's appear to be a plausible representation of the asymptotic behaviour of the sample paths of the above stochastic difference equations:

$$\frac{d\theta(\tau)}{d\tau} = \bar{R}^{-1}(\tau)E_{\theta(\tau)}(\phi y), \qquad \frac{d\bar{R}(\tau)}{d\tau} = E_{\theta(\tau)}(\phi\phi^T) - \bar{R}(\tau)$$

where we have also used the time rescaling $\tau = \log t$ to eliminate the factor $1/t$. Here $E_\theta(\phi y)$ is the expected value, assuming stationarity, of $\phi_t y_{t+1}$ and $E_\theta(\phi\phi^T) = E(\phi_t\phi_t^T)$, when the fixed control law corresponding to an estimate θ is used.

The exact technical justification for the above procedure is, at the least, very complex, and in any case, for the particular problem at hand, some boundedness conditions etc., have to be either assumed or proved by other methods.

However, this should not obscure the *great* value of the above ODE's. They were initially obtained to help in the study of a problem of great complexity and have provided the essential breakthroughs which have permitted further development of the field. At the moment these ODE's are an irreplaceable tool for problems in which there is no theory. An analysis of the ODE's suggests what one may expect without the need for extensive simulations. Moreover, because of the time scaling $\tau = \log t$, phenomena which in simulation would be observed for very large values of t would occur in the ODE solutions for modest values of τ. Ljung [93], [94] presents many convincing examples of the usefulness of the ODE's.

Kushner [95] also addresses the problem of obtaining ODE's to model the behaviour of stochastic difference equations. This approach is based on the weak convergence of measures; also see Kushner and Clark [96]. For a recent, very elegant martingale approach, the reader is referred to Metivier and Priouret [139].

A more direct approach, also recent, is due to Kushner and Shwartz [149], [150]. Besides the usefulness of the invariant measure approach, a distinct advantage of [149] is that it addresses the problem of projecting the estimates into D_s, see below.

We now study the behaviour of the above ODE's for the STR.

THEOREM 8.3 (Ljung). *Let*

i) $Ay_t = zBu_t^\theta + Cw_t$ *where* $A(z) := 1 - a_0 z - a_1 z - \cdots - a_n z^{n+1}$, $B(z) := b_0 + b_1 z + \cdots + b_n$ *and* $C(z) := 1 + c_0 z + \cdots + c_n z^{n+1}$;

ii) $\phi_t := (y_t, y_{t-1}, \cdots, y_{t-m}, u_{t-1}, u_{t-2}, \cdots, u_{t-p})^T$;

iii) $u_t^\theta := -\frac{1}{b_0}\theta^T\phi_t$;

iv) $E_\theta(\phi y)$ *is the expectation, assuming a steady state, of* $\phi_{t-1}y_t$. *Thus* θ *is restricted*

to the subset $D_s \subseteq \mathbb{R}^{2n+1}$ *for which the system is strictly stable. Similarly* $E_\theta(\phi\phi^T)$ *is the expectation, in steady state, of* $\phi_t\phi_t^T$.

Suppose that

v) *Real* $(1/C(e^{i\omega})-\frac{1}{2})>0$ *for all* ω *and all roots of* $C(z)$ *are strictly outside the unit circle*;

vi) $B(z)$ *is minimum phase, i.e., all roots of* $B(z)$ *are strictly outside the unit circle*;

vii) $(a_0+c_0)+(a_1+c_1)z+\cdots+(a_n+c_n)z^n$ *and* $B(z)$ *are exactly of degree* n *and contain no common factors.*

Suppose that the ODE*'s*

$$\frac{d\theta(\tau)}{d\tau} = R(\tau)^{-1}E_{\theta(\tau)}(\phi y), \qquad \theta(0)\in D_s,$$

$$\frac{dR(\tau)}{d\tau} = E_{\theta(\tau)}(\phi\phi^T)-R(\tau), \qquad R(0)=I,$$

are such that $\theta(\tau)\in D_s$ *for all* $\tau\geqq 0$. *Define* $V(\theta, R) := (\theta-\theta_{MV})^T R(\theta-\theta_{MV})$ *where* $\theta_{MV} := (a_0+c_0, a_1+c_1, \cdots, a_n+c_n, b_1, b_2, \cdots, b_n)$. *Then*

i) θ_{MV} *is the unique equilibrium point of the first* ODE.

ii) $V(\theta(\tau), R(\tau))>0$ *whenever* $\theta(\tau)\neq\theta_{MV}$.

iii) $(d/d\tau)V(\theta(\tau), R(\tau))<0$ *whenever* $\theta(\tau)\neq\theta_{MV}$.

Proof. First we need to calculate $E_\theta(\phi y)$ and $E_\theta(\phi\phi^T)$, i.e., we need to calculate $E(\phi_{t-1}y_t)$ and $E(\phi_t\phi_t^T)$ assuming (i), (ii) and (iii) are in steady state. Let $\theta^0 := (a_0, a_1, \cdots, a_n, b_1, \cdots, b_p)^T$. Then $y(t)=\phi^T(t-1)\theta^0+b_1u(t-1)+C(z)w(t)$ and substituting from (iii) gives $y(t)=\phi^T(t-1)(\theta^0-\theta)+C(z)w(t)=\phi^T(t-1)(\theta^0-\theta_{MV})+\phi^T(t-1)(\theta_{MV}-\theta)+C(z)w(t)$. But since $\phi^T(t-1)(\theta^0-\theta_{MV})=-\sum_{i=0}^n c_iy(t-i)=(1-C(z))y(t)$, we get $C(z)y(t)=\phi^T(t-1)(\theta_{MV}-\theta)+C(z)w(t)$ or equivalently, $y(t)=\tilde{\phi}^T(t-1)(\theta_{MV}-\theta)+w(t)$ where $\tilde{\phi}(t) := C(z)^{-1}\phi(t)$. Hence $E(\phi_{t-1}y_t)=E(\phi_{t-1}\tilde{\phi}_{t-1}^T)(\theta_{MV}-\theta)$.

Now we want to show that $E(\phi_t\tilde{\phi}_t^T+\tilde{\phi}_t\phi_t^T)$ is positive definite, but this is true since $1/C(z)$ is positive real, and $\phi_t=(1/C(z))\tilde{\phi}(t)$. Similarly $E(\phi_t\tilde{\phi}_t^T+\tilde{\phi}_t\phi_t^T-\phi_t\phi_t^T)$ is positive definite because $1/C(z)-\frac{1}{2}$ is positive real.

Now we see that the ODE's can be written as

$$\frac{d\theta(\tau)}{d\tau} = R(\tau)^{-1}E_{\theta(\tau)}(\phi\tilde{\phi}^T)(\theta_{MV}-\theta(\tau)),$$

$$\frac{dR(\tau)}{d\tau} = E_{\theta(\tau)}(\phi\phi^T)-R(\tau).$$

Clearly $\theta=\theta_{MV}$ is an equilibrium point of the first ODE and now we show that it is unique. Suppose θ^* is any other equilibrium point, then $E_{\theta^*}(\phi\tilde{\phi}^T)(\theta_{MV}-\theta^*)=0$ and $(\theta_{MV}-\theta^*)^TE_{\theta^*}(\phi\tilde{\phi}^T+\phi\tilde{\phi}^T)(\theta_{MV}-\theta^*)=0$. By positive realness, again, it follows that $(\theta_{MV}-\theta^*)^T\phi_{t-1}=0$, but then both θ^* and θ_{MV} are minimum variance control laws. But by (vii), there is a unique such law, and so $\theta^*=\theta_{MV}$, showing (i). (ii) is clearly true, and by simple computation

$$\frac{d}{d\tau}V(\theta(\tau), R(\tau)) = -(\theta(\tau)-\theta_{MV})^T[E_{\theta(\tau)}(\phi\tilde{\phi}^T+\tilde{\phi}\phi^T-\phi\phi^T)+R(\tau)](\theta(\tau)-\theta_{MV})$$
$$<0. \qquad \square$$

(Note that the above theorem is associated with the situation where the parameter b_0

in the model is exactly known, and so only the other parameters are estimated by a least squares procedure.)

This theorem points to the central role played by the positive real condition on the polynomial $C(z)$ in the convergence of the STR.

To show that $\theta(\tau) \to \theta_{MV}$ as $\tau \to \infty$, one could, for example, show that

$$\beta(\|\theta(\tau) - \theta_{MV}\|) \leqq V(\theta(\tau), R(\tau)) \leqq \alpha(\|\theta(\tau) - \theta_{MV}\|) > 0$$

and

$$\frac{dV}{d\tau}(\theta(\tau), R(\tau)) \leqq -\delta(\|\theta(\tau) - \theta_{MV}\|) < 0$$

where α, β and δ are nondecreasing continuous functions such that $\alpha(0) = \beta(0) = \delta(0) = 0$. This would (for example) prove the uniform asymptotic stability of θ_{MV}.

8.5. Martingale methods to exhibit asymptotic cost optimality. We now show asymptotic optimality of the incurred cost in self-tuning schemes by using martingale methods.

8.5.1. To start, we consider a slightly restricted model.

$$y(t+1) = \sum_{i=0}^{n} a_i y(t-i) + \sum_{i=0}^{n} b_i u(t-i) + w(t+1) + \sum_{i=0}^{n} c_i w(t-i)$$

where the restriction lies in the fact that we have taken the delay to be exactly 1, in contrast to the general case of § 8.3.

For this model, we use a slight modification of the STR algorithm of § 8.3. Specifically, the recursions

$$\hat{\theta}(N+1) = \hat{\theta}(N) + \frac{\gamma}{r(N)} \phi(N)[y(N+1) - \hat{\theta}^T(N)\phi(N)], \qquad \hat{\theta}(0), \quad \gamma > 0,$$

$$r(N+1) = 1 + \sum_{i=0}^{N} \phi^T(i)\phi(i)$$

where $\phi(i) := (y(i), y(i-1), \cdots, y(i-n), u(i), u(i-1), \cdots, u(i-n))^T$, are used. The difference with the scheme of § 8.3 lies in the fact that $R(N)$ has been replaced by its *trace* $r(N)$. This recursion is called the *stochastic approximation* (or stochastic gradient) algorithm; also see [71].

The controls are, as before, generated by

$$u(t) := -\frac{1}{\hat{\beta}_0}(\hat{\alpha}_0 y(t) + \cdots + \hat{\alpha}_n y(t-n) + \hat{\beta}_1 u(t-1) + \cdots + \hat{\beta}_n u(t-n))$$

where $\hat{\theta}(t) =: (\hat{\alpha}_0, \hat{\alpha}_1, \cdots, \hat{\alpha}_n, \hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_n)^T$. Or, implicitly, $u(t)$ is defined through the relation $\hat{\theta}^T(t)\phi(t) = 0$. (It will follow, under the assumptions listed below, that $\hat{\beta}_0 = 0$ is a zero probability event, and so the scheme is well defined.)

THEOREM 8.4 (Goodwin, Ramadge, Caines). *Suppose*

i) $\{w(t)\}$ *satisfies* $E(w(t+1)|\mathcal{F}_t) = 0$, $E(w^2(t+1)|\mathcal{F}_t) = \sigma^2$ *and* $E(|w(t+1)|^{2+\delta}|\mathcal{F}_t) < +\infty$ *for some* $\delta > 0$, *for all* t. *Here* $\mathcal{F}_t := \sigma(w(0), \cdots, w(t))$ *is the* σ*-algebra generated by the "past". The probability distribution of* $w(t)$ *is mutually absolutely continuous with respect to Lebesgue measure.*

ii) *The polynomials* $b_0 + b_1 z + \cdots + b_n z^n$ *and* $1 + c_0 z + \cdots + c_n z^{n+1}$ *have all their roots strictly outside the unit circle. Further,* $\operatorname{Re}(1 - (\gamma/2) + c_0 e^{i\omega} + \cdots + c_n e^{ni\omega + i\omega}) > 0$ *for all* ω. ($i = \sqrt{-1}$.)

Then $\lim (1/N) \sum_1^N y^2(t) = \sigma^2$ *and* $\limsup (1/N) \sum_1^N u^2(t) < +\infty$ *a.s.*

Proof. Consider the stochastic Lyapunov function, $V(t) = \|\tilde{\theta}(t)\|^2$ where $\tilde{\theta}(t) := \hat{\theta}(t) - \theta^0$ and $\theta^0 := (a_0 + c_0, \cdots, a_n + c_n, b_0, b_1, \cdots, b_n)^T$. A simple calculation shows that

$$\begin{aligned}
E(V(t+1)|\mathcal{F}_t) &= V(t) + \frac{2\gamma}{r(t)}\phi^T(t)\tilde{\theta}(t)E(y(t+1)|\mathcal{F}_t) \\
&\quad + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)(E(y(t+1)|\mathcal{F}_t))^2 + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 \\
&\leqq V(t) + \frac{2\gamma}{r(t)}\phi^T(t)\tilde{\theta}(t)E(y(t+1)|\mathcal{F}_t) \\
&\quad + \frac{\gamma^2}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2 + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 \\
&= V(t) - \frac{2\gamma}{r(t)}\left\{\phi^T(t)\tilde{\theta}(t) - \frac{(\gamma+\varepsilon)}{2}E(y(t+1)|\mathcal{F}_t)\right\}E(y(t+1)|\mathcal{F}_t) \\
&\quad - \frac{\varepsilon\gamma}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2 + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2,
\end{aligned}$$

where $\varepsilon > 0$ is chosen so small that the inequality in (ii) is still true when γ is replaced by $(\gamma + \varepsilon)$. Then

$$\begin{aligned}
\left[C(z) - \frac{\gamma+\varepsilon}{2}\right]E(y(t+1)|\mathcal{F}_t) &= C(z)(y(t+1) - w(t+1)) - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(t+1)|\mathcal{F}_t) \\
&= y(t+1) - w(t+1) + [C(z) - 1](y(t+1) - w(t+1)) - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(t+1)|\mathcal{F}_t) \\
&= \phi^T(t)\theta^0 - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(t+1)|\mathcal{F}_t) \\
&= -\phi^T(t)\tilde{\theta}(t) - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(t+1)|\mathcal{F}_t).
\end{aligned}$$

The right-hand side above can thus be viewed as $E(y(t+1)|\mathcal{F}_t)$ "filtered" through the system with transfer function $[C(z) - (\gamma+\varepsilon)/2]$. Because of the property of positive real transfer functions, it follows that

$$S(t) := 2\gamma \sum_{n=1}^{t}\left\{-\phi^T(n)\tilde{\theta}(n) - \left(\frac{\gamma+\varepsilon}{2}\right)E(y(n+1)|\mathcal{F}_n)\right\}E(y(n+1)|\mathcal{F}_n) + K \geqq 0$$

for all t, for some K, a.s. Now it follows that if $M(t) := V(t) + S(t-1)r^{-1}(t-1)$, then

$$\begin{aligned}
E(M(t+1)|\mathcal{F}_t) &\leqq V(t) - \frac{1}{r(t)}(S(t) - S(t-1)) + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 \\
&\quad - \frac{\varepsilon\gamma}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2 + \frac{S(t)}{r(t)} \\
&= V(t) + \frac{S(t-1)}{r(t)} + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 - \frac{\varepsilon\gamma}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2 \\
&\leqq M(t) + \frac{\gamma^2}{r^2(t)}\phi^T(t)\phi(t)\sigma^2 - \frac{\varepsilon\gamma}{r(t)}(E(y(t+1)|\mathcal{F}_t))^2.
\end{aligned}$$

Thus, $\{M_t, \mathcal{F}_t\}$ is "nearly a positive supermartingale" if $\sum (\phi^T(t)\phi(t)/r^2(t)) < \infty$, see [72]. But this is true because

$$\sum_1^N \frac{\phi^T(t)\phi(t)}{r^2(t)} = \sum_1^N \frac{r(t)-r(t-1)}{r^2(t)} \leqq \sum_1^N \frac{1}{r(t-1)} - \frac{1}{r(t)} \leqq 1.$$

Hence, by [73], $\{M_t\}$ converges a.s., and furthermore,

$$\sum \frac{(E(y(t+1)|\mathcal{F}_t))^2}{r(t)} < +\infty \quad \text{a.s.}$$

If $r(t) \to +\infty$, then $(1/r(N)) \sum_1^N E(y(t+1)|\mathcal{F}_t)^2 \to 0$ by Kronecker's lemma. If not, then $y^2(t) \to 0$ and $u^2(t) \to 0$ and so $C(z)w(t) \to 0$, which happens only on a null set. Due to the minimum phase condition (that all roots of $b_0 + b_1 z + \cdots + b_n z^n$ are strictly outside the unit circle) it follows that for some k_1, k_2

$$\frac{1}{N}\sum_1^N u^2(t) \leqq \frac{k_1}{N}\sum_1^N y^2(t+1) + \frac{k_2}{N}.$$

Hence

$$\frac{r(N)}{N} \leqq \frac{k_3}{N}\sum_1^N y^2(t+1) + \frac{k_4}{N}.$$

Since $E(y(t+1)|\mathcal{F}_t) = y(t+1) - w(t+1)$, and since $(1/N)\sum_1^N w^2(t+1) \to \sigma^2$, it follows that

$$\frac{r(N)}{N} \leqq \frac{k_5}{N}\sum_1^N (E(y(t+1)|\mathcal{F}_t))^2 + k_6,$$

$$\frac{1}{r(N)}\sum_1^N (E(y(t+1)|\mathcal{F}_t))^2 \geqq \frac{r(N) - k_6 N}{k_5 r(N)}.$$

Suppose $\{r(N)/N\}$ is unbounded, then along some subsequence, $(1/r(N_k))\sum_1^{N_k} (E(y(t+1)|\mathcal{F}_t))^2 \geqq 1/2k_5 > 0$ which is a contradiction, and so $\{r(N)/N\}$ is bounded, which in turn shows that $(1/N)\sum_1^N (E(y(t+1)|\mathcal{F}_t))^2 \to 0$ a.s. Now

$$\begin{aligned}\frac{1}{N}\sum_1^N y^2(t+1) &= \frac{1}{N}\sum_1^N [E(y(t+1)|\mathcal{F}_t)^2 + w^2(t+1) + 2w(t+1)E(y(t+1)|\mathcal{F}_t)] \\ &= \frac{1}{N}\sum_1^N w^2(t+1) + \frac{1}{N}\sum_1^N E(y(t+1)|\mathcal{F}_t)^2 \\ &\quad + 2\alpha(N)\left(\frac{1}{N}\sum_1^N E(y(t+1)|\mathcal{F}_t)^2\right)^{1/2}\left(\frac{1}{N}\sum_1^N w^2(t+1)\right)^{1/2}\end{aligned}$$

for some $|\alpha(N)| \leqq 1$ by the Cauchy Schwarz inequality. Taking limits gives $(1/N)\sum_1^N y^2(t) \to \sigma^2$ a.s. □

The above result shows that optimality is achieved for this NACP in the sense of § 6.2. (Actually we have used the slightly stronger condition $E|w_t|^{2+\delta} < +\infty$ to show a result slightly stronger and slightly more relevant than the conclusion $(1/N)\sum_1^N E(y^2(t+1)|\mathcal{F}_t) \to \sigma^2$ a.s. of [75].)

8.5.2. Let us call the problem of minimizing $\lim (1/N)\sum_1^N y^2(t)$, the *regulation* problem. Consider now the slightly different problem of minimizing $\lim (1/N)\sum_1^N (y(t) - y^*(t))^2$ where $\{y^*(t)\}$ is some prespecified reference trajectory

which we want the output of the system to follow. We shall call this the *tracking* problem. Note that the regulation problem is a special case of the tracking problem where one wishes to track the identically zero trajectory.

For the general tracking problem, a new twist arises. Algorithms for which proofs of asymptotic optimality exist do require estimation, in one form or another, of the coefficients $\{c_0, \cdots, c_n\}$ describing the spectrum of the noise. Hence, while in § 8.5.1, the vector $\hat{\theta}(t)$ was of dimension $(2n+2)$, now it is of dimension $(3n+3)$.

THEOREM 8.5 (Goodwin, Ramadge and Caines). *Let the algorithm be as in Theorem* 8.4, *subject only to the following changes*:

i) $\phi(t) := (y(t), y(t-1), \cdots, y(t-n), u(t), u(t-1), \cdots, u(t-n), -y^*(t), \cdots, -y^*(t-n))^T$;

ii) $\hat{\theta}(t) \in \mathbb{R}^{3n+3}$;

iii) $\{y^*(t)\}$ *is bounded*;

iv) $$u(t) = -\frac{1}{\hat{\beta}_0}(\hat{\alpha}_0 y(t) + \cdots + \hat{\alpha}_n y(t-n) + \hat{\beta}_1 u(t-1) + \cdots + \hat{\beta}_n u(t-n) - y^*(t+1) - \hat{\gamma}_0 y^*(t) - \cdots - \hat{\gamma}_n y^*(t-n))$$

or when $\hat{\theta}(t) =: (\hat{\alpha}_0, \cdots, \hat{\alpha}_n, \hat{\beta}_0, \cdots, \hat{\beta}_n, \hat{\gamma}_0, \cdots, \hat{\gamma}_n)$, $u(t)$ *is implicitly specified by* $\phi^T(t)\hat{\theta}(t) = y^*(t+1)$.

Then

$$\lim \frac{1}{N} \sum_1^N (y(t) - y^*(t))^2 = \sigma^2 \quad a.s.$$

and

$$\limsup \frac{1}{N} \sum_1^N u^2(t) < +\infty \quad a.s.$$

Proof. Similar to Theorem 8.4, or the reader may refer to [75] for explicit details. □

It should be noted that if $y^*(t) = 0$ for all t, then the last $(n+1)$ components of $\phi(\cdot)$ and $\theta(\cdot)$ make no contribution at all to the algorithm, and the vectors can therefore be collapsed, giving the same algorithm as was used in § 8.5.1 for the regulation problem.

8.5.3. How is one to understand *why* the algorithm of this section or that of the previous § 8.5.1 leads to asymptotic optimality? We proceed as follows.

Define $\hat{y}(t+d) = E(y(t+d)|\mathcal{F}_t)$ the d-step ahead prediction of $y(t+d)$. We first obtain a recursive formula for $\hat{y}(t)$ by quick formal manipulation of polynomials; these can be done rigorously by the techniques of Åström [89]. Consider the system of § 8.2.3 with $p = n$. Then

$$y(t+1) = z^d \frac{B}{A} u(t+1) + \frac{C}{A} w(t+1)$$

$$= z^d \frac{B}{A} u(t+1) + Fw(t+1) + z^d \frac{G}{A} w(t+1) \quad \text{(remembering } C = AF + z^d G\text{)}.$$

Thus the d-step ahead predictor is $\hat{y}(t+1) = y(t+1) - Fw(t+1)$. Hence

$$\hat{y}(t+1) = z^d \frac{B}{A} u(t+1) + z^d \frac{G}{A} w(t+1).$$

Since the error of prediction is $y(t+1) - \hat{y}(t+1) = Fw(t+1)$, we can substitute for

$w(t+1)$ to obtain $y(t+1) = z^d(B/A)u(t+1) + z^d(G/AF)(y(t+1) - \hat{y}(t+1))$. Multiplying through by AF, and using $C = AF + z^dG$ gives

$$C\hat{y}(t+1) = z^dBFu(t+1) + z^dGy(t+1).$$

Now consider the tracking problem if the system is known. Clearly one will choose $u(s)$ at each instant of time so that $\hat{y}(s+d) = y^*(s+d)$. Hence we will want the relation $BFu(t) + Gy(t) = Cy^*(t+d)$, i.e., one should choose $u(t)$ so that

$$\begin{aligned} u(t) = -\frac{1}{(bf)_0}[&g_0y(t) + \cdots + g_ny(t-n) + (bf)_1u(t-1) + \cdots \\ &+ (bf)_{n+d-1}u(t-n-d+1) - c_0y^*(t+d-1) - \cdots \\ &- c_ny^*(t+d-n-1) - y^*(t+d)] \end{aligned}$$

where $B(z)F(z) =: (bf)_0 + \cdots + (bf)_{n+d-1}z^n$. Moreover since $y(t+1) - y^*(t+1) = Fw(t+1)$, we can write

$$\begin{aligned} y(t+1) = &g_0y(t-d+1) + \cdots + g_ny(t-d-n+1) + (bf)_0u(t-d+1) + \cdots \\ &+ (bf)_{n+d-1}u(t-d-n+1) - c_0y^*(t) - \cdots - c_ny^*(t-n) + Fw(t+1). \end{aligned}$$

Now consider three cases.

Case 1. *Regulation, unit delay, coloured noise.* Here $y^*(t) = 0$ for all t. So under *optimal* control, the true system looks like

$$y(t+1) = g_0(t)y(t) + \cdots + g_ny(t-n) + (bf)_0u(t) + \cdots + (bf)_nu(t-n) + w(t+1),$$

$$u(t) = -\frac{1}{(bf)_0}[g_0y(t) + \cdots + g_ny(t-n) + (bf)_1u(t-1) + \cdots + (bf)_nu(t-n)].$$

Now it is transparent that in Theorem 8.4 we are actually identifying the coefficients of the first equation, and then adopting these estimated coefficients to calculate the control law as in the second equation. Since the noise in the system is now *white*, one can expect that this scheme does work (see §§ 8.1.3 and 8.1.4). (Actually since $d = 1$, we obtain $F = 1$, $zG = C - A$ and $BF = B$.)

Schemes of this type are called *implicit* or direct because they identify some *closed loop* parameters (here $g_0, \cdots, g_n, (bf)_0, \cdots, (bf)_n$) and *not* the open loop parameters ($a_0, \cdots, a_n, b_0, \cdots, b_n, c_0, \cdots, c_n$). For a discussion of this point, see [75] and Fuchs [97].

Case 2. *Tracking, unit delay, coloured noise.* Here $y^*(t)$ is *not* identically zero. Hence one *cannot* neglect $y^*(\cdot)$ as we did in Case 1. So under *optimal* control, the true system can be modeled as:

$$\begin{aligned} y(t+1) = &g_0y(t) + \cdots + g_ny(t-n) + (bf)_0u(t) + \cdots \\ &+ (bf)_nu(t-n) - c_0y^*(t) - \cdots - c_ny^*(t-n) + w(t+1), \end{aligned}$$

$$\begin{aligned} u(t) = -\frac{1}{(bf)_0}[&g_0y(t) + \cdots + g_ny(t-n) + (bf)_1u(t-1) + \cdots \\ &+ (bf)_nu(t-n) - c_0y^*(t) - \cdots - c_ny^*(t-n) - y^*(t+1)]. \end{aligned}$$

Again, the adaptive scheme of Theorem 8.5 *assumes* the system is in this form, and estimates the coefficients of this assumed model, and chooses a control input which is optimal for the estimated parameters. Again the noise in the system is *white* and one can expect convergence at least under some conditions, which is the content of Theorem 8.5. Again, this is an implicit or direct scheme.

Case 3. *Interlacing.* Consider the system only at d time unit intervals. Then

$$y(t+1)=\sum_{i=0}^{n} g_i y(t-d-i+1)+\sum_{i=0}^{n+d-1} (bf)_i u(t-d-i+1)$$
$$+\sum_{i=0}^{n}(-c_i)y^*(t-i)+v_{t+1},$$
$$y(t+d+1)=\sum_{i=0}^{n} g_i y(t-i+1)+\sum_{i=0}^{n+d-1} (bf)_i u(t-i+1)$$
$$+\sum_{i=0}^{n}(-c_i)y^*(t+d-i)+v_{t+d+1},$$
$$y(t+2d+1)=\sum_{i=0}^{n} g_i y(t-i+d+1)+\sum_{i=0}^{n+d-1} (bf)_i u(t+d-i+1)$$
$$+\sum_{i=0}^{n}(-c_i)y^*(t+2d-i)+v_{t+2d+1}$$
$$\vdots$$

where

$$v_t := f_0 w_t + f_1 w_{t-1} + \cdots + f_{d-1} w(t-d+1).$$

Now we see that $\{v_{t+1}, v_{t+d+1}, v_{t+2d+1}, \cdots\}$ are *independent* of each other. Hence one can estimate the parameters and even hope for asymptotic consistency. Thus one updates $\hat{\theta}(t+1)$ to $\hat{\theta}(t+d+1)$ to $\hat{\theta}(t+2d+1)\cdots$ etc. This procedure is called *interlacing* because one needs to store $(\hat{\theta}(t+1), \hat{\theta}(t+2), \cdots, \hat{\theta}(t+d))$ and then update it to $(\hat{\theta}(t+d+1), \hat{\theta}(t+d+2), \cdots, \hat{\theta}(t+2d))$ etc. Goodwin, Ramadge and Caines [75] present a scheme based on this for the case where $C=1$, $d \geqq 1$ while Goodwin, Sin and Saluja [98] present one for the general case. Both use a stochastic approximation scheme to update the parameters.

8.5.4. Fuchs [97], [99], [100] has considered *indirect* or *explicit* schemes where one first estimates the coefficients of the (open loop) model and *not* the coefficients in the *prediction* form, as is done in [98]. In [99], [100] the case when the delay is $d \geqq 1$ is considered and a scheme is used which does *not* involve the cumbersome interlacing procedure.

THEOREM 8.6 (Fuchs). *Consider the system with the assumptions of Theorem* 8.5, *the only change being that* $d \geqq 1$. *Let* $\{y^*(t)\}$ *be a desired* bounded *reference trajectory. Consider the algorithm*:

$$\hat{\theta}(t+1)=\hat{\theta}(t)+\frac{\gamma}{r(t)}\phi(t)e(t+1), \qquad \gamma>0,$$
$$e(t+1) := y(t+1)-\hat{\theta}(t)^T\phi(t),$$
$$\phi^T(t) := (y(t),\cdots,y(t-n),u(t-d+1),\cdots,u(t-n-d+1),e(t),\cdots,e(t-n))$$

and

$$r(t+1)=r(t)+\phi^T(t+1)\phi(t+1), \qquad r(0)=1.$$

Let $\hat{y}(t+d|\hat{\theta}(t))=E(y(t+d)|\mathcal{F}_t, \hat{\theta}(t))$ *be the prediction of* $y(t+d)$ *if* $\hat{\theta}(t)$ *is the true parameter, as in* § 8.5.3. *Choose* $u(t)$ *so that* $\hat{y}(t+d|\hat{\theta}(t))=y^*(t+d)$. *Then* $\lim (1/N)\sum_1^N (y(t+d)-y^*(t+d))^2=0$ *and* $\limsup (1/N)\sum_1^N u^2(t)<+\infty$ *a.s.*

Proof. See [100]. □

The main difference, as explained above, is that one attempts to estimate *all* the coefficients of the true system. For this purpose an identification scheme (called RML_1) has been used (or rather, a stochastic approximation variant of it). The identification method estimates $w(t)$ by $e(t)$, the residuals $y(t)-\phi^T(t-1)\hat{\theta}(t-1)$, see § 8.1.5.

Another treatment of a stochastic approximation based adaptive control scheme is given by Kushner and Kumar [101]. This is based on the use of (truncated) bounded estimates and control inputs. Because of the latter, it is assumed that the system is (open loop) *stable*, i.e., all roots of A are outside the unit circle.

8.5.5. The most serious problem with the stochastic approximation based schemes of §§ 8.5.2 and 8.5.4 is that their *rate* of convergence has been observed in practice to be *very* slow compared to least squares based algorithms. (The rates of convergence of these algorithms are, to the author's knowledge, yet to be established rigorously.)

8.5.6. The asymptotic optimality of a strict least squares type scheme such as, say, in § 8.3 (the original self-tuning regulator) has yet to be established. However, algorithms *closely* related to a least squares algorithm have been analyzed by Kumar and Moore [102], Kumar [146] and Sin and Goodwin [103]. [102] considers an algorithm which gives more weight to past measurements than recent measurements. We now briefly discuss the scheme of [103]. They consider an estimation algorithm of the form

$$\hat{\theta}(t+1)=\hat{\theta}(t)+\gamma^{-1}(t)R^{-1}(t)\phi(t)[y(t+1)-\phi^T(t)\hat{\theta}(t)],$$
$$R(t+1)=\gamma(t+1)^{-1}[R(t)+\phi(t+1)\phi(t+1)^T]$$

where

$$\phi(t):=(y(t),\cdots,y(t-n),u(t-1),\cdots,u(t-n),-\bar{y}(t),\cdots,-\bar{y}(t-n))^T,$$
$$\bar{y}(t):=\phi^T(t-1)\hat{\theta}(t)$$

and a rule for calculating $\gamma(t)$, $0<\gamma(t)\leqq 1$, is specified.

Comparing this with the least squares recursion of §§ 8.1.2 and 8.1.3 we see the following salient differences. First $\gamma(t+1)$ is *not* always 1. Second, in the vector $\phi(t)$ occur the predictions $\bar{y}(t)=\phi^T(t-1)\hat{\theta}(t)$. The distinctive feature is that in calculating this prediction we use $\hat{\theta}(t)$ and *not* $\hat{\theta}(t-1)$. Thus this prediction is made on the basis of a *more recent* estimate, *and* is called an a posteriori prediction (contrast this with the recursions in §§ 8.5.1–8.5.3). Proofs and even convergence of identification schemes can rest on such fine differences, see Solo [88].

The use of a posteriori predictions (as opposed to a priori predictions) does indeed add one new wrinkle with respect to the *regulation* problem. Note that we choose $u(t)$ so that $\phi^T(t)\hat{\theta}(t)=0$. However $\bar{y}(t)\neq 0$ and so $(\bar{y}(t),\cdots,\bar{y}(t-n))$ *cannot* be dropped from the vector $\phi(t)$—as they could before, see Case 1 of § 8.5.3.

The computation of $\gamma(t)$ involves a cumbersome process and it would be of interest to obtain schemes which do not involve such computations.

You-Hong [104] extends the results of [103] to the general delay case. Gawthrop [105] contains an analysis of least squares based schemes. Kumar [106] exhibits the consequences of assuming that a certain "regularity condition" is satisfied by the closed loop system.

8.6. Convergence of parameter estimates and control laws. So far, we have not justified the use of the *phrase* (adjective) "self-tuning" in the names of these schemes. Specifically, we have not proved that the control law converges to the optimal control

law. We have only shown that the cost incurred is optimal (i.e., $\lim (1/N)\sum_1^N y^2(t) = \sigma^2$ a.s.).

The question has recently been addressed for the scheme of § 8.5.1 in [107].

THEOREM 8.7 (Becker, Kumar and Wei). *Consider the assumptions of Theorem* 8.4 *along with the additional assumption*:

$$(a_0+c_0)+(a_1+c_1)z+\cdots+(a_n+c_n)z^n \quad \textit{and} \quad b_0+b_1z+\cdots+b_nz^n$$

have no common factors and $|a_n+c_n|+|b_n|>0$. *Then* $\lim \hat{\theta}(t) = k(a_0+c_0,\ a_1+c_1, \cdots, a_n+c_n, b_0, b_1, \cdots, b_n)$ *where k is a random scalar.*

Proof. Since $u(t)$ is chosen to render $\phi^T(t)\hat{\theta}(t)=0$, it follows that $\phi(t)$ and $\hat{\theta}(t)$ are orthogonal. However the recursion $\hat{\theta}(t+1)=\hat{\theta}(t)+(\text{scalar})\ \phi(t)$ shows that $\hat{\theta}(t+1)-\hat{\theta}(t)$ is parallel to $\phi(t)$. These two facts together show that the $\hat{\theta}(t+1)-\hat{\theta}(t)$ is orthogonal to $\hat{\theta}(t)$. Thus the jumps $((\hat{\theta}(t+1)-\hat{\theta}(t)))$ in the parameter estimates are always orthogonal to their value $(\hat{\theta}(t))$ before the jump. Pythagoras' theorem now shows that $\|\hat{\theta}(t)\|$ is an increasing quantity.

By using the Schwarz inequality, the results of [75] can be refined to show that the stochastic Lyapunov function $\|\hat{\theta}(t)-\theta^0\|^2$ converges a.s. This shows, first, that $\{\hat{\theta}(t)\}$ converges to a sphere of random radius with center at θ^0. Second, it also shows that $\{\|\hat{\theta}(t)\|\}$ is bounded, and therefore also converges, since its geometric properties of the preceding paragraph have already shown that it is increasing. Thus $\{\hat{\theta}(t)\}$ also converges to a random sphere centered at the origin.

Since two spheres in Euclidean space (here $\mathbb{R}^{2n+1}$) can intersect either at a point (when they are tangential to each other) or in a hypersphere of dimension $2n$, we wish to show that the latter cannot happen. This is done by showing that there is a subsequence $\{\hat{\theta}(t_k)\}$ which converges to the line connecting the origin and θ^0, i.e., by showing that $\theta_i^0\hat{\theta}_p(t_k)-\theta_p^0\hat{\theta}_i(t_k)\to 0$ for every $i=1,2,\cdots,2n$.

This latter is done by showing that $\lim (1/N)\sum_{t=1}^N (\theta_i^0\hat{\theta}_p(t)-\theta_p^0\hat{\theta}_i(t))^2=0$ a.s. for $i=1,2,\cdots,2n$.

This last part of the proof is illustrated by using an example. Suppose $y(t+1)=a_0y(t)+a_1y(t-1)+b_0u(t)+b_1u(t-1)+w(t+1)+c_0w(t)$. Since the control law is

$$u(t) = -\frac{1}{\hat{b}_0(t)}(\hat{a}_0(t)y(t)+\hat{a}_1(t)y(t-1)+\hat{b}_1(t)u(t-1))$$

(where $\hat{\theta}(t)=:(\hat{a}_0(t), \hat{a}_1(t), \hat{b}_0(t), \hat{b}_1(t))$), the closed loop system is

$$y(t+1) = \left(a_0-b_0\frac{\hat{a}_0(t)}{\hat{b}_0(t)}\right)y(t)+\left(a_1-b_0\frac{\hat{a}_1(t)}{\hat{b}_0(t)}\right)y(t-1)+\left(b_1-b_0\frac{\hat{b}_1(t)}{\hat{b}_0(t)}\right)u(t-1)$$
$$+w(t+1)+c_0w(t).$$

Since

$$\lim\frac{1}{N}\sum_1^N y^2(t+1) = \lim\frac{1}{N}\sum_1^N w^2(t+1)=\sigma^2 \quad \text{a.s.},$$

it can be shown that

$$\lim\frac{1}{N}\sum_1^N\left[\left(a_0-b_0\frac{\hat{a}_0(t)}{\hat{b}_0(t)}\right)y(t)+\left(a_1-b_0\frac{\hat{a}_1(t)}{\hat{b}_0(t)}\right)y(t-1)\right.$$
$$\left.+\left(b_1-b_0\frac{\hat{b}_1(t)}{\hat{b}_0(t)}\right)u(t-1)\right]^2=0 \quad \text{a.s.}$$

Since $\{\hat{b}_0(t)\}$ is bounded, it follows that

$$\lim \frac{1}{N}\sum_1^N [(a_0\hat{b}_0(t)-b_0\hat{a}_0(t))y(t)+(a_1\hat{b}_0(t)-b_0\hat{a}_1(t))y(t-1) + (b_1\hat{b}_0(t)-b_0\hat{b}_1(t))u(t-1)]^2=0.$$

Since $\|\hat{\theta}(t)\|^2=\sum_{p=1}^t \|\hat{\theta}(p)-\hat{\theta}(p-1)\|^2$ (by Pythagoras' theorem), we can replace $\hat{a}_0(t)$, $\hat{a}_1(t)$, $\hat{b}_0(t)$, $\hat{b}_1(t)$ by $\hat{a}_0(t-4)$, $\hat{a}_1(t-4)$, etc., to give

$$\lim \frac{1}{N}\sum_1^N [(a_0\hat{b}_0(t-4)-b_0\hat{a}_0(t-4))y(t)+(a_1\hat{b}_0(t-4)-b_0\hat{a}_1(t-4))y(t-1) + (b_1\hat{b}_0(t-4)-b_0\hat{b}_1(t-4))u(t-1)]^2=0.$$

Denote the square root of the expression in the above summand by $x(t)$. Now denote by $z(t)$ the expression in $x(t)$ where we change (only) each of $y(t)$, $y(t-1)$, $u(t-1)$ to $y(t-1)$, $y(t-2)$, $u(t-2)$ respectively. It can be shown that $(1/N)\sum_1^N x^2(t)\to 0$ implies $(1/N)\sum_1^N z^2(t)\to 0$.

By using a local convergence theorem for martingales, it can be shown that if $b_0x(t)+b_1z(t)=:\alpha(t)y(t)+\beta(t)y(t-1)+\gamma(t)y(t-2)+\eta(t)u(t-1)+\mu(t)u(t-2)$, then $\lim (1/N)\sum_1^N(\alpha^2(t)+\beta^2(t)+\gamma^2(t)+\eta^2(t)+\mu^2(t))=0$.

By using some algebra and condition (i) we get what we want. □

Some points need to be made. It is *not* true that $\lim \hat{\theta}(t)=(a_0+c_0,\cdots,a_n+c_n, b_0,\cdots,b_n)$. In fact, in [107] it is shown that such a limiting value can have *zero* probability. Thus the parameter estimates *do converge*, but *not* to their true values (with regard to the model of Case 1, § 8.5.3).

However the control law is (with $\hat{\theta}(t)=:(\hat{\alpha}_0(t),\cdots,\hat{\alpha}_n(t),\ \hat{b}_0(t),\cdots,\hat{b}_n(t))^T$), $u(t)=-(1/\hat{b}_0(t))[\hat{\alpha}_0(t)y(t)+\cdots+\hat{\alpha}_n(t)y(t-n)+\hat{b}_1(t)u(t-1)+\cdots+\hat{b}_n(t)u(t-n)]$ and it *is* true that

$$\left(\frac{\hat{\alpha}_0(t)}{\hat{b}_0(t)},\cdots,\frac{\hat{\alpha}_n(t)}{\hat{b}_0(t)},\frac{\hat{b}_1(t)}{\hat{b}_0(t)},\cdots,\frac{\hat{b}_n(t)}{\hat{b}_0(t)}\right)$$

converges to

$$\left(\frac{a_0+c_0}{b_0},\cdots,\frac{a_n+c_n}{b_0},\frac{b_1}{b_0},\cdots,\frac{b_n}{b_0}\right).$$

Thus the parameters of the control law do converge to the *optimal* values and self-tuning *does* occur. So here we finally have a demonstration of the result of Example 8.2 of § 8.3.1 and a justification of the description of the adaptive regulator as *self-tuning.*

An important point to note is that this proof does not rely on a "persistency of excitation" assumption. In fact, as we now demonstrate, a persistency of excitation condition does *not* hold. Let $\{N_k\}$ be a subsequence along which $\lim (1/N_k)\sum_1^{N_k}\phi(t)\phi(t)^T$ has a limit. We show that this limit is not positive definite. To see this, let $\hat{\theta}(\infty)$ be the limit of the parameter estimates. Then

$$\hat{\theta}(\infty)^T\left[\lim\frac{1}{N}\sum_1^{N_k}\phi(t)\phi(t)^T\right]\hat{\theta}(\infty) = \lim\frac{1}{N_k}\sum_1^{N_k}\hat{\theta}(\infty)^T\phi(t)\phi(t)^T\hat{\theta}(\infty)$$

$$= \lim \left[\frac{1}{N_k} \sum_1^{N_k} \hat{\theta}(t)^T \phi(t) \phi(t)^T \hat{\theta}(t) \right.$$

$$+ \frac{1}{N_k} \sum_1^{N_k} (\hat{\theta}(\infty) - \hat{\theta}(t))^T \phi(t) \phi^T(t) (\hat{\theta}(\infty) - \hat{\theta}(t))$$

$$\left. + \frac{2}{N_k} \sum_1^{N} \hat{\theta}(t)^T \phi(t) \phi^T(t) (\hat{\theta}(\infty) - \hat{\theta}(t) \right].$$

Since $\hat{\theta}(t)^T \phi(t) = 0$ by the specification of the control law and since $\hat{\theta}(t) \to \hat{\theta}(\infty)$, we see that

$$\hat{\theta}(\infty)^T \left[\lim \frac{1}{N_k} \sum_1^{N_k} \phi(t) \phi(t)^T \right] \hat{\theta}(\infty) = 0$$

showing the *singularity* of $\lim (1/N_k) \sum_1^{N_k} \phi(t) \phi(t)^T$. So we see that one should not *assume* persistency of excitation conditions to hold in *adaptive control.*

Convergence results such as the above need to be obtained in more general situations where least squares estimates are used, etc.

Caines and Lafortune [108] consider a stochastic approximation based scheme where a randomized control law is used, i.e., noise is injected into the system. By verifying that a "persistency of excitation" type condition holds, an auxiliary identification algorithm (running in parallel with the adaptive control algorithm) is shown to provide estimates which converge to the true values (an indirect or explicit scheme is used). Chen [109] also proves the strong consistency of a randomized control scheme with a modified least squares type parameter estimator as in [103]. Chen and Caines [110] reconsider the problem of [108] and show that one does not really need an additional parameter estimator in parallel.

8.7. Other proposed schemes. We now examine briefly various extensions of the basic self-tuning regulator (STR) which have been proposed to cope with various practical problems.

8.7.1. If the system is nonminimum phase, i.e., $b_0 + b_1 z + \cdots + b_n z^n$ has roots on or inside the unit circle, then we have seen in § 8.2.4 that the standard minimum variance regulator can have severe practical problems.

Åström and Wittenmark [111] consider a self-tuning scheme based on the *constrained* minimum variance regulator of Peterka [91], see Theorem 8.1 of § 8.2.5.

Clarke and Gawthrop [112] have proposed a generalization of the basic STR which incorporates models with nonzero steady state offset value (i.e., steady state output nonzero when input zero), tracking and an ability to cope with some nonminimum phase systems. The heart of the approach is the choice of $u(t)$ to minimize the d-step finite horizon cost criterion:

$$E\left[\left[\sum_{i=0}^{m} p_i y(t+d-i) - r_i y^*(t+d-i) \right]^2 + \sum_{i=0}^{m} q_i u(t-i)^2 \middle| \mathcal{F}_t \right].$$

Here $\{p_i, r_i, q_i\}$ are weighting coefficients and $y^*(t)$ can be a desired reference trajectory. By varying the coefficients one can obtain control laws which are stable for some nonminimum phase systems. It is also shown that the desired control law may be an equilibrium point (rest point) of the scheme. Gawthrop [113] and Clarke and Gawthrop [114] deal with generalizations of this.

Koivo [115] considers the multivariable version of this algorithm which in turn generalizes the multivariate version of the self-tuning regulator of Borisson [116].

Gawthrop [105] and Chen [117] analyze these schemes, the latter making some assumptions on the behaviour of the closed loop system. Goodwin, Ramadge and Caines [75] also allow for multivariable systems.

Wellstead, Edmunds, Prager and Zanker [118] propose a scheme based on the assignment of poles/zeros rather than on a cost criterion. This scheme is also based on practical considerations, the desired goal being the ability to deal with nonminimum phase systems, unknown delay d etc. It is shown that these schemes also have the property that the desired control law is an equilibrium point. Gawthrop [119] brings attention to some similarities between one of the schemes of [118] and the self-tuning controller [112]. Wellstead and Sanoff [120] extends [118], while Wellstead and Zanker [121] treats the tracking problem. Allidina and Hughes [122] presents a cost criterion approach. Åström and Wittenmark [123] contains an extensive treatment of the pole-zero assignment problem and the tracking problem.

Another approach is to use a cost criterion of the form $\lim (1/N)\sum_1^N y^2(t)+\rho u^2(t)$ where $\rho>0$ weights the control used. One can attempt to solve on-line, for every current set of estimated parameters, either a Riccati equation or perform a spectral factorization to obtain a control input which is optimal for the estimated parameters, see Åström, Borisson, Ljung and Wittenmark [124]. Mandl [125], [126] examines this scheme in a state space format and proves asymptotic optimality. The chief restriction is that the state is assumed to be completely observed and *only* the control gain matrix is unknown. [78] removes the latter restriction, but only allows the unknown parameter value to lie in a *finite* parameter set. The approach is based on § 7.3.

Grimble [127] proposes a different scheme, which is easier to implement, and which in contrast to some other schemes possesses the property that for the limiting values of some of the adjustable parameters (control weighting term etc.) the control law described in § 7.2.5 is obtained. See also Grimble [128].

Yet another approach is given by Kumar and Moore [129], [130], see also Clarke and Gawthrop [147].

8.7.2. In practical applications, the self-tuning regulator is implemented *not* on *constant* and unknown systems, but rather on *time varying* and unknown systems. It is hoped, in such cases, that the rate of convergence of the parameter estimates will be rapid in comparison with the rate of change of the system. To "keep up" with the changing system, one can make various modifications to the recursive least squares parameter estimation scheme. One can use a moving "window" of time, see Goodwin and Payne [87]. Alternatively, one can geometrically (exponentially) "forget" past observations on the system, i.e., one chooses $\hat{\theta}(t)$ so that it minimizes

$$\sum_{n=0}^{t} \lambda^{t-n}(y(t)-\phi^T(t)\theta)^2$$

over all θ. The factor λ, $0<\lambda\leqq 1$, is called the "exponential forgetting factor". The case $\lambda=1$ is the case that has been studied so far in this paper. Even if $\lambda<1$, one can obtain recursive schemes for the parameter estimates. The only change from § 8.1.2 is that

$$\hat{\theta}(t+1)=\hat{\theta}(t)+R_\lambda^{-1}(t)\phi(t)[y(t+1)-\hat{\theta}(t)^T\phi(t)]$$

where

$$R_\lambda(t)=\lambda R_\lambda(t-1)+\phi(t)\phi^T(t).$$

Alternatively, one can also obtain a recursive expression

$$R_\lambda^{-1}(t+1)=\frac{1}{\lambda}R_\lambda^{-1}(t)-\frac{1}{\lambda}\frac{R_\lambda^{-1}(t)\phi(t+1)\phi(t+1)^TR_\lambda^{-1}(t)}{\lambda+\phi^T(t+1)R_\lambda^{-1}(t)\phi(t+1)}.$$

During time intervals when the system under control is not changing, one would like to keep $\lambda \approx 1$ whereas if the system is changing, then one wants to keep $\lambda < 1$. Some practical long term problems (burst, blow up) can result from the choice of a constant forgetting factor, see Fortescue, Kershenbaum and Ydstie [131] and [123], Sanoff and Wellstead [144] and Saelid and Foss [145]. [131] proposes an "adaptive" selection of $\lambda(t)$ so that a measure of "information content" is kept constant. Latawiec and Chyra [132] also discuss this problem and offer some solutions. Lozano L. [133] obtains a bound on the asymptotic variance of the error in the parameter estimates. Zarrop [134] investigates the rate at which the forgetting factor sequence $\{\lambda(t)\}$ should converge to 1 so that asymptotic consistency of estimates can still be obtained.

A different approach to the situation of time varying parameters is taken by Caines [142] and Chen and Caines [143]. [142] analyzes the situation where the parameters form a converging martingale, while [143] analyzes the situation where the parameters constitute a uniformly bounded martingale difference sequence plus a constant.

Before leaving the topic of this section, we mention the work of Kalman [135], who as early as 1957, made a very strong case for considering schemes of this sort, and actually built a computer to implement them. Least squares estimates, forgetting factors, deadbeat control laws, efficient recursive least squares estimates, etc. all are ingredients of his scheme. See also the work of Peterka [136] and Peterka and Åström [137].

9. Conclusions. Clearly, much has been done and still much more remains to be done. For the Bayesian problems, efficient computational methods or analytic solutions to new problems are still needed. In the area of "dual control" of linear quadratic (Gaussian) systems, one needs approximations for which rigorous bounds on the quality of the approximations are available. For the adaptive control of Markov chains, one is faced with spaces of huge cardinality when the state spaces, control spaces etc. are large but finite. Further, efficient schemes to implement the algorithms are needed as well as studies of rates of convergence. In the self-tuning area, we still do not have theoretical tools to analyze all the schemes which have been proposed; in fact the original self-tuning regulator has yet to be fully analyzed. Rates of convergence have not been adequately established yet. The problem of *robustness* of the *adaptive* scheme has not been rigorously examined. An analysis of the steady state of the self-tuning regulator with a forgetting factor or periodic resetting is not available.

Clearly much more is needed in the way of theory.

Acknowledgments. The author is grateful to A. Becker, W. Lin, S. Mitter, S. Sastry, U. Shaked, P. Varaiya, J. Walrand and R. Weber for useful discussions. The author also wishes to acknowledge the special role played by H. Kushner in the development of this paper. Without his original suggestion and subsequent encouragement this paper would not have been written. The author also wishes to thank P. Varaiya and J. Walrand for their friendly hospitality during his visit to the University of California, Berkeley and S. Mitter during his visit to M.I.T.

REFERENCES

[1] D. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[2] R. BELLMAN, *Adaptive Control Processes: A Guided Tour*, Princeton Univ. Press, Princeton, NJ, 1961.

[3] E. B. DYNKIN, *Controlled random sequences*, Theory Prob. Appl., 10 (1965), pp. 1-14.

[4] M. AOKI, *Optimal control of partially observable Markovian systems*, J. Franklin Institute, 280 (1965), pp. 367-386.

[5] K. ÅSTRÖM, *Optimal control of Markov processes with incomplete state information*, J. of Math. Anal. Appl., 10 (1965), pp. 174-205.
[6] A. N. SHIRYAEV, *Some new results in the theory of controlled random processes*, English translation in Selected Translations in Mathematical Statistics and Probability, 8 (1970), pp. 49-130.
[7] C. STRIEBEL, *Sufficient statistics in the optimal control of stochastic systems*, J. of Math. Anal. Appl., 12 (1965), pp. 576-592.
[8] ———, *Optimal Control of Discrete Time Stochastic Systems*, Springer-Verlag, New York, 1975.
[9] K. HINDERER, *Foundations of Nonstationary Dynamic Programming with Discrete Time Parameter*, Springer-Verlag, New York, 1970.
[10] Y. SAWARAGI AND T. YOSHIKAWA, *Discrete time Markovian decision processes with incomplete state information*, Ann. Math. Stat., 41 (1970), pp. 78-86.
[11] D. RHENIUS, *Incomplete information in Markovian decision models*, Ann. Stat., 2 (1974), pp. 1327-1334.
[12] J. J. MARTIN, *Bayesian Decision Problems and Markov Chains*, John Wiley, New York, 1967.
[13] U. RIEDER, *Bayesian dynamic programming*, Adv. Appl. Prob., 7 (1975), pp. 330-348.
[14] K. M. VAN HEE, *Bayesian control of Markov chains*, Mathematical Center Tracts 95, Mathematisch Centrum, Amsterdam, 1978.
[15] J. C. GITTINS AND D. M. JONES, *A dynamic allocation index for the sequential design of experiments*, in Colloquia Mathematica Societatis Janos Bolyai 9, Progress in Statistics, European Meeting of Statisticians, pp. 241-266, J. Gani, K. Sarkadi and I. Vincze, eds., North-Holland, London, 1972.
[16] J. C. GITTINS AND K. D. GLAZEBROOK, *On Bayesian models in stochastic scheduling*, J. of Appl. Prob., 14 (1977), pp. 556-565.
[17] P. NASH, *Optimal allocation of resources between research projects*, Ph.D. Thesis, Cambridge University, 1973.
[18] P. WHITTLE, *Multi-armed bandits and the Gittins index*, J. Royal Stat. Soc., 42B (1980), pp. 143-149.
[19] K. D. GLAZEBROOK, *Optimal strategies for families of alternative bandit processes*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 858-861.
[20] P. WHITTLE, *Optimization Over Time: Dynamic Programming and Stochastic Control*—I, John Wiley, New York, 1982.
[21] S. ROSS, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.
[22] J. C. GITTINS, *Bandit processes and dynamic allocation indices*, J. Royal Stat. Soc., 41B (1979), pp. 148-177.
[23] K. D. GLAZEBROOK, *On a sufficient condition for superprocesses due to Whittle*, J. of Appl. Prob., 19 (1982), pp. 99-110.
[24] P. WHITTLE, *Arm-acquiring bandits*, Ann. of Prob., 9 (1981), pp. 284-292.
[25] P. VARAIYA, J. WALRAND AND C. BUYUKKOC, *Extensions of the multi-armed bandit problem: The discounted case*, Univ. California, Berkeley, 1983.
[26] M. ROTHSCHILD, *A two-armed bandit theory of market pricing*, J. Economic Theory, 9 (1974), pp. 185-202.
[27] F. P. KELLY, *Multi-armed bandits with discount factor near one: The Bernoulli case*, Ann. Stat., 9 (1981), pp. 897-1001.
[28] K. D. GLAZEBROOK, *On randomized dynamic allocation indices for the sequential design of experiments*, J. Royal Stat. Soc., 42B (1980), pp. 342-346.
[29] J. BATHER, *Randomized allocation of treatments in sequential trials*, Adv. Appl. Prob., 12 (1980), pp. 174-182.
[30] K. D. GLAZEBROOK, *On the evaluation of suboptimal strategies for families of alternative bandit processes*, J. of Appl. Prob., 19 (1982), pp. 716-722.
[31] P. R. KUMAR AND T. I. SEIDMAN, *On the optimal solution of the one-armed bandit adaptive control problem*, IEEE Trans. on Automat. Control, AC-26 (1981), pp. 1176-1184.
[32] D. BLACKWELL, *Discounted dynamic programming*, Ann. Math. Stat., 36 (1965), pp. 226-235.
[33] R. HOWARD, *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, 1960.
[34] J. K. SATIA AND R. E. LAVE, *Markov decision processes with uncertain transition probabilities*, Oper. Res., 21 (1973), pp. 728-740.
[35] A. M. MAKOWSKI, *Results on the filtering problem for linear systems with non-Gaussian initial conditions*, Proc. 21st IEEE Conference on Decision and Control, 1982, pp. 201-204.
[36] K. J. ÅSTRÖM AND B. WITTENMARK, *Problems of identification and control*, J. Math. Anal. Appl., 34 (1971), pp. 90-113.
[37] A. A. FELDBAUM, *The theory of dual control* IV, Automation and Remote Control, 22 (1961), pp. 109-121.
[38] O. L. R. JACOBS AND J. W. PATCHELL, *Caution and probing in stochastic control*, Internat. J. Control, 16 (1972), pp. 189-199.

[39] E. TSE AND Y. BAR-SHALOM, *Wide-sense adaptive dual control for nonlinear stochastic systems*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 98–108.

[40] ———, *An actively adaptive control for linear systems with random parameters via the dual control approach*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 109–117.

[41] ———, *Generalized certainty equivalence and dual effect in stochastic control*, IEEE Trans. Automat. Control, AC-20 (1975), pp. 817–819.

[42] ———, *Dual effect certainty equivalence, and separation in stochastic control*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 494–500.

[43] B. WITTENMARK, *Stochastic adaptive control methods: a survey*, Internat. J. of Control, 21 (1975), pp. 705–730.

[44] C. J. WENK AND Y. BAR-SHALOM, *A multiple model adaptive dual control algorithm for stochastic systems with unknown parameters*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 703–710.

[45] Y. BAR-SHALOM, *Stochastic dynamic programming: caution and probing*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1184–1194.

[46] J. G. DESHPANDE, T. N. UPADHYAY AND D. G. LAINIOTIS, *Adaptive control of linear stochastic systems*, Automatica, 9 (1973), pp. 107–115.

[47] D. G. LAINIOTIS, *Partitioning: a unifying framework for adaptive systems*, II: *control*, Proc. IEEE, 64 (1976), pp. 1182–1198.

[48] P. L. DERSIN, M. ATHANS AND D. A. KENDRICK, *Some properties of the dual adaptive stochastic control algorithm*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 1001–1008.

[49] H. ROBBINS, *Some aspects of the sequential design of experiments*, Bull. Amer. Math. Soc., 58 (1952), pp. 527–537.

[50] V. BORKAR AND P. VARAIYA, *Adaptive control of Markov chains*, I: *finite parameter set*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 953–958.

[51] P. R. KUMAR AND A. BECKER, *A new family of optimal adaptive controllers for Markov chains*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 137–146.

[52] B. SAGALOVSKY, *Adaptive control and parameter estimation in Markov chains: a linear case*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 414–419.

[53] P. R. KUMAR, *Adaptive control with a compact parameter set*, this Journal, 20 (1982), pp. 9–13.

[54] V. BORKAR AND P. VARAIYA, *Identification and adaptive control of Markov chains*, this Journal, 20 (1982), pp. 470–489.

[55] P. MANDL, *Estimation and control in Markov chains*, Adv. Appl. Prob., 6 (1974), pp. 40–60.

[56] ———, *On the adaptive control of finite state Markov processes*, Z. Wahrsch. Verw. Geb., 27 (1973), pp. 263–276.

[57] ———, *On the control of a Markov chain in the presence of unknown parameters*, Trans. Sixth Prague Conference on Information Theory, Random Processes and Statistical Decision Functions, Prague, 1971, pp. 601–612.

[58] M. KURANO, *Discrete-time Markovian decision processes with an unknown parameter-average return criterion*, J. Oper. Res. Soc. Japan, 15 (1972), pp. 67–76.

[59] M. KOLONKO, *Strongly consistent estimation in a controlled Markov renewal model*, J. Appl. Prob., 19 (1972), pp. 532–545.

[60] ———, *The average-optimal adaptive control of a Markov renewal model in presence of an unknown parameter*, Math. Operations Forsch. Statist. Ser. Optim., 13 (1982), pp. 567–591.

[61] J. P. GEORGIN, *Estimation et contrôles des chaînes de Markov sur des espaces arbitraires*, in Lecture Notes in Mathematics, 636, Springer-Verlag, Berlin, 1978.

[62] V. V. BARANOV, *Recursive algorithms of adaptive control in stochastic systems*, Cybernetics, 17 (1981), pp. 815–824.

[63] M. SCHÄL, *Estimation and control in stochastic dynamic programming: finite and asymptotic results*, Report No. 521, Univ. Bonn, July 1982.

[64] B. L. FOX AND J. E. ROLPH, *Adaptive policies for Markov renewal programs*, Ann. Stat., 1 (1973), pp. 334–341.

[65] B. DOSHI AND S. SHREVE, *Strong consistency of a modified maximum likelihood estimator for controlled Markov chains*, J. Appl. Prob., 17 (1980), pp. 726–734.

[66] M. SATO, K. ABE AND H. TAKEDA, *Learning control of finite Markov chains with unknown transition probabilities*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 502–505.

[67] L. M. LYUBCHIK AND A. S. POZNYAK, *Learning automata in stochastic plant control problems*, Automation and Remote Control, 6 (1974), pp. 777–789.

[68] Y. M. EL-FATTAH, *Recursive algorithms for adaptive control of finite Markov chains*, IEEE Trans. Systems, Man and Cybernetics, SMC-11 (1981), pp. 135–144.

[69] Y. M. EL-FATTAH, *Gradient approach for recursive estimation and control in finite Markov chains*, Adv. Appl. Prob., 13 (1981), pp. 778–803.

[70] Y. Z. TSYPKIN, *Adaptation and Learning in Automatic Systems*, Academic Press, New York, 1971.
[71] Y. Z. TSYPKIN, *Foundations of the Theory of Learning Systems*, Academic Press, New York, 1973.
[72] J. NEVEU, *Discrete-Parameter Martingales*, North-Holland, Amsterdam, 1975.
[73] H. ROBBINS AND D. SIEGMUND, *A convergence theorem for nonnegative almost super-martingales and some applications*, in Optimization Methods in Statistics, J. S. Rustagi, ed., Academic Press, New York, 1971, pp. 233-257.
[74] B. T. POLYAK AND Y. Z. TSYPKIN, *Pseudogradient adaptation and training algorithms*, Automation and Remote Control, 23 (1973), pp. 377-397.
[75] G. GOODWIN, P. RAMADGE AND P. CAINES, *Discrete time stochastic adaptive control*, this Journal, 19 (1981), pp. 829-853.
[76] P. R. KUMAR AND W. LIN, *Optimal adaptive controllers for unknown Markov chains*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 765-774.
[77] P. R. KUMAR, *Simultaneous identification and adaptive control of unknown systems over finite parameter sets*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 68-76.
[78] ———, *Optimal adaptive control of linear-quadratic-Gaussian systems*, this Journal, 21 (1983), pp. 163-178.
[79] A. BECKER AND P. R. KUMAR, *Optimal strategies for the N-armed bandit problem*, Math. Research Report No. 81-1, Univ. Maryland Baltimore County, 1981.
[80] J. RIORDAN, *An adaptive automaton controller for discrete time Markov processes*, Automatica, 5 (1969), pp. 721-730.
[81] F. M. D'HULSTER, R. M. C. DEKEYSER AND A. R. VAN CAUWENBERGHE, *Simulations of adaptive controllers for a paper machine headbox*, Automatica, 19 (1983), pp. 407-414.
[82] C. KIPARISSIDES AND S. L. SHAH, *Self tuning and stable adaptive control of a batch polymerization reactor*, Automatica, 19 (1983), pp. 223-224.
[83] G. DUMONT, *Self-tuning control of a chip refiner motor load*, Automatica, 18 (1982), pp. 307-314.
[84] H. BOEHM, *Adaptive control to a dry etch process by microcomputer*, Automatica, 18 (1982), pp. 665-673.
[85] K. Y-J. KO, B. C. MCINNIS AND G. C. GOODWIN, *Adaptive control and identification of the dissolved oxygen process*, Automatica, 18 (1982), pp. 727-730.
[86] T. L. LAI AND C. Z. WEI, *Least squares estimates in stochastic regression with applications to identification and control of dynamic systems*, Ann. Stat., 10 (1982), pp. 154-166.
[87] G. C. GOODWIN AND R. L. PAYNE, *Dynamic System Identification*, Academic Press, New York, 1977.
[88] V. SOLO, *The convergence of* AML, IEEE Trans. Automat. Control, AC-24 (1979), pp. 958-962.
[89] K. J. ÅSTRÖM, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.
[90] ———, *Lectures on the identification problem—the least squares method*, Report 6806, Lund Institute of Technology, 1968.
[91] V. PETERKA, *On steady state minimum variance control strategy*, Kybernetika, 8 (1972), pp. 218-231.
[92] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica, 9 (1973), pp. 185-199.
[93] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 551-575.
[94] ———, *On positive real transfer functions and the convergence of some recursive schemes*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 539-551.
[95] H. KUSHNER, *Convergence of recursive adaptive and identification procedures via weak convergence theory*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 921-930.
[96] H. KUSHNER AND D. S. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, 1978.
[97] J.-J. J. FUCHS, *Explicit self-tuning methods*, Proc. IEE, 127 (1980), pp. 259-264.
[98] G. C. GOODWIN, K. S. SIN AND K. K. SALUJA, *Stochastic adaptive control and prediction—the general delay-colored noise case*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 946-949.
[99] J.-J. J. FUCHS, *Indirect stochastic adaptive control: the general delay-white noise case*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 219-223.
[100] ———, *Indirect stochastic adaptive control: the general delay-colored noise case*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 470-472.
[101] H. KUSHNER AND R. KUMAR, *Convergence and rate of convergence of a recursive identification and adaptive control scheme which uses truncated estimators*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 775-782.
[102] R. KUMAR AND J. B. MOORE, *Convergence of adaptive minimum variance algorithms via weighting coefficient selection*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 146-153.
[103] K. S. SIN AND G. C. GOODWIN, *Stochastic adaptive control using a modified least squares algorithm*, Automatica, 18 (1982), pp. 315-321.
[104] Z. YOU-HONG, *Stochastic adaptive control and prediction based on a modified least squares—the general delay-colored noise case*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 1257-1260.

[105] P. J. GAWTHROP, *On the stability and convergence of a self-tuning controller*, Internat. J. Control, 31 (1980), pp. 973-998.

[106] R. KUMAR, *Almost sure convergence of adaptive identification prediction and control algorithms*, LCDS 81-8, Div. Applied Mathematics, Brown Univ., Providence, RI, 1981.

[107] A. BECKER, P. R. KUMAR AND C. Z. WEI, *Adaptive control with the stochastic approximation algorithm: geometry and convergence*, Univ. Maryland Baltimore County Mathematics Research Report No. 83-8, 1983, IEEE Trans. Automat. Control, to appear.

[108] P. E. CAINES AND S. LAFORTUNE, *Adaptive optimization with recursive identification for stochastic linear systems*, McGill Univ., Montreal, 1981.

[109] H. F. CHEN, *Recursive system identification and adaptive control by use of the modified least squares algorithm*, McGill Univ., Montreal, 1983.

[110] H. F. CHEN AND P. E. CAINES, *The strong consistency of the stochastic gradient algorithm of adaptive control*, McGill Univ., Montreal, 1983.

[111] K. ÅSTRÖM AND B. WITTENMARK, *Analysis of a self-tuning regulator for non-minimum phase systems*, Proc. IFAC Symposium on Stochastic Control, Budapest, Hungary, 1974.

[112] D. W. CLARKE AND P. J. GAWTHROP, *Self-tuning controller*, Proc. IEEE, 122 (1975), pp. 929-934.

[113] P. J. GAWTHROP, *Some interpretations of the self-tuning controller*, Proc. IEEE, 124 (1977), pp. 889-894.

[114] D. W. CLARKE AND P. J. GAWTHROP, *Self-tuning control*, Proc. IEEE, 126 (1979), pp. 633-640.

[115] H. N. KOIVO, *A multivariable self-tuning controller*, Automatica, 16 (1980), pp. 351-366.

[116] U. BORISSON, *Self-tuning regulators for a class of multivariable systems*, Automatica, 15 (1979), pp. 209-215.

[117] CHEN HAN-FU, *Self-tuning controller and its convergence under correlated noise*, Internat. J. Control, 35 (1982), pp. 1051-1059.

[118] P. E. WELLSTEAD, J. M. EDMUNDS, D. PRAGER AND P. ZANKER, *Self-tuning pole/zero assignment regulators*, Internat. J. Control, 30 (1979), pp. 1-26.

[119] P. J. GAWTHROP, *A comment on "self-tuning pole/zero assignment regulators*, Internat. J. Control, 31 (1980), pp. 999-1002.

[120] P. E. WELLSTEAD AND S. P. SANOFF, *Extended self-tuning algorithm*, Internat. J. Control, 34 (1981), pp. 433-455.

[121] P. E. WELLSTEAD AND P. ZANKER, *Servo self-tuners*, Internat. J. Control, 30 (1979), pp. 27-36.

[122] A. Y. ALLIDINA AND F. M. HUGHES, *Generalized self-tuning controller with pole assignment*, Proc. IEEE, 127D (1980), pp. 13-18.

[123] K. J. ÅSTRÖM AND B. WITTENMARK, *Self-tuning controllers based on pole-zero placement*, Proc. IEEE, 127D (1980), pp. 120-130.

[124] K. J. ÅSTRÖM, U. BORISSON, L. LJUNG AND B. WITTENMARK, *Theory and applications of self-tuning regulators*, Automatica, 13 (1977), pp. 457-476.

[125] P. MANDL, *The use of optimal stationary policies in the adaptive control of linear systems*, Proc. Symposium to Honour Jerzy Neyman, Warsaw, 1974, pp. 223-243.

[126] ———, *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.

[127] M. J. GRIMBLE, *A control weighted minimum-variance controller for non-minimum phase systems*, Internat. J. Control, 33 (1981), pp. 751-762.

[128] ———, *Weighted minimum-variance self-tuning control*, Internat. J. Control, 36 (1982), pp. 597-609.

[129] R. KUMAR AND J. B. MOORE, *Minimum variance control harnessed for non-minimum-phase plants*, Technical Report No. EE8014, July 1980.

[130] ———, *On adaptive minimum variance regulation for non-minimum phase plants*, Automatica, 19 (1983), pp. 449-451.

[131] T. R. FORTESCUE, L. S. KERSHENBAUM AND B. E. YDSTIE, *Implementation of self-tuning regulators with variable forgetting factors*, Automatica, 17 (1981), pp. 831-835.

[132] K. LATAWIEC AND M. CHYRA, *On low frequency and long-run effects in self-tuning control*, Automatica, 19 (1983), pp. 419-424.

[133] R. LOZANO L., *Convergence analysis of recursive identification algorithms with forgetting factor*, Automatica, 19 (1983), pp. 95-97.

[134] M. B. ZARROP, *Variable forgetting factors in parameter estimation*, Automatica, 19 (1983), pp. 295-298.

[135] R. KALMAN, *Design of a self-optimizing control system*, Trans. ASME, 80 (1958), pp. 468-478.

[136] V. PETERKA, *Adaptive digital regulation of noisy systems*, Proc. 2nd IFAC Symposium on Identification and Process Parameter Estimation, Prague, 1970.

[137] V. PETERKA AND K. J. ÅSTRÖM, *Control of multivariate systems with unknown but constant parameters*, Proc. 3rd IFAC Symposium, Hague/Delft, 1973.

[138] U. SHAKED AND P. R. KUMAR, *Minimum variance control of multivariable* ARMAX *systems*, LIDS Report, Massachusetts Institute of Technology, Cambridge, 1984.

[139] M. METIVIER AND P. PRIOURET, *Applications of a Kushner and Clark lemma to general classes of stochastic algorithms*, IEEE Trans. Inform. Theory, IT-30, Part 1 (1984), pp. 140–151.

[140] R. RISHEL, *An exact formula for a linear quadratic adaptive stochastic optimal control law*, Preprint, 1984.

[141] O. HIJAB, *Optimal adaptive control and stabilization of families of linear systems*, Preprint, 1983, Systems and Control Letters, to appear.

[142] P. CAINES, *Stochastic adaptive control: randomly varying parameters and continually disturbed controls*, Proc. 8th IFAC Congress, Kyoto, 1981, pp. 925–930.

[143] H. F. CHEN AND P. E. CAINES, *On the adaptive control of a class of systems with random parameters and disturbances*, Preprint, 1983.

[144] S. P. SANOFF AND P. E. WELLSTEAD, *Comments on "Implementation of self-tuning regulators with variable forgetting factors,"* Automatica, 19 (1983), pp. 345–346.

[145] S. SAELID AND B. FOSS, *Adaptive controllers with a vector variable forgetting factor*, Proc. 22nd IEEE Conference on Decision and Control 3, San Antonio, 1983, pp. 1488–1494.

[146] R. KUMAR, *Simultaneous adaptive control and identification via the weighted least-square algorithm*, IEEE Trans. Automat. Control, AC-29 (1984), pp. 259–263.

[147] D. W. CLARKE AND P. J. GAWTHROP, *Comments on "On adaptive minimum variance regulation for non-minimum phase plants"*, Automatica, 20 (1984), pp. 261.

[148] V. SOLO, *The convergence of an instrumental-variable-like recursion*, Automatica, 17 (1981), pp. 545–547.

[149] H. KUSHNER AND A. SHWARTZ, *An invariant measure approach to the convergence of stochastic approximations with state dependent noise*, this Journal, 22 (1984), pp. 13–27.

[150] H. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes*, MIT Press, Cambridge, MA, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

PROPERTIES OF MATRICES USED IN UNCERTAIN LINEAR CONTROL SYSTEMS*

L. G. CHOUINARD†, J. P. DAUER‡ AND G. LEITMANN§

Abstract. In this paper a number of questions arising in the design of state feedback controls for uncertain dynamical systems are answered. These results are algebraic in nature and include results on positive definite solutions of Lyapunov's equation and on the rotation of switching surfaces and attractive surfaces for min-max controls.

Key words. Lyapunov's equation, feedback control, uncertain linear systems, guaranteed performance

1. Introduction. A number of algebraic questions arise in the study of stabilizing feedback controllers for uncertain dynamical systems (see e.g. [2], [3], [4]). Briefly, the construction of proper quadratic Lyapunov functions for stability presents problems on the positive definite properties of matrices in Lyapunov's equation. Also the attractivity of the switching surface of a feedback controller (i.e. solutions tend to move to the surface) corresponds to properties of the null space of a matrix. In this paper we approach these algebraic problems and show their relationship to the theory of stabilizing uncertain dynamical systems.

In § 2 we consider the problem; given a stable matrix $\bar{A}$, for which positive definite matrices, P, is the matrix

$$Q = -(P\bar{A} + \bar{A}^T P)$$

also positive definite. This problem arises in the contruction of a positive definite Lyapunov function, $L(x) = x^T P x$, which is decreasing along solutions of a feedback system. This is a central point in Lyapunov-type stability or boundedness results. The results of Breinl and Leitmann [3] show that for uncertain linear control systems this problem is particularly important when $P = P_R$ is the solution of a Riccati equation. The first result of § 2 gives a necessary and sufficient condition in this case. If P_R does not yield a positive definite Q, then the second result shows that appropriate linear modifications, $P = P_R + P_N$, also will not yield a positive definite Q. The last two results in this section are necessary and sufficient conditions and techniques for constructing matrices P for which the resulting Q is positive definite. These results can be used to construct a matrix P which in some case retains the natural sensitivity properties of P_R. The results of § 2 are applied to uncertain dynamical systems in § 3.

The result in § 4 is concerned with the attractivity of the switching surface of a nonlinear feedback control for an uncertain linear control system. This result gives necessary and sufficient conditions for the control matrix which, when applied to min-max controllers [4], guarantees that the switching surface is attractive in the presence of uncertainties.

* Received by the editors April 19, 1983, and in revised form March 15, 1984.

† Department of Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588.

‡ Department of Mathematics and Statistics, University of Nebraska, Lincoln, Nebraska 68588. The work of this author was supported by the Office of Water Research and Technology, U.S. Department of the Interior.

§ Department of Mechanical Engineering, University of California, Berkeley, California 94720. The work of this author was supported by the National Science Foundation under grant ECS-7813931 and carried out while the author was the recipient of a U.S. Senior Scientist Award of the Alexander von Humboldt Foundation.

2. Properties of the solution of Lyapunov's equation. The basic control system to be considered is linear of the form

$$\dot{x} = Ax + Bu, \qquad x(0) = x_0, \tag{1}$$

with A an $n \times m$ matrix, B an $n \times m$ matrix, and $\dot{x} = dx/dt$. The performance objective is given by a quadratic cost functional

$$J[x(t), u(t)] = \int_0^\infty [x^T(t) Q_R x(t) + u^T(t) R u(t)]\, dt, \tag{2}$$

where $Q_R = D^T D$ is symmetric, positive semidefinite and R is symmetric, positive definite. We assume that (A, B) is completely controllable and (A, D) is completely observable. The unique optimal control for system (1), (2) is given by

$$u_l = -R^{-1} B^T P_R x, \tag{3}$$

where the symmetric, positive definite matrix P_R solves the Riccati equation

$$A^T P_R + P_R A - P_R B R^{-1} B^T P_R = Q_R = 0 \tag{4}$$

[1]. The optimal control (3) in system (1) yields the asymptotically stable closed-loop system

$$\dot{x} = (A - BR^{-1} B^T P_R) x, \qquad x(0) = x_0, \tag{5}$$

where the matrix

$$\bar{A} = A - BR^{-1} B^T P_R \tag{6}$$

is stable; i.e., all eigenvalues of $\bar{A}$ have negative real parts. Therefore the optimal control u_l of (3) stabilizes system (1) through the feedback system (5). Further, $(\bar{A}, B)$ is completely controllable [1, p. 48].

The problem of guaranteed stability (or ultimate boundedness) of the solutions of a control system is more complex when the parameters in system (1) are uncertain. A typical approach is to construct a nonlinear feedback control corresponding to a specific Lyapunov function, $V(x) = x^T P x$. This Lyapunov function is then used to analyze the solution behavior of the system [2]–[4]. Of fundamental importance in this type of analysis is Lyapunov's equation

$$P\bar{A} + \bar{A}^T P + Q = 0, \tag{7}$$

where an appropriate matrix Q, or perhaps P, is usually chosen to be symmetric, positive definite.

If we let $\bar{A}$ and Q be given matrices, with $\bar{A}$ having eigenvalues $\lambda_1, \cdots, \lambda_n$, then equation (7) has a unique solution P if and only if $\lambda_i + \lambda_j \neq 0$ for all $1 \leqq i, j \leqq n$, [5] (see also [6], [7]). It is easy to see that if (7) has any symmetric solution P, then Q must also be symmetric. Further, if Q is symmetric and the solution P of (7) exists and is unique, then P must also be symmetric.

PROPOSITION 2.1. *If $\bar{A}$ is stable and Q is positive semidefinite, then* (7) *has the unique positive semidefinite solution*

$$P = \int_0^\infty (\exp \bar{A}^T t) Q (\exp \bar{A} t)\, dt. \tag{8}$$

It is easy to see that if Q is positive definite, then P is positive definite. However, one problem in control design that arises (for references see [3]), and will be approached in this work, is that if $\bar{A}$ is stable and P is a given symmetric, positive definite matrix,

it need not follow that Q is positive definite. Of particular interest in applications is this problem for $P = P_R$, the solution of (4), or matrices P related to P_R.

THEOREM 2.1. *Suppose P_R solves* (4); *then the matrix Q given by*

$$P_R\bar{A} + \bar{A}^T P_R + Q = 0 \tag{9}$$

is positive semidefinite. Q is positive definite if and only if rank $(Q_R, P_R B) = n$.

For any matrix H, the null space of H is denoted by $N(H) = \{y: Hy = 0\}$.

Proof. Taking $P = P_R$, equations (9), (6) and (4) yield

$$\begin{aligned} -Q = \bar{A}^T P + P\bar{A} &= A^T P - PBR^{-1}B^T P + PA - PBR^{-1}B^T P \\ &= -Q_R - PBR^{-1}B^T P. \end{aligned}$$

Since the matrices Q_R and $PBR^{-1}B^T P$ are both positive semidefinite (R is positive definite), we have that Q is positive semidefinite. Therefore, Q is not positive definite if and only if there is a nonzero y such that

$$0 = y^T Q y = y^T Q_R y + y^T PBR^{-1}B^T P y. \tag{10}$$

Since Q_R and $PBR^{-1}B^T P$ are both positive semidefinite and R^{-1} is positive definite, equation (10) is satisfied when $y \in N(D)$ and $y \in N(B^T P)$. Since $N(D) = N(Q_R)$, we must have $Q_R y = 0$ and $B^T P_R y = 0$. Using the symmetry of Q_R and P_R it follows that there is a nonzero y satisfying (10) if and only if rank $(Q_R, P_R B) < n$. □

Remark 2.1. If Q_R is positive definite, then Q in (9) is positive definite (see also [1, p. 41]). If Q_R is not positive definite (but only positive semidefinite), then Q in (9) need not be positive definite even if the system $\dot{x} = Ax + Bu$ is in canonical form [8]. Examples can be easily constructed by noting that rank $B^T P_R$ = rank B and therefore in order for Q to be positive definite it is necessary (but not sufficient) that

$$\text{rank } B + \text{rank } Q_R \geqq n.$$

Remark 2.2. In the ultimate boundedness results for uncertain linear systems of Leitmann [2] a nonlinear feedback control is used. This control is of the form $u_n(t) = p(x(t))$, where for given $\varepsilon > 0$

$$p(x) = \begin{cases} -\dfrac{B^T P x}{\|B^T P x\|}\rho(x) & \text{for } \|B^T P x\| > \varepsilon, \\ -\dfrac{B^T P x}{\varepsilon}\rho(x) & \text{for } \|B^T P x\| \leqq \varepsilon, \end{cases} \tag{11}$$

and $\rho(x)$ is a norm bound of the "lumped" uncertainty. The desired stability results can be assured using control matrix $P = P_R$ if Q in (9) is positive definite only on the switching surface $N(B^T P_R)$; i.e., $x^T Q x > 0$ for $x \in N(B^T P_R)$, $x \neq 0$. However, it follows from (10) that Q is positive definite if and only if Q is positive definite on $N(B^T P_R)$.

Suppose the feedback control matrix $P = P_R$ does not produce a Q in (9) which is positive definite. The desired robustness can still be expected [3] from a feedback matrix P_L satisfying

$$N(B^T P_L) = N(B^T P_R)$$

provided the matrix Q_L in

$$P_L\bar{A} + \bar{A}^T P_L + Q_L = 0 \tag{12}$$

is positive definite. A standard approach to accomplish this and yet to retain the stabilizing properties of P_R is to consider a positive definite control matrix of the form

$P_L = P_R + P_N$ satisfying $B^T P_L = B^T P_R$ (i.e., $B^T P_N = 0$) such that the matrix Q_L in (12) is positive definite. However, the following result shows that this is not possible.

THEOREM 2.2. *The only matrix of the form $P_L = P_R + P_N$ which satisfies $B^T P_N = 0$ and is such that P_L and P_N satisfy Lyapunov's equation* (12) *for symmetric, positive semidefinite Q is $P_L = P_R$.*

Proof. Take $P_L = P_R + P_N$, where $B^T P_N = 0$. Then (12) becomes

$$-Q_L = P_R\bar{A} + \bar{A}^T P_R + P_N\bar{A} + \bar{A}^T P_N.$$

Then it follows from Theorem 2.1 and (10) that $(P_R\bar{A} + \bar{A}^T P_R)$ is negative semidefinite and, in fact, negative definite for $y \notin N(D) \cap N(B^T P_R)$. So we need to select P_N so that $(P_N\bar{A} + \bar{A}^T P_N) = -Q_N$ is negative semidefinite. By Proposition 2.1 the solution of (7) with $Q_N = G^T G$ is the positive semidefinite matrix

$$P_N = \int_0^\infty (\exp \bar{A}t)^T G^T G(\exp \bar{A}t)\, dt.$$

Since $B^T P_N = 0$, it follows that

$$0 = \int_0^\infty B^T (\exp \bar{A}t)^T G^T G(\exp \bar{A}t)\, dt.$$

Therefore,

$$0 = \int_0^\infty B^T(\exp \bar{A}t)^T G^T G(\exp \bar{A}t) B\, dt,$$

which implies $G(\exp \bar{A}t)B = 0$ for $t \geqq 0$. Since the converse clearly follows, we have that $B^T P_N = 0$ if and only if

$$G(\exp \bar{A}t)B \equiv 0 \quad \text{for } t \geqq 0. \tag{13}$$

Using that $(\bar{A}, B)$ is completely controllable and a standard controllability space argument [8, pp. 81-82], equation (13) holds if and only if $G = 0$. Therefore, $B^T P_N = 0$ if and only if $P_N = 0$. □

The "failure" of the additive modification of P_R in Theorem 2.2 motivates us to consider an alternative approach to the construction of a feedback matrix P when the conditions of Theorem 2.1 are not satisfied. Recall that $\bar{A}$ is an $n \times n$ matrix whose eigenvalues have negative real parts and suppose that P is a specified $n \times n$ symmetric positive definite matrix. Let Q be the $n \times n$ symmetric matrix defined by Lyapunov's equation (7),

$$P\bar{A} + \bar{A}^T P = -Q.$$

Then $-\frac{1}{2}Q$ is the symmetric part of the matrix $P\bar{A}$. Therefore, Q is positive definite if and only if $P\bar{A}$ is negative definite (and usually nonsymmetric).

We now determine conditions for the construction of a positive definite P with the property that Q is positive definite. This construction, at times, allows us to retain some of the "natural" stabilizing properties of P_R (see [3]), whereas fixing Q to be positive definite and solving Lyapunov's equation for P yields a matrix which is not related to P_R. Without loss of generality we will restrict ourselves to showing that $P\bar{A}$ is negative definite since this is equivalent to

$$Q = -(P\bar{A} + \bar{A}^T P)$$

being positive definite.

Consider the system $\dot{x} = \bar{A}x$ in a coordinate system where $\bar{A}$ is block upper triangular,

$$\bar{A} = \begin{pmatrix} \lambda_1 & & & & & & \\ & \ddots & & & & S & \\ & & \lambda_p & & & & \\ & & & \begin{pmatrix} \alpha_1 & b_1 \\ -c_1 & \alpha_1 \end{pmatrix} & & & \\ & & & & \ddots & & \\ & 0 & & & & \begin{pmatrix} \alpha_k & b_k \\ -c_k & \alpha_k \end{pmatrix} \end{pmatrix}, \tag{14}$$

where $\lambda_1, \cdots, \lambda_p$ are the real eigenvalues of $\bar{A}$, $\alpha_1, \cdots, \alpha_k$ are the real parts of the complex eigenvalues of $\bar{A}$, and $b_i c_i > 0$ for $i = 1, \cdots, k$. The matrix $\bar{A}$ can always be so transformed by $U^T \bar{A} U$ using an orthogonal matrix U.

The first result is a constructive procedure for P so that both P and $-P\bar{A}$ are positive definite when $\bar{A}$ has real eigenvalues. It is inductive in nature and establishes a necessary and sufficient criterion. A similar procedure when $\bar{A}$ has at least one pair of complex eigenvalues will follow. We assume that we have partitioned the matrices as

$$P = \begin{pmatrix} P_1 & M \\ M^T & c \end{pmatrix}, \qquad \bar{A} = \begin{pmatrix} A_1 & \Gamma \\ 0 & \lambda_n \end{pmatrix},$$

where the $(n-1)\times(n-1)$ matrices P_1 and A_1 have the same structure as P and $\bar{A}$, respectively, and P_1, $-P_1 A_1$ are positive definite. Clearly in order for P and $-P\bar{A}$ to be positive definite we must construct P_1 so that this holds.

THEOREM 2.3. *Suppose P_1 and $-P_1 A_1$ are positive definite. Then, given any M, Γ and negative number λ_n, there exists a constant K such that both P and $-P\bar{A}$ are positive definite if and only if $c > K$.*

Proof. To construct K consider the quadratic function

$$q(v) = -(v^T P_1 A_1 v + M^T A_1 v + v^T P_1 \Gamma + v^T M \lambda_n + M^T \Gamma)$$

which has a positive definite Hessian and therefore has a minimum value q^*. Choosing $\bar{K} = \max\{-q^*/\lambda_n, 0\}$ makes $-P\bar{A}$ positive definite if and only if $c > \bar{K}$. Choosing $K = \max\{-q^*/\lambda_n, -\hat{q}, 0\}$ where $\hat{q}$ is the minimum of $v^T P_1 v + 2M^T v$ makes both P and $-P\bar{A}$ positive definite if and only if $c > K$. □

Remark 2.3. Since M is unrestricted, for any such $\bar{A}$ we can construct the matrix P as diagonal; i.e., such a matrix $\bar{A}$ is D-stable [11], [12].

When $\bar{A}$ has a pair of complex conjugate eigenvalues, partition $\bar{A}$ as

$$\bar{A} = \begin{pmatrix} A_1 & \Gamma \\ 0 & \begin{pmatrix} \alpha & b \\ -c & \alpha \end{pmatrix} \end{pmatrix},$$

and P as

$$P = \begin{pmatrix} P_1 & M \\ M^T & \begin{pmatrix} z & k \\ k & y \end{pmatrix} \end{pmatrix},$$

where A_1 and P_1 are $(n-2)\times(n-2)$ and Γ and M are $(n-2)\times 2$. Again, P_1A_1 and P_1 clearly must be positive definite, so we assume P_1 has been so constructed.

THEOREM 2.4. *Suppose P_1 and $-P_1A_1$ are positive definite. Then, given any M, Γ and real numbers α, b, c and k satisfying $\alpha<0$, $bc>0$, there exists an unbounded region in the zy-plane consisting of values which make P and $-P\bar{A}$ (simultaneously) positive defiite.*

Proof.

$$-Q=P\bar{A}+\bar{A}^TP=\begin{pmatrix} P_1A_1+A_1^TP_1 & P_1\Gamma+M\begin{pmatrix}\alpha & b\\ -c & \alpha\end{pmatrix}+A^TM \\ \Gamma^TP_1+\begin{pmatrix}\alpha & -c\\ b & \alpha\end{pmatrix}M^T+M^TA_1 & M^T\Gamma+\Gamma^TM+\begin{pmatrix}2z\alpha-2kc & zb-yc-2k\alpha\\ zb-yc+2k\alpha & 2y\alpha+2kb\end{pmatrix}\end{pmatrix}.$$

Note that the quadratic terms in $\det(Q)$ are just $\det(-P_1A_1-A_1^TP_1)(4zy\alpha^2-(zb-yc)^2)$. The determinant of the $(n-1)\times(n-1)$ leading minor of Q is linear in z (y does not appear) with coefficient $(-2\alpha)\cdot\det(-P_1A_1-A_1^TP_1)>0$ of z. Likewise, $\det(P)$ has quadratic term $zy\cdot\det(P_1)$ and its lead $(n-1)\times(n-1)$ minor is linear in z with coefficient $\det(P_1)$ of the z term. All four of these determinants are therefore positive for sufficiently large z and y satisfying $zb-yc=0$ (which can be found since $bc>0$). The determinant test for symmetric positive definite matrices then applies. □

Remark 2.4. The construction in the proof of Theorem 2.4 specifies four polynomials whose values have to be positive for (z, y) in order to make P and $-P\bar{A}$ positive definite. These polynomials can be calculated by standard techniques for finding determinants using row and/or column operations.

3. Applications to uncertain linear systems. Consider a control system of the form (1) where the parameters are uncertain. (See [2]-[4].) Simulation and analysis [2], [3] have indicated that in the design of stable feedback controls of the form (11) robustness, i.e., insensitivity of the response to parameter variation, for many uncertain systems can be expected if the feedback control matrix is chosen with $P=P_R$, where P_R is the solution of the Riccati equation (4) and $\rho(x)$ is a norm bound of the "lumped" uncertainty. However, guaranteed asymptotic stability or ultimate boundedness of the resulting closed-loop system requires that the positive definite Lyapunov function

$$V(x)=x^TP_Rx$$

be decreasing along trajectories of the system. This can be assured if the resulting Q in Lyapunov's equation (7) is positive definite for $P=P_R$.

Example 3.1. Bryson and Ho [15, pp. 168-170] present the example of a roll attitude regulator for a missile. A feedback controller, as in (3), is designed for a missile using hydraulic-powered ailerons that will keep the roll attitude ϕ close to zero, while staying within the physical limits of aileron deflection δ and aileron deflection rate $\dot{\delta}$. This third order linear system is described by

$$A=\begin{pmatrix}0 & 0 & 0\\ \alpha/\tau & -1/\tau & 0\\ 0 & 0 & 0\end{pmatrix},\qquad B=\begin{pmatrix}1\\ 0\\ 0\end{pmatrix},$$

$$Q_R=\begin{pmatrix}1/\delta_0^2 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 1/\phi_0^2\end{pmatrix},\qquad R=\frac{1}{u_0^2}.$$

Here τ is the roll-time constant, α the aileron effectiveness, δ_0 and ϕ_0 are the maximum

available values of δ and ϕ, respectively, and u_0 is the maximum available value of the command signal to aileron actuators. For a missile system which includes uncertainties, such as noise, the control (3) is modified by the nonlinear feedback control (11) as in [2], [3]. Since

$$P_R B = \begin{pmatrix} \sigma/u_0 \\ \tau(\sigma^2 - 1/\delta_0^2)/2\alpha \\ 1/u_0\phi_0 \end{pmatrix},$$

the matrix $(Q_R, P_R B)$ has rank 3. Theorem 2.1 implies that Q in (9) is positive definite. The results of Breinl and Leitmann [3, § 5] show that by choosing $P = P_R$ in (11) we obtain a feedback control which guarantees practical stability for uncertainties and which considerably reduces the sensitivity of the system.

Example 3.2. The regulator problem for single-input-single-output systems (see [9, § 9-8] has

$$B = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \qquad Q_R = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

For $n \geqq 3$ we have rank $(Q_R, PB) \leqq 2$ for any matrix P. Theorem 2.1 shows that $P = P_R$ will not produce a positive definite Q in (9) and therefore cannot be used in (11) to guarantee stability when uncertainties are present in the system. On the other hand, the results of Theorems 2.3 and 2.4 can be used to exert preferences in the design of the control matrix P. Since the system is single-input, the feedback control (11) becomes

$$p(x) = \begin{cases} \dfrac{-\rho(x)}{\left\| \sum_{i=1}^{n} P_{ni} x_i \right\|} \sum_{i=1}^{n} P_{ni} x_i & \text{for } \left\| \sum_{i=1}^{n} P_{ni} x_i \right\| > \varepsilon, \\ \dfrac{-\rho(x)}{\varepsilon} \sum_{i=1}^{n} P_{ni} x_i & \text{for } \left\| \sum_{i=1}^{n} P_{ni} x_i \right\| \leqq \varepsilon, \end{cases}$$

where $(P_{n1}, P_{n2}, \cdots, P_{nn})$ is the nth row of P and $P_{n1}, P_{n2}, \cdots, P_{n,n-1}$ can be chosen arbitrarily.

4. Insensitivity via switching surface. The nonlinear portions of min-max controllers [4],

$$p(x) = \begin{cases} \dfrac{-B^T P x}{\| B^T P x \|} \bar{\rho}(x) & \text{for } \| B^T P x \| \neq 0, \\ \{u \in E^m : \| u \| \leqq \bar{\rho}\} & \text{for } \| B^T P x \| = 0, \end{cases}$$

are designed to "return" the trajectories of uncertain linear systems to the switching surface

$$N(B^T P) = \{x : B^T P x = 0\}.$$

The stability properties of $\bar{A}$ then guarantee the necessary asymptotic stability of the basic (nominal) linear control system on $N(B^T P)$. Recent results by Gutman and Palmor [4] for min-max controllers show that states in the subset

$$\Gamma_\Delta = \{x : \| B^T P \bar{A} x \| \leqq \Delta\},$$

where Δ is an appropriate constant, possess natural asymptotic attractivity properties to the surface $N(B^TP)$.

Since P and $\bar{A}$ are nonsingular, the subspaces $N(B^TP)$ and

$$\Gamma_0 = N(B^TP\bar{A}).$$

have the same dimension but are, in general, rotations of each other. To achieve maximum effectiveness of the attractivity of Γ_Δ, we need P to be selected so that $\Gamma_0 = N(B^TP)$ (in addition to the properties examined in § 2). We now address this problem.

We begin with a result on positive definite matrices. For any subspace V of E^n, Euclidean n-space, we let the orthogonal complement of V be denoted by

$$V^\perp = \{x: v^Tx = 0 \text{ for all } v \in V\}.$$

LEMMA 4.1. *Let V and W be subspaces of E^n of the same dimension. There exists a positive definite matrix $\hat{P}$ such that $\hat{P}$ maps V onto W if and only if $W \cap V^\perp = \{0\}$. In this case $\hat{P}$ can be selected to be symmetric.*

Proof. Suppose $\hat{P}$ maps V onto W and is positive definite. Let $x \in W \cap V^\perp$. Then $x \in W$ and so there exists $v \in V$ such that $x = \hat{P}v$. But $\hat{P}v = x \in V^\perp$ gives $v^T\hat{P}v = v^Tx = 0$, and $\hat{P}$ positive definite implies $v = 0$. Then $x = \hat{P}v = \hat{P}0 = 0$ and $W \cap V^\perp = \{0\}$.

Conversely, suppose $W \cap V^\perp = \{0\}$. Denote the orthogonal projection map onto the subspace V by P_V. Note that for any subspace V, P_V is symmetric positive semidefinite. Let $\hat{P} = P_W + P_{V^\perp}$. Clearly $\hat{P}$ is (symmetric) positive semidefinite and if $v \in V$, then $\hat{P}v = P_Wv \in W$, so $\hat{P}V \subseteq W$, and it suffices to prove that $\hat{P}x = 0$ implies $x = 0$. Let $x = w + w^\perp$, where $w \in W$ and $w^\perp \in W^\perp$. Then $\hat{P}x = P_Ww + P_{V^\perp}(w + w^\perp)$. Now since $E^n = W \oplus V^\perp$, $P_Ww = 0$ and $P_{V^\perp}(w + w^\perp) = 0$. But $P_Ww = w$ so $w = 0$, and $P_{V^\perp}w^\perp = 0$. Thus $w^\perp \in (V^\perp)^\perp = V$ and since $w^\perp \in W^\perp$, $w^\perp \in V \cap W^\perp = (V^\perp \oplus W)^\perp = (E^n)^\perp = 0$. So $x = w + w^\perp = 0 + 0 = 0$, and $\hat{P}$ is positive definite. □

THEOREM 4.1. *There exists a positive definite P such that*

$$N(B^TP) = N(B^TP\bar{A}) \tag{15}$$

if and only if there is a subspace W of dimension (n-rank B), which is invariant under $\bar{A}$ and satisfies

$$R(B) \cap W = \{0\}, \tag{16}$$

where $R(B) = N(B^T)^\perp$ is the range of B, i.e., the column space of B.

Proof. Since $\bar{A}$ is nonsingular, (15) is satisfied if and only if

$$\bar{A}[N(B^TP)] = N(B^TP). \tag{17}$$

Since $N(B^TP) = P^{-1}[N(B^T)]$, (17) holds if and only if

$$\bar{A}P^{-1}[N(B^T)] = P^{-1}[N(B^T)].$$

So, equation (15) holds if and only if the space $P^{-1}[N(B^T)]$ is invariant under $\bar{A}$. Let $\alpha = n$-rank B. Then the dimension of $N(B^T)$ is α. Let W be a subspace of dimension α which is invariant under $\bar{A}$. By Lemma 4.1 there exists a positive definite matrix P^{-1} which maps $N(B^T)$ onto W if and only if $W \cap N(B^T)^\perp = \{0\}$. P^{-1} positive definite is equivalent to P being positive definite. □

Example 4.1. Suppose the rank of B is 1 (e.g., if system (1) has a scalar control), and suppose that each eigenvalue of $\bar{A}$ has a nonzero complex part. (All trajectories of the stable system $\dot{x} = \bar{A}x$ oscillate.) Then n is even and every (real) $\bar{A}$-invariant subspace has even dimension. Since the dimension of $N(B^TP)$ is $n-1$ for any

nonsingular P, it cannot be invariant under A. Hence there exists no positive definite P satisfying equation (15).

Example 4.2. Suppose the matrix $\bar{A}$ has eigenvectors $v_1, \cdots, v_n$ corresponding to distinct eigenvalues. For such a matrix the set of all $\bar{A}$-invariant subspaces can be enumerated using subsets of these eigenvectors as bases. Let B have rank k. Equation (16) shows that a given positive definite P satisfies (15) if and only if $N(B^TP)$ is $\bar{A}$-invariant; i.e., if and only if $B^TPv_i = 0$ for $n-k$ eigenvectors v_i of $\bar{A}$. This is a criterion which can be used with the matrices of § 2. Note that this result is sufficient for P to satisfy (15) even if $\bar{A}$ does not have distinct eigenvalues.

Suppose the rank of B is $k=1$, so the dimension of $N(B^TP)$ must be $n-1$. If $Bu = v_i$ has a solution for some eigenvector v_i of $\bar{A}$, then the positive definite matrix P satisfying (15) is defined as above for the subspace W generated by the eigenvectors $\{v_j: 1 \leqq j \leqq n, j \neq i\}$. In general, if we select $x \in R(B)$, then

$$x = \sum_{j=1}^{n} \alpha_j v_j,$$

for a unique set of scalars $\alpha_1, \cdots, \alpha_n$. Select v_i such that $\alpha_i \neq 0$ and define

$$W = \langle v_j: 1 \leqq j \leqq n, j \neq i \rangle,$$

the linear subspace spanned by $\{v_j: 1 \leqq j \leqq n, j \neq i\}$. Then W is an $\bar{A}$ invariant subspace of dimension $n-1$ which satisfies (16). The corresponding sum of projections defined above satisfies (15). There is a generalization of this result when $1 < k < n$.

REFERENCES

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

[2] G. LEITMANN, *On the efficacy of nonlinear control in uncertain linear system*, J. Dynamic Systems, Measurement and Control, 102 (1981), pp. 95-102.

[3] W. BREINL AND G. LEITMANN, *Zustandrückführung für dynamische Systeme mit Parameterunsicherheiten*, Regelungstechnik, 31 Heft 3, 1983.

[4] S. GUTMAN AND Z. PALMOR, *Properties of min-max controllers in uncertain dynamical systems*, this Journal, 20 (1982), pp. 85-861.

[5] J. SNYDERS AND M. ZAKAI, *On nonnegative solutions of the equation* $AD + BA' = C^*$, SIAM J. Appl. Math., 18 (1970), pp. 704-714.

[6] S. AVRAHAM AND R. LOEWRY, *On the inertia of the Lyapunov transform* $AH + HA^*$ *for* $H > 0$, Linear and Multilinear Algebra, 6 (1978), pp. 1-21.

[7] D. CARLSON AND R. HILL, *Generalized controllability and inertia theory*, Linear Algebra Appl., 5 (1976), pp. 177-187.

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

[9] M. ATHANS AND P. L. FALB, *Optimal Control*, McGraw-Hill, New York, 1966.

[10] G. STRANG, *Linear Algebra and its Applications*, Academic Press, New York, 1976.

[11] B. N. DATTA, *Stability and D-stability*, Linear Algebra and Appl., 21 (1978), pp. 135-141.

[12] G. P. BARKER, A. BERMAN AND R. J. PLEMMONS, *Positive diagonal solutions to the Lyapunov equations*, Linear and Multilinear Algebra, 5 (1978), pp. 249-256.

[13] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators Part* I: *General Theory*, Interscience, New York, 1966.

[14] W. BREINL, *Entwurf eines unempfindlichen Tragregelsystems fur ein Magnetschwebefahrzeug*, Fortschr. VDI Z., 8, 34, 1980.

[15] A. E. BRYSON, JR. AND Y. C. HO, *Applied Optimal Control*, Hemisphere Publishing, New York, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

003

STOCHASTIC CONTROL OF ONE-DIMENSIONAL DIFFUSIONS WHOSE GENERATORS HAVE DISCONTINUOUS COEFFICIENTS*

HIDEO NAGAI†

Abstract. Stochastic control of one-dimensional singular diffusion processes on an open interval (α, β) is studied. Coefficients $m_i(x)$, $s_i(x)$ of those generators $(1/m_i(x))(d/dx)(1/s_i(x))(d/dx)$ are so singular that controlled processes cannot be defined by solutions of stochastic differential equations. We take up Markovian switching controlled processes and characterize the pay-off function, which is defined by additive functionals of the diffusion processes corresponding to given measures μ^i, by the unique solution of

$$\mathscr{E}^i(u, v-u) \geqq \langle \mu^i, v-u \rangle \quad \forall v \in \mathscr{S}, \quad i=1,2,3,\cdots,$$

$$u \in \mathscr{S}.$$

Here

$$\mathscr{E}^i(u, v) = \int_\alpha^\beta \frac{du}{dx}\frac{dv}{dx}\frac{1}{s_i(x)}\,dx,$$

$$\mathscr{S} = \{v \in H_0^1(\alpha, \beta);\ \mathscr{E}^i(u, \phi) \leqq \langle \mu^i, \phi \rangle \forall \phi \geqq 0,\ \in H_0^1(\alpha, \beta)\}.$$

Key words. stochastic control, singular diffusions, system of quasi-variational inequalities, pay-off functions, switching controlled processes

Introduction. The problem of Bellman-Dirichlet

$$\begin{aligned} &\sup\{\mathscr{G}^i u - f^i\} = 0 \quad \text{a.s. in } D, \\ &u = 0 \quad \text{on } \partial D \end{aligned} \tag{0.1}$$

has been studied by L. C. Evans and A. Friedman [2] and P. L. Lions [6], where $\mathscr{G}^i$, $i=1,2,3,\cdots$, are uniformly elliptic operators with regular coefficients, f^i, $i=1,2,3,\cdots$, are given functions with some regularities and D is an open subset of R^N. On the other hand P. L. Lions and J. L. Menaldi [7] considered the "weak solution" u^* of the equation (0.1) to characterize the pay-off function in stochastic control of stochastic integrals, where diffusion and drift coefficients are merely Lipschitz continuous. "Weak solution" u^* of (0.1) means that u^* is the maximum element of

$$\mathscr{S} = \{v \in H_0^1(D);\ \mathscr{G}^i v \leqq f^i \text{ in } \mathscr{D}' \text{ for all } i\}. \tag{0.2}$$

Now we consider in this article stochastic control of one-dimensional diffusions whose generators have discontinuous coefficients as a generalization of the work of P. L. Lions and J. L. Menaldi. More precisely we treat with the case that $D=(\alpha, \beta)$, $-\infty < \alpha, \beta < \infty$, f^i, $i=1,2,3,\cdots$, are bounded signed measures and generators $\mathscr{G}^i$ of diffusions are the following:

$$\mathscr{G}^i = \frac{1}{m_i(x)}\frac{d}{dx}\left(\frac{1}{s_i(x)}\frac{d}{dx}\right) \tag{0.3}$$

where $m_i(x)$, $s_i(x)$, $i=1,2,3,\cdots$, are uniformly positive bounded measurable functions and $s_i(x)$, $i=1,2,3,\cdots$, are of bounded variation.

Our first problem is to prove the existence of the maximum element u^* of $\mathscr{S}$ (cf. Theorem 1.1). Next we prove in Theorem 3.1 that the pay-off function of stochastic

* Received by the editors September 2, 1983, and in revised form January 1, 1984.

† Department of Mathematics, Fakulty of Science, Tokyo Metropolitan University, Fukasawa, Setagaya-ku, Tokyo, Japan.

control of absorbing diffusion processes with generators $\mathscr{G}^i$ of the formula (0.3) equals to u^*.

We note in this case controlled processes cannot be defined as solutions of S.D.E. because the diffusion and drift coefficients are very singular. Instead we take up Markovian switching controlled processes which are used by J. Zabczyk [11] and M. Nisio [10].

1. An analytical result. Let $I=(\alpha, \beta)$ be a finite interval. We consider for each $c \geqq 1$ the class $\mathbf{G}_c$ of second order differential operators each of which has the following form

$$\mathscr{G}=\frac{1}{m(x)}\frac{d}{dx}\left(\frac{1}{s(x)}\frac{d}{dx}\right) \tag{1.1}$$

where $m(x)$ is a measurable function on I satisfying $1/c \leqq m(x) \leqq c$ and $s(x)$ is a function on I of bounded variation which also satisfies $1/c \leqq s(x) \leqq c$. We put

$$\mathbf{G}=\bigcup_{c \geqq 1} \mathbf{G}_c. \tag{1.2}$$

Let us assume that a countable subset

$$\{\mathscr{G}^i\}_{i=1}^{\infty}=\left\{\frac{1}{m_i(x)}\frac{d}{dx}\left(\frac{1}{s_i(x)}\frac{d}{dx}\right)\right\}_{i=1}^{\infty}$$

of $\mathbf{G}$ and a sequence $\{\mu^i\}_{i=1}^{\infty}$ of finite Borel measures which satisfy the following conditions are given:

(1.3) There exists $c_0 \geqq 1$ such that each $\mathscr{G}^i \in \mathbf{G}_{c_0}$.

(1.4) There exists a finite Borel measure μ on I such that $|\mu^i| \leqq \mu$ for each i, where $|\mu^i|$ indicates the total variation measure of μ^i.

For each $\mathscr{G}^i$ we consider the Dirichlet space $(H_0^1(\alpha, \beta), \mathscr{E}^i)$ on $L^2(m_i(x)\,dx)$ defined by

$$\mathscr{E}^i(u, v)=\int_{\alpha}^{\beta}\frac{1}{s_i(x)}\frac{du}{dx}\frac{dv}{dx}\,dx, \qquad u, v \in H_0^1(\alpha, \beta). \tag{1.5}$$

Here $H_0^1(\alpha, \beta)=\{v \in H^1(\alpha, \beta);\ v(\alpha)=v(\beta)=0\}$, $H^1(\alpha, \beta)=\{v \in L^2(\alpha, \beta);\ v$ is absolutely continuous on I and $dv/dx \in L^2(\alpha, \beta)\}$. Let us put

$$\mathscr{S}=\{v \in H_0^1(\alpha, \beta);\ \mathscr{E}^i(v, \phi) \leqq \langle \mu^i, \phi\rangle, \forall \phi \geqq 0, \in H_0^1(\alpha, \beta) \text{ for each } i\}. \tag{1.6}$$

We at first consider the existence of the maximum element of $\mathscr{S}$. For we will formulate in § 3 stochastic control of diffusions with generators $\{\mathscr{G}^i\}$, where the pay-off function is defined by additive functionals corresponding to given measures $\{\mu^i\}$, and then characterize it by the maximum element of the class $\mathscr{S}$.

DEFINITION 1. $\mathscr{G}^1 < \mathscr{G}^2$ for $\mathscr{G}^1, \mathscr{G}^2 \in \mathbf{G}$ means that $ds_1/s_1(x-) \leqq ds_2/s_2(x-)$, where ds_i means a Borel measure on (α, β) defined by a function $s_i(x)$ of bounded variation.

THEOREM 1. *If there exist $\underline{\mathscr{G}}$ and $\bar{\mathscr{G}}$ belonging to $\mathbf{G}$ such that $\underline{\mathscr{G}} < \mathscr{G}^i < \bar{\mathscr{G}}$ for all i, then $\mathscr{S}$ has a maximum element.*

2. Proof of Theorem 1. Let $G^i(x, y)$ be a Green function of $(H_0^1(\alpha, \beta), \mathscr{E}^i)$ for each i defined by the equality $\mathscr{E}^i(G^i(\cdot, y), v)=v(y)$, $v \in H_0^1$ and G^i be the Green operator defined by the following:

$$G^i\nu(x)=\int G^i(x, y)m_i(y)\nu(dy)$$

where ν is any Borel measure of bounded variation.

LEMMA 2.1. *Let* $\mathcal{G}^1, \mathcal{G}^2 \in \mathbf{G}$, *then* $\mathcal{G}^1 < \mathcal{G}^2$ *if and only if* s_1/s_2 *is nonincreasing.*

LEMMA 2.2. *Under the assumptions of Theorem* 1 *there exist nonpositive functions* w_1 *and* w_2 *belonging to* $H^1(\alpha, \beta)$ *such that* $w_1(\alpha) = 0$, $w_2(\beta) = 0$ *and* $\mathcal{E}^i(G^i\mu^i - w_j, v) \geqq 0$ $\forall v \geqq 0, \in H_0^1$ *for all* i *and* $j = 1, 2$.

PROPOSITION 2.1. *Under the assumptions of Theorem* 1 *the following system of quasi-variational inequalities has a maximum solution for each* n *and* ε:

$$\text{(SQV.1)} \qquad \begin{aligned} &\mathcal{E}^1(u^1, v-u^1) \geqq \langle \mu^1, v-u^1\rangle \quad \forall v \leqq u^2+\varepsilon,\ u^1 \leqq u^2+\varepsilon,\\ &\mathcal{E}^2(u^2, v-u^2) \geqq \langle \mu^2, v-u^2\rangle \quad \forall v \leqq u^3+\varepsilon,\ u^2 \leqq u^3+\varepsilon,\\ &\qquad\vdots \qquad\qquad\qquad\qquad\qquad \vdots\\ &\mathcal{E}^n(u^n, v-u^n) \geqq \langle \mu^n, v-u^n\rangle \quad \forall v \leqq u^1+\varepsilon,\ u^n \leqq u^1+\varepsilon. \end{aligned}$$

Proof of Proposition 2.1. Put $u_0^1 = G^1\mu^1$; then the following variational inequality has the unique solution u_1^n:

$$\mathcal{E}^n(u_1^n, v-u_1^n) \geqq \langle \mu^n, v-u_1^n\rangle \quad \forall v \leqq u_0^1+\varepsilon,\ u_1^n \leqq u_0^1+\varepsilon. \tag{2.1}$$

Therefore we can take inductively the solutions u_1^i of the variational inequalities:

$$\mathcal{E}^i(u_1^i, v-u_1^i) \geqq \langle \mu^i, v-u_1^i\rangle \quad \forall v \leqq u_1^{i+1}+\varepsilon,\ u_1^i \leqq u_1^{i+1}+\varepsilon, \tag{2.2}$$

$i = 1, 2, \cdots, n-1$. Then, replacing u_0^1 or u_1^{i+1} by u_k^1 or u_{k+1}^{i+1} respectively in (2.1) or (2.2), we can take the solutions u_{k+1}^n, u_{k+1}^i, $i = 1, 2, \cdots, n-1$, $k = 1, 2, \cdots$, of variational inequalities (2.1) or (2.2). For each i and k solutions u_k^i have the following properties:

$$u_{k+1}^i \leqq u_k^i, \tag{2.3}$$

$$\mathcal{E}^i(G^i\mu^i - u_k^i, v) \geqq 0 \quad \forall v \geqq 0, \in H_0^1(\alpha, \beta), \tag{2.4}$$

$$u_k^i \geqq w_j, \qquad j = 1, 2, \tag{2.5}$$

where w_1 and w_2 are the functions appearing in Lemma 2.1. (2.4) is easily seen, for u_k^i satisfies that

$$\mathcal{E}^i(G^i\mu^i - u_k^i, \bar{v} - u_k^i) \leqq 0 \quad \forall \bar{v} \leqq u_k^{i+1}+\varepsilon; \tag{2.6}$$

hence if we put in (2.6) $\bar{v} = u_k^i - v$, $v \geqq 0$, $\in H_0^1$ we have (2.4). In order to see (2.3) we at first remark that $u_0^1 - u_1^1 = G^1\mu^1 - u_1^1 \geqq 0$ because u_1^1 satisfies (2.4) (cf. [3]). By standard argument it follows that $u_2^n \leqq u_1^n$ from $u_1^1 + \varepsilon \geqq u_0^1 + \varepsilon$ (cf. [1], [8]). Therefore by induction it follows that $u_{k+1}^i \leqq u_k^i$, $i = 1, \cdots, n-1$ or $u_{k+1}^n \leqq u_k^n$ from $u_{k+1}^{i+1} \leqq u_k^{i+1}$, $i = 1, \cdots, n-1$ or $u_{k+1}^1 \leqq u_k^1$ respectively in the same way as above for each k. (2.5) is due to Lemma 2.3 below. The conclusion of the present proposition will follow from (2.3), (2.4) and (2.5) as follows.

From (2.3) and (2.4) it follows that

$$\mathcal{E}^i(G^i\mu^i - u_k^i, G^i\mu^i - u_k^i) \leqq \mathcal{E}^i(G^i\mu^i - u_{k+1}^i, G^i\mu^i - u_{k+1}^i) \tag{2.7}$$

for each i and k. Let us put $w = w_1 \vee w_2$; then we have $w \in H_0^1$ and $w \leqq u_k^i$. Therefore it follows from (2.4) that

$$\mathcal{E}^i(G^i\mu^i - u_k^i, u_k^i - w) \geqq 0 \tag{2.8}$$

for all i and k. (2.8) means that

$$\mathcal{E}^i(G^i\mu^i - u_k^i, G^i\mu^i - u_k^i) \leqq \mathcal{E}^i(G^i\mu^i - u_k^i, G^i\mu^i - w).$$

After applying the Schwarz inequality we have

$$\mathscr{E}^i(G^i\mu^i - u_k^i,\ G^i\mu^i - u_k^i) \leqq \mathscr{E}^i(G^i\mu^i - w,\ G^i\mu^i - w). \tag{2.9}$$

It follows from (2.4), (2.7) and (2.9) that

$$\mathscr{E}^i(u_k^i - u_j^i,\ u_k^i - u_j^i) \to 0 \quad \text{as } k, j \to \infty$$

for each i. H_0^1 is complete with respect to $\mathscr{E}^i$-norm, so there exists $u^i \in H_0^1$ such that

$$\mathscr{E}^i(u_k^i - u^i,\ u_k^i - u^i) \to 0 \quad \text{as } k \to \infty$$

for each i. It can be seen in a similar way to [8] that u^i, $i = 1, 2, \cdots, n$ is the maximum solution of (SQV. 1).

LEMMA 2.3. *We have* $u_k^i \geqq w_j$, $i = 1, \cdots, n$, $k = 1, 2, \cdots, j = 1, 2$.

Proof of Lemma 2.3. It is known that a Markov process $\mathbf{X}^i = (\Omega, \mathscr{B}, P_x^i, X_t)$ corresponds to the Dirichlet space $(H_0^1, \mathscr{E}^i)$ on $L^2(m_i(x)\,dx)$ and continuous additive functional A_t^i of bounded variation of $\mathbf{X}^i$ to given measure μ^i for each i (cf. [3], [5]) such that $\lim_{t \downarrow 0} 1/t \int (v(x) - E_x^i[v(X_t)])v(x)m_i(x)\,dx = \mathscr{E}^i(v, v)$, $v \in H_0^1$, and $E_x^i[A_t^i] = E_x^i[G^i\mu^i(X_t)] - G^i\mu^i(x)$, $t > 0$. Let us consider the following optimal stopping problems:

$$\bar{u}_k^n(x) = \inf_\tau E_x[A_\tau^n + u_{k-1}^1(X_\tau) + \varepsilon],$$

$$\bar{u}_k^i(x) = \inf_\tau E_x^i[A_\tau^i + u_k^{i+1}(X_\tau) + \varepsilon]$$

$i = 1, \cdots, n$, $k = 1, 2, \cdots$. Then we know that $\bar{u}_k^i(x) = u_k^i(x)$ for each i and k (cf. [8], [2]) because there exists no exceptional set in the present case. Moreover we know that an optimal stopping time τ_k^i exists for each i and k. We now prove that $u_1^n(x) \geqq w_j(x)$, $j = 1, 2$. We at first note that it follows that $G^1\mu^1(x) \geqq w_j(x)$, $j = 1, 2$, $x \in I$ from the maximum principle because it follows that $G^1\mu^1(\alpha) - w_j(\alpha) \geqq 0$, $G^1\mu^1(\beta) - w_j(\beta) \geqq 0$, $j = 1, 2$ from $w_j(\alpha) \leqq 0$, $w_j(\beta) \leqq 0$, $j = 1, 2$ and $G^1\mu^1(\alpha) = G^1\mu^1(\beta) = 0$ and w_j satisfies $\mathscr{E}^1(G^1\mu^1 - w_j, v) \geqq 0$, $\forall v \geqq 0, \in H_0^1$, $j = 1, 2$. Therefore we have

$$\begin{aligned} u_1^n(x) &= E_x^n[A_{\tau_1^n}^n + u_0^1(X_{\tau_1^n}) + \varepsilon] \\ &\geqq E_x^n[A_{\tau_1^n}^n + w_j(X_{\tau_1^n})] \\ &= E_x^n[G^n\mu^n(x) - w_j(x) - \{G^n\mu^n(X_{\tau_1^n}) - w_j(X_{\tau_1^n})\} + w_j(x)] \\ &\geqq w_j(x) \end{aligned}$$

for $j = 1, 2$. It can be seen that $u_k^i \geqq w_j$ for all i, k and j by induction.

Proof of Lemma 2.1. Let $\mathscr{G}^1$, $\mathscr{G}^2 \in \mathbf{G}$ and s_1 and s_2 correspond to them respectively, then we have

$$\begin{aligned} \frac{s_1}{s_2}(y) - \frac{s_1}{s_2}(x) &= \frac{s_1(y)s_2(x) - s_1(x)s_2(y)}{s_2(y)s_2(x)} \\ &= \frac{s_1(x)}{s_2(y)} \left\{ \frac{s_1(y) - s_1(x)}{s_1(x)} - \frac{s_2(y) - s_2(x)}{s_2(x)} \right\}, \qquad y > x, \end{aligned}$$

from which Lemma 2.1 follows.

Proof of Lemma 2.2. We at first assume that $\bar{\mathscr{G}}$, $\underline{\mathscr{G}} \in \mathbf{G}_c$ for some $c \geqq 1$. Put $B(x) = \mu((\alpha, x])$, $B_1(x) = cc_0(B(x) - B(\beta -))$ and $B_2(x) = cc_0B(x)$. Let us define w_1

and w_2 by the following:

$$w_1(x) = \int_\alpha^x B_1(x)\underline{s}(x)\,dx,$$

$$w_2(x) = \int_\alpha^x B_2(x)\bar{s}(x)\,dx - \int_\alpha^\beta B_2(x)\bar{s}(x)\,dx$$

where $\underline{s}$ or $\bar{s}$ corresponds to $\underline{\mathcal{G}}$ or $\bar{\mathcal{G}}$ by

$$\underline{\mathcal{G}} = \frac{1}{\underline{m}(x)}\frac{d}{dx}\left(\frac{1}{\underline{s}(x)}\frac{d}{dx}\right) \quad \text{or} \quad \bar{\mathcal{G}} = \frac{1}{\bar{m}(x)}\frac{d}{dx}\left(\frac{1}{\bar{s}(x)}\frac{d}{dx}\right)$$

respectively. Then we have

$$\begin{aligned} -d\left(\frac{1}{s_i(x)}\frac{d}{dx}(G^i\mu^i - w_1)\right) &= d\mu^i + d\left(\frac{1}{s_i(x)}\frac{d}{dx}w_1\right) \\ &= d\mu^i + d\left(\frac{B_1(x)\underline{s}(x)}{s_i(x)}\right) \\ &= d\mu^i + B_1(x-)d\left(\frac{\underline{s}}{s_i}\right) + \frac{\underline{s}(x+)}{s_i(x+)}\,dB_1. \end{aligned}$$

As $\underline{s}/s_i$ is nonincreasing and $B_1(x-) \leqq 0$, we obtain $B_1(x-)d(\underline{s}/s_i) \geqq 0$. On the other hand we have

$$d|\mu^i| \leqq \frac{1}{cc_0}\,dB_1 \leqq \frac{\underline{s}(x^+)}{s_i(x^+)}\,dB_1.$$

Hence $-d((1/s_i(x))(d/dx)(G^i\mu^i - w_1))$ is a positive measure for each i. As for $-d(1/s_i(x))(d/dx)(G^i\mu^i - w_2))$ we also have the same conclusion by a similar argument making use of the fact that $B_2(x-) \geqq 0$, $d|\mu^i| \leqq (\bar{s}(x+)/s_i(x+))\,dB_2$ and $\bar{s}/s_i$ is nondecreasing.

Now we give the proof of Theorem 1.1 which follows from Proposition 2.1.

Proof of Theorem 1.1. Owing to Proposition 2.1 there exists a maximum solution $(u^1, u^2, \cdots, u^n)$ of (SQV. 1) for each n and ε. We can see in the same way as the proof of Proposition 2.1 that all u^i have the same limit u_n in H_0^1 as ε tends to 0 because it holds that

$$\mathcal{E}^i(G^i\mu^i - u^i, G^i\mu^i - u^i) \leqq \mathcal{E}^i(G^i\mu^i - w, G^i\mu^i - w)$$

and $u^i \downarrow$ as $\varepsilon \downarrow 0$. Moreover this u_n is the maximum element of $\mathcal{S}_n$ defined by

$$\mathcal{S}_n = \{v \in H_0^1;\ \mathcal{E}^i(v, \phi) \leqq \langle \mu^i, \phi\rangle, \forall \phi \geqq 0, \in H_0^1, i = 1, 2, \cdots n\}.$$

It can be seen in the same way as above that $\{u_n\}$ is a Cauchy sequence in H_0^1. The fact that the limit u of u_n in H_0^1 is the maximum element of $\mathcal{S}$ also follows.

The proof of Theorem 1.1 implies that u has another expression as the solution of infinite variational inequalities as follows.

Remark. The above u is the unique solution of the following:

$$\mathcal{E}^i(u, v-u) \geqq \langle \mu^i, v-u\rangle \quad \forall v \in \mathcal{S}, u \in \mathcal{S}, \quad i = 1, 2, \cdots.$$

3. Stochastic control. Let $\mathbf{X}^i = (\Omega^i, \mathcal{B}^i, \mathcal{B}_t^i, P_x^i, X_t)$ be a diffusion process on $I \cup \{\delta\}$ associated with the generator $\mathcal{G}^i$ given in § 1 for each i, where δ is the terminal point

added to an open interval I. We define a class V_m for each $m=1, 2, 3, \cdots$, by

$$V_m=\{v=\{v_{m,k}(x)\}_{k=0}^{\infty};\ v_{m,k}(x) \text{ is a integer-valued Borel function on } I\cup\{\delta\} \text{ for each } k\}. \tag{3.1}$$

We call each $v\in V_m$ a control. For each v a Markovian switching controlled process $\mathbf{X}^v=(\Omega_\infty, \mathscr{F}, P_x^v, Y_t)$ can be defined as follows. Let $\Omega_\infty=\prod_{i=1}^{\infty}\Omega^i$, $\mathscr{F}=\bigotimes_{i=1}^{\infty}\mathscr{B}^i$ and $\Delta=1/2^m$. We define a stochastic kernel $\pi_n(\omega_n, \cdot)$ by

$$\pi_n(\omega_n, A)=P_{X_\Delta(\omega_n)}^{v_{m,n}(X_\Delta(\omega_n))}(A), \qquad \omega_n\in\Omega^n, \quad A\in\mathscr{B}^{n+1}.$$

Then we can see that by virtue of a theorem of Ionescu-Tulcea there exists a unique system $\{P_x^v, x\in I\cup\{\delta\}\}$ or probability measures on $(\Omega_\infty, \mathscr{F})$ such that

$$P_x^v\left(\prod_{j=1}^{n+1}B_j\right)=\int_{B_1}P_x^{v_{m,0}(x)}(d\omega_1)\int_{B_2}\pi_1(\omega_1, d\omega_2)\int_{B_3}\pi_2(\omega_2, d\omega_3)\cdots\int_{B_{n+1}}\pi_n(\omega_n, d\omega_{n+1}),$$

$B_i\in\mathscr{B}^i, \qquad i=1, 2, \cdots, n+1$

in the same way as the construction of the branching process (cf. [4]). Let us set

$$Y_t(\boldsymbol{\omega})=X_{t-k\Delta}(\omega_{k+1}), \quad k\Delta\leqq t<(k+1)\Delta, \quad k=0, 1, 2, \cdots,$$
$$\boldsymbol{\omega}=(\omega_1, \omega_2, \cdots)\in\Omega_\infty.$$

Then (P_x^v, Y_t) is the controlled process by a control v which we are going to consider. Now we noted in § 2 that for each i to a given measure μ^i there corresponds a continuous additive functional A_t^i of bounded variation of $\mathbf{X}^i$. Connecting these additive functionals A_t^i, $i=1, 2, 3, \cdots$, we define $A_t^v(\boldsymbol{\omega})$ by

$$A_t^v(\boldsymbol{\omega})=\begin{cases} A_t^{v_{m,0}(Y_0)}(\omega_1), & 0\leqq t\leqq\Delta, \\ A_\Delta^v+A_{t-\Delta}^{v_{m,1}(Y_\Delta)}(\omega_2), & 0\leqq t\leqq 2\Delta, \\ \quad\vdots & \\ A_{k\Delta}^v+A_{t-k\Delta}^{v_{m,k}(Y_{k\Delta})}(\omega_{k+1}), & k\Delta\leqq t\leqq(k+1)\Delta, \\ \quad\vdots & \end{cases} \tag{3.2}$$

for $\boldsymbol{\omega}=(\omega_1, \omega_2, \cdots)\in\Omega_\infty$. Then the pay-off function is defined by

$$\bar{u}(x)=\inf_{v\in V_m,\ m\in\mathbf{N}}E_x^v[A_\infty^v], \tag{3.3}$$

where $\mathbf{N}$ is the set of all integers. Our purpose is to prove Theorem 3.1.

THEOREM 3.1. *Under the assumptions of Theorem* 1.1 *we have* $u(x)=\bar{u}(x)$ *where* u *is the maximum element of* $\mathscr{S}$ *whose existence is insured by Theorem* 1.1.

4. Proof of Theorem 3.1. Let u be the maximum element of $\mathscr{S}$. Since $G^i\mu^i-u$ is excessive with respect to $(H_0^1, \mathscr{E}^i, L^2(m_i(x)\,dx))$ there exists positive Radon measure ν^i such that $G^i\mu^i-u=G^i\nu^i$. Let us denote by $\bar{A}_t^i$ the continuous additive functional of bounded variation of $\mathbf{X}^i$ corresponding to ν^i for each i and by $\bar{A}_t^v$ the additive functional of controlled process $\mathbf{X}^v$ defined by the formula (3.2) from $\{\bar{A}_t^i\}_i$. Now the following lemma is useful.

LEMMA 4.1. *We have for all* t *and* $v\in V_m$

$$u(x)+E_x^v[\bar{A}_t^v]=E_x^v[A_t^v+u(Y_t)], \tag{4.1}$$

$$u(x)+E_x^v[\bar{A}_\infty^v]=E_x^v[A_\infty^v], \tag{4.2}$$

$$u(x)\leqq\inf_{v\in V_m,\ m\in\mathbf{N}}E_x^v[A_t^v+u(Y_t)]. \tag{4.3}$$

Proof of Lemma 4.1. We will write for brevity $E_{\cdot}^{m,k}$ or $G^{m,k}\mu^{m,k}(\cdot)$ or $A_{\cdot}^{m,k}$ in place of $E_{\cdot}^{v_{m,k}(\cdot)}$ or $G^{v_{m,k}(\cdot)}\mu^{v_{m,k}(\cdot)}(\cdot)$ or $A_{\cdot}^{v_{m,k}(\cdot)}$ respectively. Let $t\in[k\Delta,(k+1)\Delta)$ and $v\in V_m$, then

$$
\begin{aligned}
&E_x^v[A_{t-k\Delta}^{m,k}(\omega_{k+1})]\\
&\quad=E_x^v[E_{Y(k\Delta)}^{m,k}[G^{m,k}\mu^{m,k}(X_0)-G^{m,k}\mu^{m,k}(X_{t-k\Delta})]\\
&\quad=E_x^v[G^{m,k}\mu^{m,k}(Y_{k\Delta})-u(Y_{k\Delta})-\{G^{m,k}\mu^{m,k}(Y_t)-u(Y_t)\}+u(Y_{k\Delta})-u(Y_t)]\\
&\quad=E_x^v[G^{m,k}\nu^{m,k}(Y_{k\Delta})-G^{m,k}\nu^{m,k}(Y_t)+u(Y_{k\Delta})-u(Y_t)]\\
&\quad=E_x^v[\bar{A}_{t-k\Delta}^{m,k}(\omega_{k+1})+u(Y_{k\Delta})-u(Y_t)].
\end{aligned}
$$

Similarly we have for $j=0,1,2,\cdots,k-1$,

$$
E_x^v[A_\Delta^{m,j}(\omega_{j+1})]=E_x^v[\bar{A}_\Delta^{m,j}(\omega_{j+1})+u(Y_{j\Delta})-u(Y_{(j+1)\Delta})].
$$

Then we have

$$
\begin{aligned}
E_x^v[A_t^v]&=E_x^v\left[\sum_{j=0}^{k-1}A_\Delta^{m,j}(\omega_{j+1})+A_{t-k\Delta}^{m,k}(\omega_{k+1})\right]\\
&=E_x^v[\bar{A}_t^v-u(Y_t)+u(Y_0)].
\end{aligned}
$$

(4.1) has been proved.

(4.3) is a direct consequence of (4.1) because $\bar{A}_t$ is a nonnegative additive functional.

For the proof of (4.2) we at first remark that there exists positive constant λ such that $P_x^i[\zeta<\infty]=P_t^i1(x)\leqq e^{-\lambda t}$ for all t, x and i, where P_t^i is a transition semi-group of the process $\mathbf{X}^i$. For the principal eigenvalue of the generator in the sense of L^2 of diffusion process $\mathbf{X}^i$ is uniformly (with respect to i) and strictly positive because of the condition that $1/c_0\leqq s_i(x)$, $m_i(x)\leqq c_0$, which implies the existence of the above λ. Let $A_t^{\mu,i}$ be the continuous additive functional of bounded variation of $\mathbf{X}^i$ corresponding to μ and $A_t^{v,\mu}$ be the additive functional of $\mathbf{X}^v$ defined by (3.2) from $\{A_t^{\mu,i}\}$, then it follows that total variation of A_t^v is dominated by $A_t^{v,\mu}$ from $|\mu^i|\leqq\mu$ for all i. On the other hand we have

$$
\begin{aligned}
E_x^v[A_{\zeta\infty}^{v,\mu}(\boldsymbol{\omega})]&=\sum_{j=0}^{\infty}E_x^v[A_\Delta^{\mu,v_{m,j}(Y_{j\Delta})}(\omega_{j+1})]\\
&=\sum_{j=0}^{\infty}E_x^v[G^{m,j}\mu(Y_{j\Delta})-G^{m,j}\mu(Y_{(j+1)\Delta})]\\
&\leqq M\sum_{j=0}^{\infty}e^{-\lambda j\Delta}\\
&=M\frac{1}{1-e^{-\lambda\Delta}}
\end{aligned}
$$

where M is a positive number. Because $G^i(x,y)$ is bounded uniformly with respect to x, y and i and $\mu((\alpha,\beta))<\infty$. Hence (4.2) follows from (4.1) by virtue of dominated convergence theorem and monotone convergence theorem.

Now let us put

$$
u(x,t)=\inf_{v\in V_m,\,m\in\mathbf{N}}E_x^v[A_t^v+u(Y_t)]. \tag{4.4}
$$

Then we have the following Lemma 4.2.

LEMMA 4.2. *We have for all s, t, x and i*

$$u(x,t) \leqq E_x^i[A_s^i + u(X_s, t)]. \tag{4.5}$$

Proof of Lemma 4.2. Let us put

$$V_{m,l} = \{v \in V_m;\ v_{m,k}(x) \in \{1, 2, \cdots, l\} \text{ for all } x \text{ and } k\};$$

then we have for all x and t

$$\inf_{v \in V_m} E_x^v[A_t^v + u(Y_t)] = \lim_{l \to \infty} \inf_{v \in V_{m,l}} E_x^v[A_t^v + u(Y_t)]. \tag{4.6}$$

Let us assume that $t \in [k\Delta, (k+1)\Delta)$. We prove (4.6) by induction with respect to k. At first for $k=0$, we have

$$E_x^v[A_t^v + u(Y_t)] \geqq \inf_{v \in V_{m,l}} E_x^v[A_t^v + u(Y_t)]$$

if $v_{m,0}(x) \in \{1, 2, \cdots, l\}$. Therefore we obtain

$$E_x^v[A_t^v + u(Y_t)] \geqq \lim_{l \to \infty} \inf_{v \in V_{m,l}} E_x^v[A_t^v + u(Y_t)]$$

for all $v \in V_m$ and x. Hence we have

$$\inf_{v \in V_m} E_x^v[A_t^v + u(Y_t)] \geqq \lim_{l \to \infty} \inf_{v \in V_{m,l}} E_x^v[A_t^v + u(Y_t)].$$

The converse inequality is obvious. In order to prove (4.6) for general k we remark the following: Owing to the measurable selection theorem there exists $v^* \in V_{m,l}$ such that

$$E_x^{v^*}[A_t^{v^*} + u(Y_t)] = \inf_{v \in V_{m,l}} E_x^v[A_t^v + u(Y_t)]$$

(cf. [10]). Let us assume that (4.6) holds for $k-1$. We put $\bar{v} = \{v_{m,j+1}\}_{j=0}^{\infty} \in V_m$ for $v = \{v_{m,j}\}_{j=0}^{\infty} \in V_m$; then we have

$$\begin{aligned} E_x^v[A_t^v + u(Y_t)] &= E_x^v[A_\Delta^v + E_y^{\bar{v}}[A_{t-\Delta}^{\bar{v}} + u(Y_{t-\Delta})]|_{y=Y_\Delta}] \\ &\geqq E_x^v[A_\Delta^v + \inf_{v \in V_m} E_y^v[A_{t-\Delta}^v + u(Y_{t-\Delta})]|_{y=Y_\Delta}] \\ &= E_x^v[A_\Delta^v + \lim_{l \to \infty} \inf_{v \in V_{m,l}} E_y^v[A_{t-\Delta}^v + u(Y_{t-\Delta})]|_{y=Y_\Delta}] \\ &= \lim_{l \to \infty} E_x^v[A_\Delta^v + E_y^{v^*}[A_{t-\Delta}^{v^*} + u(Y_{t-\Delta})]|_{y=Y_\Delta}]. \end{aligned}$$

We furthermore assume that $v_{m,0}(x) \in \{1, 2, \cdots, l\}$ and put $\underline{v} = \{v_{m,0}(x), v_{m,0}^*(x), v_{m,1}(x), \cdots, v_{m,j}(x), \cdots\}$; then we have

$$\begin{aligned} E_x^v[A_\Delta^v + E_y^{v^*}[A_{t-\Delta}^{v^*} + u(Y_t)]|_{y=Y_\Delta}] &= E_x^{\underline{v}}[A_t^{\underline{v}} + u(Y_t)] \\ &\geqq \inf_{v \in V_{m,l}} E_x^v[A_t^v + u(Y_t)]. \end{aligned}$$

Therefore we obtain for $v \in V_m$

$$E_x^v[A_t^v + u(Y_t)] \geqq \lim_{l \to \infty} \inf_{v \in V_{m,l}} E_x^v[A_t^v + u(Y_t)].$$

Hence we have

$$\inf_{v\in V_m} E_x^v[A_t^v+u(Y_t)]\geqq \lim_{l\to\infty}\inf_{v\in V_{m,l}} E_x^v[A_t^v+u(Y_t)].$$

Since the converse inequality is obvious we arrive at (4.6) for general k. By virtue of (4.6) we have our present lemma as follows. Since $\inf_{v\in V_m} E_{X_s}^v[A_t^v+u(Y_t)]$ is monotone nonincreasing with respect to m and dominated by an integrable function of ω, we have

$$\begin{aligned} E_x^i[A_s^i+u(X_s,t)] &= E_x^i[A_s^i+\lim_{m\to\infty}\inf_{v\in V_m} E_{X_s}^v[A_t^v+u(Y_t)]] \\ &= \lim_{m\to\infty} E_x^i[A_s^i+\lim_{l\to\infty} E_{X_s}^{v^{*l}}[A_t^{v^{*l}}+u(Y_t)]], \\ &= \lim_{m\to\infty}\lim_{l\to\infty} E_x^i[A_s^i+E_{X_s}^{v^{*l}}[A_t^{v^{*l}}+u(Y_t)]] \end{aligned}$$

where v^{*l} is a control which minimizes $E_x^v[A_t^v+u(Y_t)]$ in $V_{m,l}$. In the same way as construction of controlled process in § 3 we can define the following process $(\underline{P}_x^v, \underline{Y}_t)$ for $v=\{v_{m,0}(x), v_{m,0}^{*l}(x), v_{m,1}^{*l}(x),\cdots, v_{m,j}^{*l}(x),\cdots\}$, where $v_{m,0}(x)=i$ for all x. $\underline{Y}_t$ behaves under the law P_x^i from time 0 to time s. After time s $\underline{Y}_t$ moves under the law $P_{Y_{k\Delta}}^{v_{m,k}(Y_{k\Delta})}$ for $t\in[s+k\Delta, s+(k+1)\Delta)$, $k=0,1,2,\cdots$, successively. $\underline{A}_t^v$ can be defined in the same way as (3.2) for given measures $\{\mu^j\}_j$. Then we have

$$E_x^i[\underline{A}_s^i+E_{X_s}^{v^{*l}}[\underline{A}_t^{v^{*l}}+u(Y_t)]]=\underline{E}_x^v[\underline{A}_{t+s}^v+u(\underline{Y}_{t+s})].$$

On the other hand it can be seen that $\underline{E}_x^v[\underline{A}_{t+s}^v+u(\underline{Y}_{t+s})]$ is monotone nondecreasing in s because (4.1) holds also in this situation. Therefore we obtain

$$E_x^i[A_s^i+u(X_s,t)]\geqq E_x^i[A_0^i+u(X_0,t)]=u(x,t).$$

LEMMA 4.3. *We have $u_t(x)=u(x)$ for all t and x.*

Proof of Lemma 4.3. Let us put $u_t=u_t(\cdot)=u(\cdot,t)$; then we have $u_t\geqq u$ by Lemma 4.1 and $u_t-G^i\mu^i\leqq P_s^i(u_t-G^i\mu^i)$ for all i and s. Therefore we obtain

$$\begin{aligned} &\frac{1}{s}(u_t-u-P_s^i(u_t-u),u_t-u) \\ &=\frac{1}{s}(u_t-G^i\mu^i-P_s^i(u_t-G^i\mu^i),u_t-u)-\frac{1}{s}(u-G^i\mu^i-P_s^i(u-G^i\mu^i),u_t-u) \\ &\leqq\frac{1}{s}(G^i\mu^i-u-P_s^i(G^i\mu^i-u),u_t-u) \\ &=\frac{1}{s}(G^i\nu^i-P_s^iG^i\nu^i,u_t-u) \\ &=\frac{1}{s}\int_0^s\langle\nu^i,P_\tau^i(u_t-u)\rangle d\tau, \end{aligned}$$

where $(\cdot,\cdot)$ is a $L^2(m_i(dx))$-inner product. Since the last member converges to $\langle\nu^i,u_t-u\rangle$ as $s\to 0$ we have $\mathscr{E}^i(u_t-u,u_t-u)<\infty$ which means that $u_t\in H_0^1$. Therefore because of Lemma 4.2 we obtain

$$\mathscr{E}^i(G^i\mu^i-u_t,\phi)\geqq 0\quad\forall\phi\geqq 0,\in H_0^1$$

for all i. Hence we conclude that $u_t=u$ for all t.

LEMMA 4.4. *We have* $\inf_{v\in V_m, m\in\mathbf{N}} E_x^v[\bar{A}_\infty^v]=0$.

Proof of Lemma 4.4. By virtue of Lemma 4.1 and Lemma 4.3 we have $\inf_{v\in V_m,m\in\mathbf{N}}E_x^v[\bar{A}_t^v]=0$. For all t. Let $t=k\Delta$ and define $V_{m,t}=\{v\in V_m;\ v_{m,j}, j\geqq k,$ are constant functions$\}$, then the following equality holds:

$$\inf_{v\in V_{m,t},m\in\mathbf{N}} E_x^v[\bar{A}_\infty^v - G^{m,k}\mu^{m,k}(Y_t)]=0.$$

Therefore we have

$$\inf_{v\in V_{m,t},m\in\mathbf{N}} E_x^v[\bar{A}_\infty^v]\leqq \inf_{v\in V_{m,t},m\in\mathbf{N}} E_x^v[-G^{m,k}\mu^{m,k}(Y_t)]\leqq M e^{-\lambda t}.$$

Hence we obtain that $\inf_{v\in V_m,m\in\mathbf{N}}E_x^v[\bar{A}_\infty^v]\leqq \lim_{t\to\infty} M e^{-\lambda t}=0$.

Proof of Theorem 3.1. This is immediate from Lemma 4.1 and Lemma 4.4.

5. Examples.

Example 5.1. Let $I=(-1,1)$,

$$s_i(x)=\exp\left(-\int_0^x \frac{b_i(y)}{a_i(y)}\,dy\right),$$

$$m_i(x)=\frac{1}{a_i(x)}\exp\left(\int_0^x \frac{b_i(y)}{a_i(y)}\,dy\right),$$

where $a_i(x)$ and $b_i(x)$ are bounded measurable functions such that $0<c_1\leqq a_i(x)\leqq c_2<\infty$ and $|b_i(x)|\leqq M<\infty$ for all i. Then it is obvious that s_i and m_i satisfy conditions in Theorem 1.1.

$$\mathcal{G}^i=\frac{1}{m_i(x)}\frac{d}{dx}\left(\frac{1}{s_i(x)}\frac{d}{dx}\right)$$

can be written as

$$\mathcal{G}^i=a_i(x)\left(\frac{d}{dx}\right)^2+b_i(x)\frac{d}{dx}.$$

Example 5.2. Let $I=(-1,1)$,

$$s_i(x)=\begin{cases}\exp\left(\displaystyle\int_0^x k_i(y)(1-y)^{-1+\alpha}dy\right), & 0<x<1,\\ \exp\left(-\displaystyle\int_0^x l_i(y)(1+y)^{-1+\beta}dy\right), & -1<x\leqq 0\end{cases}$$

and $m_i(x)=\{s_i(x)\}^{-1}$, where k_i and l_i are measurable functions such that $0\leqq k_i,\ l_i\leqq M<\infty$ for all i. It is easy to see that the above s_i and m_i satisfy conditions in Theorem 1.1. If we rewrite as $\mathcal{G}^i=(d/dx)^2+b_i(x)\,d/dx$,

$$b_i(x)=\begin{cases}-k_i(x)(1-x)^{-1+\alpha}, & 0<x<1,\\ l_i(x)(1+x)^{-1+\beta}, & -1<x\leqq 0,\end{cases}$$

then $b_i(x)$ has singularities at -1 and 1.

The following example does not satisfy the condition that there exists $\underline{\mathcal{G}}$ and $\bar{\mathcal{G}}\in\mathbf{G}$ such that $\underline{\mathcal{G}}<\mathcal{G}^i<\bar{\mathcal{G}}$.

Example 5.3. Let $I=(-1, 1)$,

$$s_i(x)=\begin{cases}\exp\left(1-k(1-x)^{k^{-1}}\right), & 1-k^{-k}\leqq x<1,\\ \exp\left(1-k(1+k)^{k^{-1}}\right), & -1<x\leqq -1+k^{-k},\\ 1, & \text{otherwise,}\end{cases}$$

and $m_i(x)=\{s_i(x)\}^{-1}$. $\mathcal{G}^i$ can be written as follows:

$$\mathcal{G}^i=\left(\frac{d}{dx}\right)^2+b_i(x)\frac{d}{dx},$$

$$b_i(x)=\begin{cases}-(1-x)^{-1+k^{-1}}, & 1-k^{-k}\leqq x<1,\\ (1+x)^{-1+k^{-1}}, & -1<x\leqq -1+k^{-k},\\ 0, & \text{otherwise.}\end{cases}$$

We note that it is not necessary that s_i are continuous.

Example 5.4. Let $I=(-1, 1)$,

$$s_i(x)=\begin{cases}2-\dfrac{1}{i}, & -1<x<\dfrac{1}{2}-\dfrac{1}{i+1},\\ 2-\dfrac{1}{i+1}, & \dfrac{1}{2}-\dfrac{1}{i+1}\leqq x<1,\end{cases}$$

and $m_i(x)=1$. Then these s_i and m_i satisfy the conditions of Theorem 1.1.

REFERENCES

[1] A. BENSOUSSAN AND J. L. LIONS, *Nouvelles methodes en contrôle impulsionnel*, Appl. Math. Optim., 1 (1975), pp. 289-312.

[2] L. C. EVANS AND A. FRIEDMAN, *Optimal stochastic switching and the Dirichlet problem for the Bellman equation*, Trans. Amer. Math. Soc., 253 (1979), pp. 365-389.

[3] M. FUKUSHIMA, *Dirichlet Forms and Markov Processes*, North-Holland/Kodansha, Tokyo, 1980.

[4] N. IKEDA, M. NAGASAWA AND S. WATANABE, *Branching Markov processes* II, J. Math. Kyoto Univ., 8 (1968), pp. 365-410.

[5] K. ITO AND H. P. MCKEAN, *Diffusion Processes and Their Sample Paths*, Springer-Verlag, New York, 1965.

[6] P. L. LIONS, *Resolution analytique des problèmes de Bellman-Dirichlet*, Acta Mathematica, 146 (1981), pp. 151-166.

[7] P. L. LIONS AND J. L. MENALDI, *Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations* I, this Journal, 20 (1982), pp. 58-81.

[8] H. NAGAI, *Impulsive control of symmetric Markov processes and quasi-variational inequalities*, Osaka J. Math, 20 (1983), pp. 863-879.

[9] ———, *On an optimal stopping problem and a variational inequality*, J. Math. Soc. Japan, 30 (1978), pp. 303-312.

[10] M. NISIO, *On stochastic optimal controls and envelope of Markov semi-groups*, Proc. International Symposium on SDE, Kyoto, 1976, pp. 297-325.

[11] J. ZABCZYK, *Optimal control by means of switchings*, Studia Math., 45 (1973), pp. 161-171.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

004

CONVERGENCE RATES OF QUASI-NEWTON ALGORITHMS FOR SOME NONSMOOTH OPTIMIZATION PROBLEMS*

EKKEHARD SACHS†

Abstract. In this paper we consider a class of nonsmooth optimization problems and investigate an algorithm which makes use of approximations of the derivative. We study a growth condition on the objective and various conditions on the step-sizes and the quasi-Newton operators to obtain linear, superlinear and quadratic rates of convergence. These results are applied to a class of Broyden updates and two inexact step-size rules.

Key words. nonsmooth optimization, quasi-Newton methods, superlinear convergence rate.

1. Introduction. In this paper we investigate the local convergence properties of methods for the numerical solution of the following problems.

Throughout the paper, let X, Y be normed linear spaces, W a nonempty bounded and convex subset of X and mappings

$$G: X \to Y, \qquad G \text{ nonlinear},$$

$$\phi: Y \to R, \qquad \phi \text{ convex},$$

be given. Find $\hat{w} \in W$ such that for all $w \in W$

$$\phi(G\hat{w}) \leqq \phi(Gw). \tag{1.1}$$

We distinguish between ϕ and G instead of using $f = \phi \circ G$, because we do not impose any differentiability requirements on ϕ, but assume that G satisfies certain smoothness conditions.

There are various examples of classes of problems which are of the type of minimization problem (1.1):

Nonlinear approximation problems: G represents a nonlinear parametrization of the approximating family of functions and ϕ is a nondifferentiable norm, such as the L_1- or L_∞-norm.

Nonlinear control problems: G is the operator which maps the input (control) into the output (state). This can be described by a nonlinear ordinary differential equation, partial differential equation or integral equation. ϕ itself represents the cost-functional which is not required to be differentiable.

For a more detailed presentation of applications see Bertsekas [2] and Sachs [22].

In order to solve these problems we linearize the cost-functional as far as possible, which means we replace $\phi(G(\cdot))$ at a point $w_i \in W$ by $\phi(Gw_i + G'_{w_i}(\cdot - w_i))$. However, this requires the computation of G'_{w_i} which could become costly as in the control problem example. Thus we use an approximation B_i of the derivative G'_{w_i} and we obtain a quasi-Newton-type method.

Let $L(X, Y)$ denote the space of all linear and continuous operators from X into Y.

ALGORITHM

Step 0: Select $w_0 \in W$, an update procedure for $B_i \in L(X, Y)$ and a step size procedure for λ_i.

* Received by the editors August 28, 1981, and in revised form January 25, 1984.

† Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27695-8205.

Step 1: Compute $v_i \in W$ such that for all $v \in W$

$$\phi(Gw_i + B_i(v_i - w_i)) \leqq \phi(Gw_i + B_i(v - w_i)).$$

Step 2: Determine the step-size $\lambda_i \in [0, 1]$.

Step 3: Set $w_{i+1} = w_i + \lambda_i(v_i - w_i)$ and determine $B_{i+1} \in L(X, Y)$ by the update formula.

Step 4: Set $i = i + 1$ and go to Step 1.

For the purpose of abbreviation we introduce a set valued mapping L_B which maps W into subsets of W.

DEFINITION. For each $w \in W$, $B \in L(X, Y)$ define

$$L_B(w) = \{v \in W: \phi(Gw + B(v - w)) \leqq \phi(Gw + B(u - w)) \text{ for all } u \in W\}. \tag{1.2}$$

L_B can be single-valued or even $L_B(w) = \varnothing$ can occur. The latter case is a question of existence of solution of optimization problems which need to be solved in Step 1 and is discussed in [22]. In the case that $w \in L_B(w)$ holds, we arrived at a stationary point and the algorithm should be stopped, see [22].

As it is evident from Step 1, the function ϕ of the objective and the convex set W have to be structured in such a way that the minimization in Step 1 can be performed. For example, linear Chebyshev approximation problems with linear constraints can be solved by various powerful routines.

The step-size determination in Step 2 has been incorporated because it is essential for global convergence results, see Sachs [22]. We consider two versions which resemble the ones given for smooth optimization problems by Armijo [1] and Goldstein [9].

Problems of the type (1.1) have been investigated numerically by Bertsekas [2] and Poljak [21]. Both authors rewrite it as a constrained problem

$$\min \phi(y), \quad y = Gw, \quad w \in W,$$

and use the augmented Lagrangian method for its solution.

For the given algorithm a global convergence analysis has been presented in Sachs [22]. Local convergence rate results for methods of this type without step-size determination are proven in Gruver and Sachs [11]. In a recent publication, Mine and Fukushima [19] consider a sequence of convex subproblems for minimizing a sum of a smooth and a nonsmooth function.

Since $\phi \circ G$ is a nonconvex and nondifferentiable, but locally Lipschitz-continuous function, certain subgradient techniques are also applicable, see e.g. Goldstein [10] and Mifflin [18].

Probably the most extensive literature can be found for nonlinear minimax problems or nonlinear Chebyshev approximation. Algorithms for these problems have been proposed since 1959 by Ishizaki and Watanabe [16], Zuhovickii, Poljak, and Primak [23], and Osborne and Watson [20] until very recent papers, e.g. by Han [14] or Hald and Madsen [12]. If we confine ourselves to methods which use quasi-Newton-updates, we find a quasi-Newton-type method for discrete Chebyshev approximation in Madsen [17], Hald and Schjaer-Jacobsen [13], and Hald and Madsen [12]. Other applications of quasi-Newton methods by Hornung [15] and Han [14] use different algorithmic schemes.

Section 2 is devoted to convergence rates. Linear convergence is obtained, if the step-sizes stay away from 0 and if the approximations B_i are close enough to $G'_{\hat{w}}$. Superlinear convergence holds, if the step-sizes converge to 1 and if the operators B_i approach $G'_{\hat{w}}$ on a sequence of one-dimensional subspaces. This condition is a well-

known characterization of superlinear convergence for quasi-Newton methods in the smooth case, see Dennis and Moré [5]. Under more stringent conditions we prove sharper convergence rate estimates such as quadratic convergence.

In the third section of this paper we consider a growth condition of the nonlinear objective $\phi \circ G$ as we move away from the minimum. This condition is essential for the convergence rate proven later in the context, see Example 2.7. It is known in Chebyshev approximation under the name of "strong uniqueness" which is closely related to Haar's systems of functions, see Cheney [3]. In the convergence rate analysis for conditional gradient methods and related Newton-type methods, similar growth conditions have been used by Dunn [7], [8]. We give two characterizations of the growth condition using the derivative of G and prove a continuous dependence for the solution of the linearized problem at the point of linearization.

Proof for the results on the convergence rates and the properties implied by the growth condition are the contents of §§ 4 and 5.

Broyden updates are shown to satisfy the linear convergence requirements and also those for superlinear convergence, provided the range space of B_i and $G'_{\hat{w}}$ is finite-dimensional. These theorems are the subject of § 6.

In the last section we give modified versions of Goldstein's and Armijo's step-size rules which have been used in Sachs [22] to prove global convergence results. Here we show that if we are close enough to the optimum, then the step-sizes are automatically set equal to 1. This implies that the condition for a superlinear convergence rate is always satisfied.

The theory has been provided for infinite-dimensional spaces in order to be able to use it for control problems and to provide a basis for a convergence analysis of adaptive methods, where the discretization error of the solution of the differential equation is subsequently reduced. However, also from a finite-dimensional viewpoint we give some more insight into the convergence behavior for these quasi-Newton-type methods.

2. Theorem on rates of convergence. In the derivation of convergence rates for optimization algorithms, various conditions such as regularity of the gradient at the optimal point or positive definiteness of the second derivative play an important role. In many cases they imply a certain growth for the objective as one moves away from the minimal point. The condition which we will use can be interpreted as a certain uniqueness property or as a growth condition.

DEFINITION 2.1 (*growth condition*). The function $\phi \circ G$ satisfies the growth condition at $\hat{w}$, if there exist numbers $\alpha, \beta > 0$ such that for all $w \in W$ with $\|w - \hat{w}\| \leqq \alpha$ we have

$$\phi(Gw) - \phi(G\hat{w}) \geqq \beta \|w - \hat{w}\|. \tag{2.1}$$

This inequality does not only imply that $\hat{w}$ is a strict minimal point but gives an estimate on the growth of $\phi \circ G$ when moving away from $\hat{w}$.

Conditions of type (2.1) have been used in various contexts. In Chebyshev approximation it is called a local strong uniqueness condition and has been used for proofs of convergence rates by Cromme [4]. In a similar form it also occurs in the convergence rate analysis of algorithms with smooth objectives. We refer to the papers by Dunn [7] and [8].

Characterizations and implications of the growth condition are presented in § 3. An example also shows the necessity of these conditions in order to obtain the convergence rates.

According to different assumptions on the approximate behavior of B_i with respect to $G'_{\hat{w}}$ and the step lengths λ_i we achieve different estimates on the rate of convergence. The weakest result is a linear convergence rate.

THEOREM 2.2 (linear convergence). *Let ϕ be convex and continuous on Y and Lipschitz-continuous on bounded sets, G Fréchet differentiable at $\hat{w} \in W$, and let $\phi \circ G$ satisfy the growth condition at $\hat{w}$. Suppose $\{\lambda_i\}_{\mathbb{N}} \subset [0, 1]$ satisfies*

$$\lambda_i \geqq \lambda^* > 0 \quad \text{for all } i \in \mathbb{N}. \tag{2.2}$$

For each $\kappa \in (1-\lambda^*, 1)$ there is an $\varepsilon > 0$, such that if $w_1 \in W$ and $\{B_i\}_{\mathbb{N}} \subset L(X, Y)$ satisfy

$$\|w_1 - \hat{w}\| \leqq \varepsilon \tag{2.3}$$

and

$$\|B_i - G'_{\hat{w}}\| \leqq \varepsilon \quad \text{for all } i \in \mathbb{N}, \tag{2.4}$$

then for each sequence $\{w_i\}_{\mathbb{N}} \subset W$ defined recursively by

$$w_{i+1} = w_i + \lambda_i(v_i - w_i), \qquad v_i \in L_{B_i}(w_i) \backslash \{w_i\} \tag{2.5}$$

we obtain

$$\|w_{i+1} - \hat{w}\| \leqq \kappa \|w_i - \hat{w}\| \tag{2.6}$$

for $i \in \mathbb{N}$.

Theorem 2.2 shows that linear convergence is achieved, if the operators B_i stay in a neighborhood of $G'_{\hat{w}}$ and the step lengths λ_i are bounded away from zero. The following corollary gives an application of this theorem.

COROLLARY 2.3. *Let the assumptions of Theorem 2.2 be satisfied and select $\lambda_i = \lambda \in (0, 1]$. Let $G'_{(\cdot)}$ be continuous at $\hat{w}$ and let*

$$w_{i+1} = w_i + \lambda(v_i - w_i), \qquad v_i \in L_{B_i}(w_i).$$

For each rate factor $\kappa \in (1-\lambda, 1)$ there exists $\varepsilon > 0$ such that

$$\|w_1 - \hat{w}\| \leqq \varepsilon$$

and

$$\|B_i - G'_{w_i}\| \leqq \varepsilon \quad \text{for } i \in \mathbb{N} \tag{2.7}$$

imply the linear convergence rate (2.6).

If G is given by a nonlinear differential equation as e.g. in control problems, then G'_{w_i} is defined through a system of linear differential equations. Usually, this is solved by a discretization scheme. Corollary 2.3 tells us that the local convergence rate is linear if the discretization error is small enough during the whole iteration process, but does not need to tend to zero as the minimal point is approached.

For superlinear convergence we have the following conditions.

THEOREM 2.4 (superlinear convergence). *Let ϕ be convex and continuous on Y and Lipschitz-continuous on bounded sets, G Fréchet differentiable at $\hat{w} \in W$, and let $\phi \circ G$ satisfy the growth condition at $\hat{w}$. There exists $\varepsilon > 0$ such that if $\{B_i\}_{\mathbb{N}} \subset L(X, Y)$, $v_i \in L_{B_i}(w_i)$, and $\{\lambda_i\}_{\mathbb{N}} \subset [0, 1]$ fulfill*

$$|\lambda_i - 1| \leqq \varepsilon, \qquad \lim_{i \to \infty} \lambda_i = 1 \tag{2.8}$$

and

$$\|w_1 - \hat{w}\| \leqq \varepsilon, \qquad \|B_i - G'_{\hat{w}}\| \leqq \varepsilon \quad \text{for all } i \in \mathbb{N}, \tag{2.9}$$

and

$$\lim_{i\to\infty}\frac{\|(B_i - G'_{\hat{w}})(v_i - w_i)\|}{\|v_i - w_i\|} = 0, \tag{2.10}$$

then we obtain for the sequence $\{w_i\}_{\mathbb{N}} \subset W$, *defined by*

$$w_{i+1} = w_i + \lambda_i(v_i - w_i), \tag{2.11}$$

$$\lim_{i\to\infty}\frac{\|w_{i+1} - \hat{w}\|}{\|w_i - \hat{w}\|} = 0. \tag{2.12}$$

Observe that in order to obtain a superlinear rate of convergence, the step-sizes have to converge to zero and the B_i have to approximate $G'_{\hat{w}}$ in the direction $v_i - w_i$ with increasing accuracy. However, it is not required that B_i has to converge to $G'_{\hat{w}}$ in the operator-norm; it only has to stay in a small neighborhood of $G'_{\hat{w}}$.

A refinement of Theorem 2.4 is the following.

THEOREM 2.5. *Let ϕ be convex and continuous on Y and Lipschitz-continuous on bounded sets, let $\phi \circ G$ satisfy the growth condition at $\hat{w} \in W$ and let G be Fréchet differentiable at $\hat{w} \in W$ such that for some ε^*, $\rho > 0$, $\nu \in (0, 1]$ the following inequality holds for all $w \in W$ with $\|w - \hat{w}\| \leqq \varepsilon^*$*

$$\|Gw - G\hat{w} - G'_{\hat{w}}(w - \hat{w})\| \leqq \rho\|w - \hat{w}\|^{1+\nu}. \tag{2.13}$$

There exists $\varepsilon \in (0, \varepsilon^*]$ such that if $\{B_i\}_{\mathbb{N}} \subset L(X, Y)$, $v_i \in L_{B_i}(w_i)$, and $\{\lambda_i\}_{\mathbb{N}} \subset (0, 1]$ satisfy for fixed $\delta > 0$ and all $i \in \mathbb{N}$

$$1 - \lambda_i \leqq \delta\|v_i - w_i\|^{\nu}, \tag{2.14}$$

$$\|(B_i - G'_{\hat{w}})(v_i - w_i)\| \leqq \delta\|v_i - w_i\|^{1+\nu}, \tag{2.15}$$

$$\|w_1 - \hat{w}\| \geqq \varepsilon, \qquad \|B_i - G'_{\hat{w}}\| \leqq \varepsilon, \tag{2.16}$$

then we obtain for the sequence $\{w_i\}_{\mathbb{N}} \subset W$ defined by

$$w_{i+1} = w_i + \lambda_i(v_i - w_i) \tag{2.17}$$

that for some $\delta_0 > 0$

$$\frac{\|w_{i+1} - \hat{w}\|}{\|w_i - \hat{w}\|^{1+\nu}} \leqq \delta_0 \tag{2.18}$$

for i large enough.

Remark 2.6. The inequality (2.13) is satisfied for $\nu = 1$ if G is Fréchet differentiable on W and $G'_{(\cdot)}$ is Lipschitz-continuous on W. If we choose

$$B_i = G'_{w_i} \text{ and } \lambda_i = \min\{\tfrac{1}{2}, 1 - \delta\|v_i - w_i\|\}, \tag{2.19}$$

then under the assumptions of Theorem 2.5 we obtain a quadratic convergence rate.

In order to see that the growth condition (2.1) is essential for local convergence rates, let us discuss an example where (2.1) is not satisfied.

Example 2.7. Let $X = \mathbb{R}$, $Y = C[-1, 1]$, $W = [-\omega, \omega]$. We want to approximate uniformly the function $1 - x^2$ by the nonlinear family of functions $w^2 - 2wx - |w|^{1+\nu}$, $0 < \nu < 1$, on the interval $[-1, 1]$. Hence define $G: W \to Y$ by

$$(Gw)(x) = 1 - (x - w)^2 + |w|^{1+\nu},$$

$$\phi(y) = \max_{-1\leqq x\leqq 1} |y(x)|, \qquad y \in C[-1, 1].$$

Since for each $w \in W$

$$1+|w|^{1+\nu} \geqq (Gw)(x) \geqq 1-(1+|w|)^2+|w|^{1+\nu} \geqq -1-|w|^{1+\nu},$$

we have

$$\phi(Gw) = 1+|w|^{1+\nu},$$

$\hat{w} = 0$ is the solution and

$$\phi(Gw) - \phi(G0) = |w|^{1+\nu}, \qquad w \in W.$$

The growth condition (2.1) is not satisfied whenever $\nu > 0$. For the linearized problems we obtain for $v \in [-1, 1]$

$$(G'_w r)(x) = r(2(x-w)+(1+\nu)|w|^{\nu} \operatorname{sgn} w),$$

and for sufficiently small ω and ν

$$\begin{aligned}
&\|Gw + G'_w(v-w)\|_\infty \\
&\quad = \max_{-1 \leqq x \leqq 1} |1+w^{1+\nu}-(x-w)^2+2(v-w)(x-w)+(1+\nu)w^{\nu}(v-w)| \\
&\quad = \max_{-1 \leqq x \leqq 1} |1+|w|^{1+\nu}+(v-w)^2-(x-v)^2+(1+\nu)|w|^{\mu} \operatorname{sgn} w(v-w)| \\
&\quad = |1+|w|^{1+\nu}+(v-w)^2+(1+\nu)|w|^{\nu}(\operatorname{sgn} w)(v-w)| \\
&\quad = \left|1+|w|^{1+\nu}+\left(v-w+\frac{1+\nu}{2}|w|^{\nu} \operatorname{sgn} w\right)^2-\frac{(1+\nu)^2}{4}|w|^{2\nu}\right|.
\end{aligned}$$

The minimum is obtained for

$$v = w - \frac{1+\nu}{2}|w|^{\nu} \operatorname{sgn} w,$$

which results with step-size 1 into the iteration rule

$$w_{i+1} = w_i - \frac{1+\nu}{2}|w_i|^{\nu} \operatorname{sgn} w_i.$$

This is the type of iteration with $\lambda_i = 1$, $B_i = G'_{w_i}$ which should give a good local convergence rate. However, we show

$$|w_i|^{1-\nu} \leqq \frac{1+\nu}{4} \Leftrightarrow |w_{i+1}| \geqq |w_i|. \tag{2.20}$$

For $w_i > 0$ we have ($w_i < 0$ is treated analogously)

$$\begin{aligned}
w_{i+1} = w_i - \frac{1+\nu}{2}w_i^{\nu} \leqq -w_i \;&\Leftrightarrow\; 2w_i \leqq \frac{1+\nu}{2}w_i^{\nu} \\
&\Leftrightarrow\; w_i^{1-\nu} \leqq \frac{1+\nu}{4}.
\end{aligned}$$

Equation (2.20) tells that there is no convergence at all, no matter how close to the minimal point we select our starting point.

3. Discussion of growth condition. The growth condition (2.1) is of local nature and can be shown to be equivalent to a global condition for the linearized problem.

Since ϕ is convex and continuous, its directional derivative defined by

$$\phi'(w, h) = \lim_{\lambda \to 0^+} \frac{1}{\lambda}(\phi(G(w+\lambda h)) - \phi(Gw)), \qquad w, h \in X, \tag{3.1}$$

exists.

LEMMA 3.1. *Let ϕ be convex and continuous, and G be Fréchet differentiable at $\hat{w} \in W$. Then all the following conditions are equivalent:*

a) *There exist $\alpha, \beta > 0$ such that for all $w \in W$ with $\|w - \hat{w}\| \leqq \alpha$*

$$\phi(Gw) - \phi(G\hat{w}) \geqq \beta \|w - \hat{w}\|. \tag{2.1}$$

b) *There exists a $\gamma > 0$ such that for all $w \in W$*

$$\phi(G\hat{w} + G'_{\hat{w}}(w - \hat{w})) - \phi(G\hat{w}) \geqq \gamma \|w - \hat{w}\|. \tag{3.2}$$

c) *There exists a $\delta > 0$ such that for all $w \in W$*

$$\phi'(G\hat{w}, G'_{\hat{w}}(w - \hat{w})) \geqq \delta \|w - \hat{w}\|. \tag{3.3}$$

The next lemma states a local Lipschitz continuity of the operator $L_B(\cdot)$ near $\hat{w}$. Furthermore we obtain an estimate of the Lipschitz constant which tends to zero under certain conditions.

LEMMA 3.2. *Let ϕ be convex and continuous on Y and Lipschitz continuous on bounded sets and G Fréchet differentiable at $\hat{w} \in W$. Suppose $\phi \circ G$ satisfies the growth condition at $\hat{w} \in W$.*

(i) *For all $\kappa > 0$ there exists $\varepsilon > 0$ such that for every $B \in L(X, Y)$ and $w \in W$ with*

$$\|B - G'_{\hat{w}}\| \leqq \varepsilon \quad \text{and} \quad \|w - \hat{w}\| \leqq \varepsilon \tag{3.4}$$

we obtain

$$\|v - \hat{w}\| \leqq \kappa \|w - \hat{w}\| \tag{3.5}$$

for all $v \in L_B(w) \backslash \{w\}$.

(ii) *In particular there are $\varepsilon_0 > 0$, $\kappa_0 > 0$, and $\eta: \mathbb{R} \to \mathbb{R}$ with $\eta(r) \to 0$ for $r \to 0$ such that if (3.4) holds with $\varepsilon = \varepsilon_0$, then*

$$\|v - \hat{w}\| \leqq \left(\kappa_0 \frac{\|(B - G'_{\hat{w}})(v - w)\|}{\|v - w\|} + \eta(\|w - \hat{w}\|) \right) \|w - \hat{w}\| \tag{3.6}$$

for all $v \in L_B(w) \backslash \{w\}$.

The growth condition also implies a growth estimate in the neighborhood of $\hat{w}$.

LEMMA 3.3. *Let ϕ be convex and continuous on Y and Lipschitz-continuous on bounded sets and let G be Fréchet differentiable at $\hat{w} \in W$. Suppose $\phi \circ G$ satisfies the growth condition at $\hat{w}$. Then there exist γ_0, $\varepsilon > 0$ with the property: If $w \in W$ and $B \in L(X, Y)$ are such that*

$$\|w - \hat{w}\| \leqq \varepsilon \quad \text{and} \quad \|B - G'_{\hat{w}}\| \leqq \varepsilon,$$

then

$$\gamma_0 \|v - w\| \leqq \phi(Gw) - \phi(Gw + B(v - w)) \tag{3.7}$$

holds for all $v \in L_B(w)$.

Remark 3.4. In both preceding lemmas, the assumptions can be relaxed if Y is a finite-dimensional space. This follows from the fact that ϕ is locally Lipschitz continuous as a convex continuous function. Without the Lipschitz continuity of ϕ on bounded sets the statements (3.5)-(3.7) still hold for v in suitable neighborhoods of $\hat{w}$.

4. Proofs of § 3.

Proof of Lemma 3.1. The Fréchet differentiability of G yields

$$\phi(Gw) = \phi(G\hat{w} + G'_{\hat{w}}(w-\hat{w}) + d_w) \tag{4.1}$$

where

$$\|d_w\| \leqq \eta(\|w-\hat{w}\|)\|w-\hat{w}\|, \qquad \lim_{r\to 0}\eta(r) = 0. \tag{4.2}$$

Since ϕ is convex and continuous, it is locally Lipschitz-continuous at $G\hat{w}$. Hence, we infer from (4.1) and (4.2) that there exist $\varepsilon, \rho > 0$ such that for all $w \in X$ with $\|w-\hat{w}\| \leqq \varepsilon$

$$|\phi(Gw) - \phi(G\hat{w} + G'_{\hat{w}}(w-\hat{w}))| \leqq \rho\eta(\|w-\hat{w}\|)\|w-\hat{w}\|. \tag{4.3}$$

Assume (a) holds. Because of (4.2), there is an $\varepsilon_1 > 0$ such that for r with $0 < r < \varepsilon_1$ we have $\eta(r) \leqq \beta/2\rho$. Together with (4.1)-(4.3) we conclude from (2.1) for all $w \in W$ with $\|w-\hat{w}\| \leqq \varepsilon_2 = \min(\varepsilon, \varepsilon_1, \alpha)$

$$\begin{aligned}
&\phi(G\hat{w} + G'_{\hat{w}}(w-\hat{w})) - \phi(G\hat{w}) \\
&\quad = \phi(Gw) - \phi(G\hat{w}) + \phi(Gw) - \phi(G\hat{w} + G'_{\hat{w}}(w-\hat{w})) \\
&\quad \geqq (\beta - \rho\eta(\|w-\hat{w}\|))\|w-\hat{w}\| \geqq \frac{\beta}{2}\|w-\hat{w}\|.
\end{aligned} \tag{4.4}$$

This inequality however, is only the local version of (3.2). Let v be an arbitrary element of W with $v \neq \hat{w}$. Define

$$\lambda = \min(1, \varepsilon_2\|v-\hat{w}\|^{-1}),$$
$$v_\lambda = \hat{w} + \lambda(v-\hat{w}) \in W.$$

Then

$$\|v_\lambda - \hat{w}\| = \lambda\|v-\hat{w}\| \leqq \varepsilon_2,$$

and (4.4) and the convexity of ϕ imply

$$\begin{aligned}
\frac{\beta}{2}\lambda\|v-\hat{w}\| &= \frac{\beta}{2}\|v_\lambda - \hat{w}\| \\
&\leqq \phi(G\hat{w} + \lambda G'_{\hat{w}}(v-\hat{w})) - \phi(G\hat{w}) \\
&\leqq \lambda(\phi(G\hat{w} + G'_{\hat{w}}(v-\hat{w})) - \phi(G\hat{w})).
\end{aligned}$$

Division by λ yields (3.2) for each $v \in W$ with $\gamma = \beta/2$.

Assume (b) holds. Then, for $w \in W$ define $w_\lambda = \hat{w} + \lambda(w-\hat{w}) \in W$, $\lambda \in [0, 1]$. (3.2) implies

$$\begin{aligned}
\phi(G\hat{w} + \lambda G'_{\hat{w}}(w-\hat{w})) - \phi(G\hat{w}) &= \phi(G\hat{w} + G'_{\hat{w}}(w_\lambda - \hat{w})) - \phi(G\hat{w}) \\
&\geqq \gamma\|w_\lambda - \hat{w}\| = \lambda\gamma\|w-\hat{w}\|.
\end{aligned}$$

Division by λ and taking the limit for $\lambda \to 0$ yields (3.3) with $\delta = \gamma$.

Assume (c) is true. (4.2) implies that there is $\varepsilon_3 > 0$ such that $\eta(r) \leqq \delta/4\rho$ for all $0 < r \leqq \varepsilon_3$. Set $\alpha = \min(\varepsilon, \varepsilon_3)$ and take any $w \in W$ with $\|w-\hat{w}\| \leqq \alpha$. (3.3) yields that for λ small enough and positive

$$\begin{aligned}
\frac{\delta}{2}\|w-\hat{w}\| &\leqq \frac{1}{\lambda}(\phi(G\hat{w} + \lambda G'_{\hat{w}}(w-\hat{w})) - \phi(G\hat{w})) \\
&\leqq \phi(G\hat{w} + G'_{\hat{w}}(w-\hat{w})) - \phi(G\hat{w}).
\end{aligned} \tag{4.5}$$

Equations (4.3) and (4.5) imply

$$\begin{aligned}(Gw)-\phi(G\hat{w}) &= \phi(G\hat{w}-G'_{\hat{w}}(w-\hat{w}))-\phi(G\hat{w})+\phi(Gw)-\phi(G\hat{w}-G'_{\hat{w}}(w-\hat{w}))\\ &\geqq\left(\frac{\delta}{2}-\rho\eta(\|w-\hat{w}\|)\right)\|w-\hat{w}\|\\ &\geqq\frac{\delta}{4}\|w-\hat{w}\|.\end{aligned}$$

Hence (2.1) holds with $\beta=\delta/4$ and the proof is completed.

Proof of Lemma 3.2. (ii) By Lemma 3.1 the growth condition for $\phi\circ G$ at $\hat{w}\in W$ implies that (3.2) holds. This estimate and the definition of L_B with $u=\hat{w}$ in (1.2) yield

$$\begin{aligned}\gamma\|v-\hat{w}\| \leqq{}& \phi(G\hat{w}+G'_{\hat{w}}(v-\hat{w}))-\phi(Gw+B(v-w))\\ &+\phi(Gw+B(\hat{w}-w))-\phi(G\hat{w})\end{aligned}\tag{4.6}$$

for all $v\in L_B(w)$ with $w\in W$ and $B\in L(X, Y)$. If we pick some $\varepsilon^*>0$ and restrict w and B to (3.4) with $\varepsilon=\varepsilon^*$, then all arguments of ϕ in (4.6) lie in a bounded set. Hence for some constant $\kappa_1>0$ the following estimate follows from (4.6)

$$\begin{aligned}\gamma\|v-\hat{w}\| &\leqq \kappa_1(\|Gw-G\hat{w}+(B-G'_{\hat{w}})(v-w)-G'_{\hat{w}}(w-\hat{w})\|\\ &\qquad+\|Gw-G\hat{w}+B(\hat{w}-w)\|)\\ &\leqq \kappa_1(\|(B-G'_{\hat{w}})(v-w)\|+\|(G'_{\hat{w}}-B)(w-\hat{w})\|+2\eta_1(\|w-\hat{w}\|)\|w-\hat{w}\|),\end{aligned}\tag{4.7}$$

with

$$\begin{aligned}&\|Gw-G\hat{w}+G'_{\hat{w}}(w-\hat{w})\|\leqq\eta_1(\|w-\hat{w}\|)\|w-\hat{w}\|,\\ &\lim_{r\to 0}\eta_1(r)=0.\end{aligned}\tag{4.8}$$

With (4.7) we estimate further

$$\begin{aligned}\gamma\|v-\hat{w}\| &\leqq \kappa_1(2\|(B-G'_{\hat{w}})(v-w)\|+\|B-G'_{\hat{w}}\|\,\|v-\hat{w}\|+2\eta_1(\|w-\hat{w}\|)\|w-\hat{w}\|)\\ &\leqq \kappa_1\left(2\left(\frac{\|(B-G'_{\hat{w}})(v-w)\|}{\|v-w\|}+\eta_1(\|w-\hat{w}\|)\right)\|w-\hat{w}\|+3\|B-G'_{\hat{w}}\|\,\|v-\hat{w}\|\right)\end{aligned}\tag{4.9}$$

for all $w\in W$, $B\in L(X, Y)$, $v\in L_B(w)\backslash\{w\}$ with (3.4). Define

$$\varepsilon_0=\min\left\{\varepsilon^*,\frac{\gamma}{6\kappa_1}\right\},\quad \kappa_0=\frac{4\kappa_1}{\gamma},\quad \eta(\cdot)=\frac{4\kappa_1}{\gamma}\eta_1(\cdot).\tag{4.10}$$

Then for all $w\in W$, $B\in L(X, Y)$, $v\in L_B(w)\backslash\{w\}$ which satisfy

$$\|w-\hat{w}\|\leqq\varepsilon_0\leqq\varepsilon^*,\qquad \|B-G'_{\hat{w}}\|\leqq\varepsilon_0\leqq\varepsilon^*\tag{4.11}$$

we deduce with (4.9)-(4.11) that the following inequalities hold:

$$\begin{aligned}\|v-\hat{w}\| &= \frac{2}{\gamma}\left(\gamma-\frac{\gamma}{2}\right)\|v-\hat{w}\|\leqq\frac{2}{\gamma}(\gamma-3\varepsilon_0\kappa_1)\|v-\hat{w}\|\\ &\leqq\frac{2}{\gamma}(\gamma-3\kappa_1\|B-G'_{\hat{w}}\|)\|v-\hat{w}\|\\ &\leqq\frac{4\kappa_1}{\gamma}\left(\frac{\|(B-G'_{\hat{w}})(v-w)\|}{\|v-w\|}+\eta_1(\|w-\hat{w}\|)\right)\|w-\hat{w}\|\\ &=\left(\kappa_0\frac{\|(B-G'_{\hat{w}})(v-w)\|}{\|v-w\|}+\eta(\|w-\hat{w}\|)\right)\|w-\hat{w}\|.\end{aligned}$$

This completes the proof of (ii). In order to show (i), choose for given $\kappa > 0$ a number $\varepsilon \in (0, \varepsilon_0)$ with

$$\kappa_0 \varepsilon + \eta(\varepsilon) \leqq \kappa. \tag{4.12}$$

Thus for all $B \in L(X, Y)$, $w \in W$, $v \in L_B(w) \backslash \{w\}$ with (3.4) we obtain from (3.6)

$$\|v - \hat{w}\| \leqq (\kappa_0 \|B - G'_{\hat{w}}\| + \eta(\varepsilon)) \|w - \hat{w}\| \leqq \kappa \|w - \hat{w}\|.$$

Proof of Lemma 3.3. The growth condition of $\phi \circ G$ at $\hat{w}$ implies by Lemma 3.1 the existence of $\gamma > 0$ such that (3.2) holds for all $w \in W$. If we add $\phi(Gw) - \phi(Gw + B(v - w))$ on both sides of (3.2), we obtain

$$\begin{aligned} \phi(Gw) - \phi(Gw + B(v-w)) \geqq{}& \gamma \|w - \hat{w}\| + \phi(G\hat{w}) \\ & - \phi(Gw + B(v-w)) \\ & + \phi(Gw) - \phi(G\hat{w} + G'_{\hat{w}}(w - \hat{w})) \end{aligned} \tag{4.13}$$

for all $w \in W$, $B \in L(X, Y)$, $v \in L_B(w)$. Choose some $\varepsilon^* > 0$ and restrict $B \in L(X, Y)$ to

$$\|B - G'_{\hat{w}}\| \leqq \varepsilon^*. \tag{4.14}$$

Then all the arguments of ϕ in (4.13) lie in a bounded set and a Lipschitz constant $\kappa_1 > 0$ exists such that (4.13) can be estimated by

$$\begin{aligned} &\phi(Gw) - \phi(Gw + B(v-w)) \\ &\quad \geqq \gamma \|w - \hat{w}\| - \kappa_1 \|G\hat{w} - Gw - B(v-w)\| - \kappa_1 \|Gw - G\hat{w} - G'_{\hat{w}}(w - \hat{w})\|. \end{aligned}$$

Hence

$$\begin{aligned} \phi(Gw) - \phi(Gw + B(v-w)) \geqq{}& \gamma \|w - \hat{w}\| - 2\kappa_1 \eta(\|w - \hat{w}\|) \|w - \hat{w}\| \\ & - \kappa_1 \|(G'_{\hat{w}} - B)(v - w)\| - \kappa_1 \|G'_{\hat{w}}(v - \hat{w})\|, \end{aligned} \tag{4.15}$$

where $\eta: (0, \infty) \to \mathbb{R}$ satisfies by the Fréchet differentiability of G at $\hat{w}$

$$\|Gw - G\hat{w} - G'_{\hat{w}}(w - \hat{w})\| \leqq \eta(\|w - \hat{w}\|) \|w - \hat{w}\|$$

and

$$\lim_{r \to 0} \eta(r) = 0.$$

Choose $\varepsilon_1 > 0$ such that

$$|\eta(r)| \leqq \frac{\gamma}{12\kappa_1} \quad \text{for all } r \in (0, \varepsilon_1] \tag{4.16}$$

and define

$$\kappa = \gamma(6\kappa_1 \|G'_{\hat{w}}\|)^{-1}. \tag{4.17}$$

By Lemma 3.2 there exists $\varepsilon_2 > 0$ such that for $B \in L(X, Y)$, $w \in W$ with

$$\|B - G'_{\hat{w}}\| \leqq \varepsilon_2, \qquad \|w - \hat{w}\| \leqq \varepsilon_2 \tag{4.18}$$

we obtain for all $v \in L_B(w) \backslash \{w\}$

$$\|v - \hat{w}\| \leqq \kappa \|w - \hat{w}\| \tag{4.19}$$

and hence

$$\|v-w\| \leqq (\kappa+1)\|w-\hat{w}\|. \tag{4.20}$$

Define

$$\varepsilon = \min\left\{\varepsilon^*, \varepsilon_1, \varepsilon_2, \frac{\gamma}{6\kappa_1(\kappa+1)}\right\}, \qquad \gamma_0 = \frac{\gamma}{2(\kappa+1)}. \tag{4.21}$$

If $w \in W$ and $B \in L(X, Y)$ are such that

$$\|w-\hat{w}\| \leqq \varepsilon \quad \text{and} \quad \|B - G'_{\hat{w}}\| \leqq \varepsilon, \tag{4.22}$$

then we deduce from (4.15) with (4.16)-(4.22) that

$$\begin{aligned}
&\phi(Gw) - \phi(Gw + B(v-w)) \\
&\quad \geqq \gamma\|w-\hat{w}\| - 2\kappa_1\frac{\gamma}{12\kappa_1}\|w-\hat{w}\| - \kappa_1\varepsilon(\kappa+1)\|w-\hat{w}\| - \kappa_1\kappa\|G'_{\hat{w}}\|\,\|w-\hat{w}\| \\
&\quad \geqq \frac{\gamma}{2}\|w-\hat{w}\| \geqq \frac{\gamma}{2(\kappa+1)}\|v-w\| = \gamma_0\|v-w\|,
\end{aligned}$$

which completes the proof.

5. Proofs of § 2.

Proof of Theorem 2.2. For given $\lambda^* \in (0, 1)$ and $\kappa \in (1-\lambda^*, 1)$ define

$$\bar{\kappa} = 1 - \frac{1-\kappa}{\lambda^*} \in (0, 1). \tag{5.1}$$

By Lemma 3.2 there exists $\varepsilon > 0$ such that (3.4) implies

$$\|v-\hat{w}\| \leqq \bar{\kappa}\|w-\hat{w}\| \tag{5.2}$$

for all $v \in L_B(w)$ and w and B with (3.4). Therefore, (2.5) implies

$$\begin{aligned}
\|w_2-\hat{w}\| &\leqq (1-\lambda_1)\|w_1-\hat{w}\| + \lambda_1\|v_1-\hat{w}\| \\
&\leqq (1-\lambda_1(1-\bar{\kappa}))\|w_1-\hat{w}\| \\
&\leqq (1-\lambda^*(1-\bar{\kappa}))\|w_1-\hat{w}\| \\
&= \kappa\|w_1-\hat{w}\|.
\end{aligned}$$

Since in particular $\|w_2-\hat{w}\| < \|w_1-\hat{w}\| \leqq \varepsilon$, an induction argument along the same lines then shows the linear convergence rate.

Proof of Theorem 2.4. Conditions (2.8) and (2.9) imply from Theorem 2.2 the linear convergence of $\{w_i\}_{\mathbb{N}}$ to $\hat{w}$. Using the estimate (3.6) in Lemma 3.2(ii) we obtain

$$\lim_{i\to\infty}\frac{\|v_i-\hat{w}\|}{\|w_i-\hat{w}\|} \leqq \kappa_0 \lim_{i\to\infty}\frac{\|(B_i - G'_{\hat{w}})(v_i-w_i)\|}{\|v_i-w_i\|} + \kappa_0 \lim_{i\to\infty}\eta(\|w_i-\hat{w}\|) = 0.$$

Hence,

$$\lim_{i\to\infty}\frac{\|w_{i+1}-\hat{w}\|}{\|w_i-\hat{w}\|} \leqq \lim_{i\to\infty}(1-\lambda_i) + \lim_{i\to\infty}\lambda_i\frac{\|v_i-\hat{w}\|}{\|w_i-\hat{w}\|} = 0.$$

Proof of Theorem 2.5. (2.13) implies for the function η in Lemma 3.2(ii) that for some fixed $\rho_1 > 0$

$$\eta(r) \leqq \rho_1 r^{\nu} \tag{5.3}$$

for all $r \in [0, \varepsilon^*]$. (3.6) in connection with (2.15) yields

$$\|v_i - \hat{w}\| \leqq (\kappa_0 \delta \|v_i - w_i\|^\nu + \rho_1 \|w_i - \hat{w}\|^\nu) \|w_i - \hat{w}\|. \tag{5.4}$$

Choose ε in (2.16) so small that Lemma 3.2(i) yields

$$\|v_i - \hat{w}\| \leqq \|w_i - \hat{w}\| \tag{5.5}$$

for all $i \in \mathbb{N}$.

Then (5.4) can be estimated with (5.5) by

$$\|v_i - \hat{w}\| \leqq (2\kappa_0 \delta + \rho_1) \|w_i - \hat{w}\|^{1+\nu}. \tag{5.6}$$

The construction of the next iterate yields with (5.5), (5.6) and (2.14)

$$\begin{aligned} \|w_{i+1} - \hat{w}\| &\leqq (1 - \lambda_i) \|w_i - \hat{w}\| + \lambda_i \|v_i - \hat{w}\| \\ &\leqq \delta \|v_i - w_i\|^\nu \|w_i - \hat{w}\| + \lambda_i (\rho_1 + 2\kappa_0 \delta) \|w_i - \hat{w}\|^{1+\nu} \\ &\leqq (2\delta + \rho_1 + 2\kappa_0 \delta) \|w_i - \hat{w}\|^{1+\nu}, \end{aligned}$$

which had to be shown.

Proof of Remark 2.6. The estimate (2.14) clearly follows from (2.19) and similarly implies (2.19) with the Lipschitz-continuity of $G'_{(\cdot)}$ for some constant $\kappa_1 > 0$

$$\|(G'_{w_i} - G'_{\hat{w}})(v_i - w_i)\| \leqq \kappa_1 \|w_i - \hat{w}\| \, \|v_i - w_i\|. \tag{5.7}$$

Equation (3.5) in Lemma 3.2 yields for some $\kappa \in (0, 1)$

$$\|w_i - \hat{w}\| \leqq \|v_i - \hat{w}\| + \|w_i - v_i\| \leqq \kappa \|w_i - \hat{w}\| + \|w_i - v_i\|$$

i.e.

$$\|w_i - \hat{w}\| \leqq \frac{1}{1 - \kappa} \|w_i - v_i\|.$$

Substitute this inequality into (5.7) to obtain (2.15) with $\nu = 1$.

6. Update formulas. In order to approximate G'_{w_i} by B_i there are in general two classes of formulas. If G'_{w_i} is given by a linear differential or integral equation, B_i could be a discretization scheme. A check of the local and global convergence requirement should be performed for each update formula individually. In this section we investigate the so-called Broyden-class of operators B_i. In order to prove (2.10) we shall make use of an inner product defined in the space X. For this section we suppose the following

Assumption 6.1. Let X be equipped with an inner product $\langle \cdot, \cdot \rangle$ and with the norm

$$\|\cdot\| = \sqrt{\langle \cdot, \cdot \rangle}.$$

For example, if we are dealing with control problems and select $X = L_\infty[0, T]$, then, in order to satisfy Assumption 6.1, we equip $L_\infty[0, T]$ with the L_2-norm and the usual inner product

$$\langle a, b \rangle = \int_0^T a(t) b(t) \, dt, \qquad a, b \in L_\infty[0, T],$$

and obtain a pre-Hilbert space.

As pointed out in [11, p. 100ff], rank-one updates with useful estimates reduce to

$$B_{i+1} = \rho_i \left(B_i - \sigma_i \frac{]B_i p_i, p_i[}{\langle p_i p_i \rangle} \right) + \sigma_i \frac{]y_i, p_i[}{\langle p_i, p_i \rangle} \tag{6.1}$$

where $y_i = Gw_{i+1} - Gw_i$, $p_i = w_{i+1} - w_i$, ρ_i, $\sigma_i \in \mathbb{R}$, and operators $]y, x[\in L(X, Y)$ defined by

$$]y, x[(r) = \langle x, r\rangle y \quad \text{for all } r \in X.$$

The factors ρ_i and σ_i are scaling factors.

THEOREM 6.2. *Let ϕ be convex and continuous on Y and Lipschitz continuous on bounded sets, G continuously Fréchet differentiable with a locally Lipschitz continuous derivative on a ball of radius ε^* around $\hat{w}$ and let $\phi \circ G$ satisfy the growth condition at $\hat{w}$. Suppose $\{\lambda_i\}_{\mathbb{N}} \subset [0, 1]$ satisfies*

$$\lambda_i \geqq \lambda^* > 0 \quad \textit{for all } i \in \mathbb{N}. \tag{6.2}$$

For each $\kappa \in (1-\lambda^*, 1)$ there is $\varepsilon > 0$, such that if $w_1 \in W$ and $B_1 \in L(X, Y)$ satisfy

$$\|w_1 - \hat{w}\| \leqq \varepsilon, \tag{6.3}$$

$$\|B_1 - G'_{\hat{w}}\| \leqq \varepsilon, \tag{6.4}$$

then we obtain a linear convergence rate for the iterates

$$w_{i+1} = w_i + \lambda_i(v_i - w_i), \qquad v_i \in L_{B_i}(w_i), \tag{6.5}$$

provided the $\{B_i\}_{\mathbb{N}} \subset L(X, Y)$ are updated by (6.1) *with*

$$0 \leqq \sigma_i \leqq 2 \tag{6.6}$$

and

$$|1 - \rho_i| \leqq \delta \|w_i - \hat{w}\| \tag{6.7}$$

for some fixed $\delta > 0$.

Proof. Select $\kappa \in (1-\lambda^*, 1)$ and define a number $\bar{\kappa} \in (0, 1)$ by

$$\bar{\kappa} = 1 - \frac{1-\kappa}{\lambda^*}.$$

By Lemma 3.2 there exists $\varepsilon_1 > 0$ such that for $\varepsilon_2 = \min(\varepsilon^*, \varepsilon_1)$

$$\|B - G'_{\hat{w}}\| \leqq \varepsilon_2 \quad \text{and} \quad \|w - \hat{w}\| \leqq \varepsilon_2 \tag{6.8}$$

imply

$$\|v - \hat{w}\| \leqq \bar{\kappa}\|w - \hat{w}\| \quad \text{for all } v \in L_B(w), \quad v \neq w. \tag{6.9}$$

By induction we prove that (6.8) holds for all $i \in \mathbb{N}$. (6.8) is true for $i = 1$ by choosing $\varepsilon \leqq \varepsilon_2$. Suppose (6.8) is satisfied for B_i, w_i, $i = 1, 2, \cdots, r-1$, $r-1 \in \mathbb{N}$ fixed. (6.1) implies

$$\|B_{i+1} - G'_{\hat{w}}\| \leqq \|\rho_i B_i - G'_w\| \left\| I - \sigma_i \frac{]p_i, p_i[}{\langle p_i, p_i\rangle} \right\| + \frac{\sigma_i}{\langle p_i, p_i\rangle} \|]y_i - G'_{\hat{w}} p_i, p_i[\|. \tag{6.10}$$

The inner product on X yields with (6.6)

$$\left\| I - \sigma_i \frac{]p_i, p_i[}{\langle p_i, p_i\rangle} \right\| = 1. \tag{6.11}$$

From (6.9) we conclude as in the proof of Theorem 2.2 that

$$\|w_{i+1} - \hat{w}\| = \|w_i + \lambda_i(v_i - w_i) - \hat{w}\| \leqq \kappa \|w_i - \hat{w}\|, \qquad i = 1, \cdots, r-1.$$

This inequality and the local Lipschitz continuity of $G'_{(\cdot)}$ yield for $i = 1, \cdots, r-1$

$$
\begin{aligned}
\|]y_i - G'_{\hat{w}}p_i, p_i[\| &\leqq \|p_i\| \, \|y_i - G'_{\hat{w}}p_i\| \\
&\leqq \|(G'_{w_i+\tau_i p_i} - G'_{\hat{w}})\| \, \|p_i\|^2 \\
&\leqq \gamma(\tau_i \|w_{i+1} - \hat{w}\| + (1-\tau_i)\|w_i - \hat{w}\|)\|p_i\|^2 \\
&\leqq \gamma \|w_i - \hat{w}\| \, \|p_i\|^2.
\end{aligned} \tag{6.12}
$$

The existence of $\tau_i \in [0, 1]$ is a consequence of the continuity of $G'_{(\cdot)}$ and the mean value theorem, e.g. Dieudonné [6, Thm. 8.6.2], and $\gamma > 0$ denotes the Lipschitz constant.

Choose

$$
\varepsilon = \varepsilon_2(1-\kappa)(1+2\gamma+\delta(\varepsilon_2+\|G'_{\hat{w}}\|))^{-1}. \tag{6.13}
$$

The linear convergence rate estimate implies for $i = 1, \cdots, r-1$

$$
\|w_{i+1} - \hat{w}\| \leqq \kappa^i \|w_1 - \hat{w}\| \leqq \kappa^i \varepsilon. \tag{6.14}
$$

Equation (6.10) can be further estimated using (6.11), (6.13), (6.6) and (6.7)

$$
\begin{aligned}
\|B_{i+1} - G'_{\hat{w}}\| &\leqq \|B_i - G'_{\hat{w}}\| + |1-\rho_i| \, \|B_i\| + \sigma_i \gamma \|w_i - \hat{w}\| \\
&\leqq \|B_i - G'_{\hat{w}}\| + (\delta \|B_i\| + 2\gamma)\|w_i - \hat{w}\|,
\end{aligned}
$$

where $i = 1, \cdots, r-1$. We use this inequality and (6.14) consecutively for $i = 1, \cdots, r-1$ to obtain with (6.14) and $0 < \kappa < 1$

$$
\begin{aligned}
\|B_r - G'_{\hat{w}}\| &\leqq \|B_1 - G'_{\hat{w}}\| + \sum_{i=1}^{r-1} (\delta \|B_i\| + 2\gamma)\|w_i - \hat{w}\| \\
&\leqq \varepsilon + \sum_{i=1}^{r-1} (\delta(\varepsilon_2 + \|G'_{\hat{w}}\|) + 2\gamma)\kappa^i \varepsilon \\
&\leqq \varepsilon + \varepsilon(\delta(\varepsilon_2 + \|G'_{\hat{w}}\|) + 2\gamma)\frac{1}{1-\kappa} = \varepsilon_2.
\end{aligned}
$$

Equation (6.14) already implies for $i = r-1$

$$
\|w_r - \hat{w}\| \leqq \kappa \|w_{r-1} - \hat{w}\| \leqq \kappa\varepsilon_2 < \varepsilon_2.
$$

Hence we can apply Theorem 2.2 which proves the linear rate of convergence.

In order to prove the superlinear convergence rate we need a stronger convergence requirement for the sequence of $\{B_i\}_{\mathbb{N}}$ converging to $G'_{\hat{w}}$. We show its validity for Broyden updates in a weak sense.

THEOREM 6.3. *Let G be continuously Fréchet differentiable with a Lipschitz continuous Fréchet derivative on a ball of radius ε^* around $\hat{w}$, and let $\{w_i\}_{\mathbb{N}} \subset W$ a sequence of iterates which converge linearly to $w \in W$, i.e.*

$$
\|w_{i+1} - \hat{w}\| \leqq \kappa \|w_i - \hat{w}\|
$$

for some $\kappa \in (0, 1)$ and all $i \in \mathbb{N}$. Suppose the associated sequence of operators $\{B_i\}_{\mathbb{N}}$ defined in (6.1) *with*

$$
\rho_i = 1, \qquad \sigma \leqq \sigma_i \leqq 2 - \sigma \tag{6.15}
$$

for all $i \in \mathbb{N}$ and some fixed $\sigma > 0$ satisfies for some $\varepsilon > 0$

$$
\|B_i - G'_{\hat{w}}\| \leqq \varepsilon.
$$

Then

$$\lim_{i\to\infty}\frac{l((B_i-G'_{\hat{w}})(p_i))}{\|p_i\|}=0 \quad \textit{for each } l\in Y^*. \tag{6.16}$$

Proof. From (6.1) we deduce for each $l\in Y^*$ and $x\in X$

$$\begin{aligned} l((B_{i+1}-G'_{\hat{w}})(x)) &= l((B_i-G'_{\hat{w}})(x))-\sigma_i\frac{\langle p_i,x\rangle}{\langle p_i,p_i\rangle}l(B_ip_i-y_i)\\ &= l\left((B_i-G'_{\hat{w}})\left(x-\sigma_i\frac{\langle p_i,x\rangle}{\langle p_i,p_i\rangle}p_i\right)\right)+\sigma_i\frac{\langle p_i,x\rangle}{\langle p_i,p_i\rangle}l(y_i-G'_{\hat{w}}p_i). \end{aligned} \tag{6.17}$$

If $\langle\cdot,\cdot\rangle$ denotes the inner product in X and $(B_i-G'_{\hat{w}})^*\colon Y^*\to X^*$ the formal adjoint operator, then with

$$b_i=(B_i-G'_{\hat{w}})^*l, \tag{6.18}$$

$$\sup_{\|x\|\leqq 1}\left\langle b_i,\, x-\sigma_i\frac{\langle p_i,x\rangle}{\langle p_i,p_i\rangle}p_i\right\rangle=\left(\langle b_i,b_i\rangle-\sigma_i(2-\sigma_i)\frac{\langle p_i,b_i\rangle^2}{\langle p_i,p_i\rangle}\right)^{1/2} \tag{6.19}$$

and as in (6.12) with $\gamma_1=(2-\sigma)\gamma>0$ for i sufficiently large

$$\sup_{\|x\|\leqq 1}\sigma_i\frac{\langle p_i,x\rangle}{\langle p_i,p_i\rangle}\|y_i-G'_{\hat{w}}p_i\|\leqq\gamma_1\|w_i-\hat{w}\|. \tag{6.20}$$

We take the supremum over all x in the unit ball in (6.17) and use (6.18)-(6.20) to obtain for $i\geqq i_0$

$$\begin{aligned} \|b_{i+1}\| &\leqq\left(\|b_i\|^2-\sigma_i(2-\sigma_i)\frac{\langle p_i,b_i\rangle^2}{\langle p_i,p_i\rangle}\right)^{1/2}+\|l\|\gamma_1\|w_i-\hat{w}\|\\ &\leqq\|b_i\|-(2-\sigma_i)\frac{\sigma_i}{2}\frac{\langle p_i,b_i\rangle^2}{\langle p_i,p_i\rangle\|b_i\|}+\|l\|\gamma_1\|w_i-\hat{w}\|. \end{aligned} \tag{6.21}$$

Since for all $i\in\mathbb{N}$

$$\|b_i\|\leqq\varepsilon\|l\|$$

holds by assumption, we obtain through a summation in (6.21)

$$0\leqq\frac{1}{2\varepsilon\|l\|}\sum_{i=i_0}^{\infty}(2-\sigma_i)\sigma_i\frac{\langle p_i,b_i\rangle^2}{\langle p_i,p_i\rangle}\leqq\|b_{i_0}\|+\gamma_1\|l\|\sum_{i=i_0}^{\infty}\|w_i-\hat{w}\|. \tag{6.22}$$

The linear convergence rate implies that

$$\sum_{i=1}^{\infty}\|w_i-\hat{w}\|<\infty$$

and using (6.15) in (6.22) we deduce that

$$\sum_{i=1}^{\infty}\frac{\langle p_i,b_i\rangle^2}{\langle p_i,p_i\rangle}\quad\text{is finite,}$$

and therefore

$$\lim_{i\to\infty}\frac{l((B_i-G'_{\hat{w}})(p_i))}{\|p_i\|}=0.$$

Condition (2.10) for superlinear convergence is satisfied, if for instance the range spaces of B_i and $G'_{\hat{w}}$ can be embedded in a finite-dimensional space, in particular if a finite-dimensional problem is being solved.

7. Step-size rules. In order to obtain global convergence results for this method, an implementation of step-size rules is necessary. An analysis of global convergence properties for three different step-size rules is carried out by the author in [22]. We will pursue this investigation with respect to local convergence in this paper for the two more practical, inexact step-size rules. This is their definition:

Modified Goldstein rule. Select $\delta \in (0, \frac{1}{2})$ for the algorithm, and let $B_i \in L(X, Y)$, $w_i \in W$, $v_i \in L_{B_i}(w_i)$ with

$$\phi(Gw_i + B_i(v_i - w_i)) < \phi(Gw_i). \tag{7.1}$$

If

$$\phi(Gv_i) - \phi(Gw_i) \leqq \delta(\phi(Gw_i + B_i(v_i - w_i)) - \phi(Gw_i)), \tag{7.2}$$

then set $\lambda_i = 1$, otherwise select $\lambda_i \in [0, 1]$ such that

$$\begin{aligned} &(1-\delta)\lambda_i(\phi(Gw_i + B_i(v_i - w_i)) - \phi(Gw_i)) \\ &\leqq \phi(G(w_i + \lambda_i(v_i - w_i))) - \phi(Gw_i) \\ &\leqq \delta\lambda_i(\phi(Gw_i + B_i(v_i - w_i)) - \phi(Gw_i)). \end{aligned} \tag{7.3}$$

Modified Armijo rule. Select $\delta \in (0, 1)$, $\beta \in (0, 1)$ for the algorithm, and let $B_i \in L(X, Y)$, $w_i \in W$, $v_i \in L_{B_i}(w_i)$ with (7.2). If (7.2) holds, set $\lambda_i = 1$. Otherwise select the smallest integer $k \in \mathbb{N}$ which satisfies

$$\phi(G(w_i + \beta^k(v_i - w_i))) - \phi(Gw_i) \leqq \delta\beta^k(\phi(Gw_i + B_i(v_i - w_i)) - \phi(Gw_i)) \tag{7.4}$$

and set $\lambda_i = \beta^k$.

For a discussion of certain properties of these step-size rules we refer to [22].

In the case of a superlinear rate of convergence it was required that the sequence of step sizes converges to 1. We show in the following theorem that this is the case for the two inexact step-size rules mentioned above.

THEOREM 7.1. *Let ϕ be convex and continuous on Y and Lipschitz-continuous on bounded sets, G continuously Fréchet differentiable on a ball of radius ε^* around $\hat{w} \in W$ and let $\phi \circ G$ satisfy the growth condition at $\hat{w}$. There exists $\varepsilon > 0$ such that if $\{B_i\}_{\mathbb{N}} \subset L(X, Y)$, $v_i \in L_{B_i}(w_i)$ fulfill* (7.1),

$$\|w_i - \hat{w}\| \leqq \varepsilon, \qquad \|B_i - G'_{\hat{w}}\| \leqq \varepsilon \quad \text{for all } i \in \mathbb{N} \tag{7.5}$$

and

$$\lim_{i\to\infty} \frac{\|(B_i - G'_{\hat{w}})(v_i - w_i)\|}{\|v_i - w_i\|} = 0, \tag{7.6}$$

then we obtain for the sequence $\{w_i\}_{\mathbb{N}} \subset W$, $\{\lambda_i\}_{\mathbb{N}}$ defined by

$$w_{i+1} = w_i + \lambda_i(v_i - w_i)$$

and either the Amijo or Goldstein step-size rule

$$\lim_{i\to\infty} \frac{\|w_{i+1} - \hat{w}\|}{\|w_i - \hat{w}\|} = 0.$$

Proof. We show that λ_i is set equal to 1 for i sufficiently large. By Lemma 3.3 there exist $\gamma_0 > 0$ and $\varepsilon_0 > 0$ such that (7.5) with $\varepsilon = \varepsilon_0$ implies for all $i \in \mathbb{N}$

$$\gamma_0 \| v_i - w_i \| \leqq \phi(Gw_i) - \phi(Gw_i + B_i(v_i - w_i)). \tag{7.7}$$

Since ϕ is Lipschitz continuous on bounded sets we also deduce from (7.5) with $\varepsilon = \varepsilon_0$ that for some $\kappa > 0$

$$\phi(Gv_i) - \phi(Gw_i + B_i(v_i - w_i)) \leqq \kappa \| Gv_i - Gw_i + B_i(v_i - w_i) \| \tag{7.8}$$

for all $i \in \mathbb{N}$. If $0 < \varepsilon \leqq \min(\varepsilon_0, \varepsilon^*)$, then the differentiability assumptions on G imply for (7.8)

$$\begin{aligned} &\phi(Gv_i) - \phi(Gw_i + B_i(v_i - w_i)) \\ &\qquad \leqq \kappa(\| (B_i - G'_{\hat{w}})(v_i - w_i) \| + \| G'_{z_i} - G'_{\hat{w}} \| \, \| v_i - w_i \|) \end{aligned} \tag{7.9}$$

with [6, Thm. 8.6.2] and

$$\| G'_{z_i} - G'_{\hat{w}} \| = \max_{0 \leqq \tau \leqq 1} \| G'_{w_i + \tau(v_i - w_i)} - G'_{\hat{w}} \|. \tag{7.10}$$

There exists $i_0 \in \mathbb{N}$ such that (7.6) implies

$$\frac{\| (B_i - G'_{\hat{w}})(v_i - w_i) \|}{\| v_i - w_i \|} \leqq \frac{\gamma_0}{2\kappa}(1 - \delta). \tag{7.11}$$

Furthermore, there exists $\varepsilon_2 > 0$ such that (7.5) yields with Lemma 3.2 and $\varepsilon = \varepsilon_2$

$$\| v_i - \hat{w} \| \leqq \| w_i - \hat{w} \|. \tag{7.12}$$

The continuity of $G'_{(\cdot)}$ implies the existence of $\varepsilon_3 > 0$ such that

$$\| u - \hat{w} \| \leqq \varepsilon_3 \Rightarrow \| G'_u - G'_{\hat{w}} \| \leqq \frac{\gamma_0}{2\kappa}(1 - \delta). \tag{7.13}$$

If $\varepsilon = \min(\varepsilon^*, \varepsilon_0, \varepsilon_1, \varepsilon_2, \varepsilon_3)$, then (7.13) holds for $u = z_i$ because of (7.12) and furthermore we have the following estimates using (7.7), (7.9), (7.11) and (7.13) and for $i \geqq i_0$

$$\begin{aligned} &\frac{\phi(Gv_i) - \phi(Gw_i)}{\phi(Gw_i + B_i(v_i - w_i)) - \phi(Gw_i)} \\ &\quad = 1 - \frac{\phi(Gv_i) - \phi(Gw_i + B_i(v_i - w_i))}{\phi(Gw_i + B_i(v_i - w_i)) - \phi(Gw_i)} \\ &\quad \geqq 1 - \frac{\kappa}{\gamma_0} \frac{\| (B_i - G'_{\hat{w}})(v_i - w_i) \| + \| G'_{z_i} - G'_{\hat{w}} \| \, \| v_i - w_i \|}{\| v_i - w_i \|} \\ &\quad = 1 - \frac{\kappa}{\gamma_0}\left(2\frac{\gamma_0}{2\kappa}(1 - \delta)\right) = \delta. \end{aligned}$$

Hence (7.2) is satisfied and we set $\lambda_i = 1$ for $i \geqq i_0$.

REFERENCES

[1] L. ARMIJO, *Minimization of functions having Lipschitz-continuous first derivatives*, Pacific J. Math., 16 (1966), pp. 1-3.

[2] D. P. BERTSEKAS, *Approximation procedure based on the method of multipliers*, J. Optim. Theory Appl., 23 (1977), pp. 487-510.

[3] E. W. CHENEY, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.

[4] L. CROMME, *Eine Klasse von Verfahren zur Ermittlung bester nichlinearer Tschebyscheff-Approximationen*, Numer. Math., 25 (1976), pp. 447-459.
[5] J. E. DENNIS AND J. J. MORÉ, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Math. Comp., 28 (1974), pp. 549-560.
[6] J. DIEUDONNÉ, *Foundations of Modern Analysis*, Academic Press, New York and London, 1960.
[7] J. C. DUNN, *Convergence rates for conditional gradient sequences generated by implicit step length rules*, this Journal, 18 (1980), pp. 473-487.
[8] ———, *Newton's method and the Goldstein step length rule for constrained minimization problems*, this Journal, 18 (1980), pp. 659-674.
[9] A. A. GOLDSTEIN, *On Newton's method*, Numer. Math., 7 (1965), pp. 391-393.
[10] ———, *Optimization of Lipschitz continuous functions*, Math. Programming, 13 (1977), pp. 14-22.
[11] W. A. GRUVER AND E. SACHS, *Algorithmic Methods in Optimal Control*, Pitman, London, 1981.
[12] J. HALD AND K. MADSEN, *Combined LP and quasi-Newton methods for minimax optimization*, Math. Programming, 20 (1981), pp. 49-62.
[13] J. HALD AND H. SCHJAER-JACOBSEN, *Linearly constrained minimax optimization without calculating derivatives*, Methods in Operations Research, 31 (1979), pp. 289-301.
[14] S. P. HAN, *Variable metric methods for minimizing a class of nondifferentiable functions*, Math. Programming, 20 (1981), pp. 1-13.
[15] R. HORNUNG, *Algorithm for the solution of a discrete minimax problem: subgradient methods and a new fast Newton method*, Proc. 9th IFIP Conference on Optimization Technology, K. Iracki, K. Malanowski, and S. Walukiewicz, eds. Part 2, Warsaw, 1979, Springer, Berlin-Heidelberg-New York, 1980, pp. 78-86.
[16] Y. ISHIZAKI AND H. WATANABE, *An iterative Chebyshev approximation method for network design*, IEEE Trans. Circuit Theory, 15 (1968), pp. 326-336.
[17] K. MADSEN, *Minimax solutions of nonlinear equations without calculating derivatives*, Math. Progr. Study, 3 (1975), pp. 110-126.
[18] R. MIFFLIN, *An algorithm for constrained optimization with semismooth functions*, Math. Oper. Res., 2 (1977), pp. 191-207.
[19] H. MINE AND M. FUKUSHIMA, *A minimization method for the sum of a convex function and a continuously differentiable function*, J. Optim. Theory Appl., 33 (1981), pp. 9-23.
[20] M. R. OSBORNE AND G. A. WATSON, *An algorithm for minimax approximation in the nonlinear case*, Comput. J., 12 (1969), pp. 64-69.
[21] B. T. POLJAK, *On the Bertsekas method for minimization of composite function*, in International Symposium on Systems Optimization and Analysis, Rocquencourt 1978, A. Bensoussan and J. L. Lions, eds., Springer, Berlin-Heidelberg-New York, pp. 179-186, 1979.
[22] E. SACHS, *Global convergence of quasi-Newton algorithms for some nonsmooth optimization problems*, J. Optim. Theory Appl., 40 (1983), pp. 201-219.
[23] S. I. ZUHOVICKII, R. A. POLJAK AND M. E. PRIMAK, *An algorithm for the solution of the problem of convex Čebyšev approximation*, Soviet Math., 4 (1963), pp. 901-904.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

005

DETERMINISTIC IMPULSE CONTROL PROBLEMS*

G. BARLES†

Abstract. We prove that the optimal cost function of a deterministic impulse control problem is the unique viscosity solution of a first-order Hamilton-Jacobi quasi-variational inequality in $\mathbb{R}^N$.

Key words. deterministic impulse control, dynamic programming principle, viscosity solution, first-order Hamilton-Jacobi equations, quasi-variational inequality

Introduction. Impulse control problems lead, via the dynamic programming principle, to the study of various kinds of quasi-variational inequalities (see for more details and for many examples A. Bensoussan and J. L. Lions [3]).

In this work, we are interested in deterministic impulse control in $\mathbb{R}^N$. More precisely, our main result states that the optimal cost function of a deterministic impulse control problem is the unique viscosity solution of a first-order Hamilton-Jacobi quasi-variational inequality in $\mathbb{R}^N$ of a particular form:

$$\max\left(H(x, u, Du), u - Mu\right) = 0 \quad \text{in } \mathbb{R}^N, \tag{P}$$

where

$$H(x, t, p) = \sup_{v \in V} \left(b(x, v) \cdot p + \lambda t - f(x, v)\right),$$

$$M\varphi(x) = \inf_{\xi \in (\mathbb{R}^+)^N} \left(\varphi(x+\xi) + c(\xi)\right);$$

V is a separable metric space, b and f are functions from $\mathbb{R}^N \times V$ into $\mathbb{R}^N$, c is a continuous positive function, $\lambda > 0$ (precise assumptions are detailed in § 2). Let us just mention that the state of the controlled system is given by the solution $y_x(t)$ of the following problem:

$$\frac{dy_x(t)}{dt} + b(y_x(t), v(t)) = 0, \qquad t \in]\theta_i, \theta_{i+1}[,$$

$$y_x(0) = x,$$

$$y_x(\theta_i + 0) = y_x(\theta_i - 0) + \xi_i,$$

where $\theta = (\theta_i)_{i \in \mathbb{N}}$ is a nondecreasing sequence of positive reals which satisfies: $\theta_n \to +\infty$ when $n \to +\infty$ and $\xi = (\xi_i)_{i \in N}$ is a sequence of elements of $(\mathbb{R}^+)^N$.

Finally, $v(t)$ is any measurable function which states its values in a compact subset of V. Finally, $K = (\theta, \xi, v(\cdot))$ is the *control.* The optimal cost function u is given by:

$$u(x) = \inf_K \left(\int_0^\infty f(y_x(t), v(t))\, e^{-\lambda t}\, dt + \sum_{i \in \mathbb{N}} c(\xi_i) c^{-\lambda \theta_i} \right).$$

We first recall below the definition and main properties of the notion of viscosity solutions for problems like (P). These results are easy extensions of those obtained by M. G. Crandall and P. L. Lions [6], M. G. Crandall, L. C. Evans and P. L. Lions [5] and P. L. Lions [10] for first-order Hamilton-Jacobi equations. Let us just mention

* Received by the editors December 1, 1983, and in revised form May 25, 1984.
† ENS St. Cloud et Ceremade, Impasse du Docteur Roux, 83150 Bandol, France.

that the form of M creates some difficulties for stability results (see for more details G. Barles [1]).

In § 2, we introduce the deterministic impulse control problem and we prove various results concerning the optimal cost function u which are similar to those obtained in the classical deterministic case (see P. L. Lions [10]). An essential tool is the dynamic programming principle (we give two forms of this result). We indicate some properties of u (its regularity, its behavior at infinity, the fact that u is the maximal subsolution of the Q.V.I. in the distribution sense). Finally, we show that u is a viscosity solution of (P).

In the last section, we prove that problem (P) has a unique viscosity solution. In this context, the particular form of H and the well-known results of optimal stopping time problems enable us to adapt a proof due to B. Hanouzet and J. L. Joly [9] for the elliptic Q.V.I.

The methods we use combine elements from the deterministic control (see P. L. Lions [10]), from the first-order Hamilton-Jacobi equations (especially the methods due to the notion of viscosity solutions—see M. G. Crandall, L. C. Evans and P. L. Lions [5], M. G. Crandall and P. L. Lions [6] or P. L. Lions [10]) and from the theory of elliptic Q.V.I. (see A. Bensoussan and J. L. Lions [3]).

1. Viscosity solutions for first-order Hamilton–Jacobi quasi-variational inequalities. We just want to recall the main equivalent definitions and to mention the most important properties of the viscosity solutions of first-order Hamilton-Jacobi quasi-variational inequalities, without proofs. More details and complete proofs can be found in G. Barles [1]. The notion of viscosity solutions of first-order Hamilton-Jacobi equations was introduced by M. G. Crandall and P. L. Lions [6] and all the results mentioned below are straightforward extensions of those obtained by M. G. Crandall and P. L. Lions [6], M. G. Crandall, L. C. Evans and P. L. Lions [5] or P. L. Lions [10].

1.1. Main definitions. We denote by BUC $(\mathbb{R}^N)$, the space of bounded and uniformly continuous functions on $\mathbb{R}^N$.

We recall the following notions of sub- and superdifferential of continuous functions considered in [5] and [6]. Let $\varphi \in C(\mathbb{R}^N)$.

DEFINITION 1.1. (i) The superdifferential of φ at $x_0 \in \mathbb{R}^N$, denoted by $D^+\varphi(x_0)$, is the set (possibly empty) defined by

$$D^+\varphi(x_0) = \left\{ p \in \mathbb{R}^N, \limsup_{x \to x_0} \frac{\varphi(x) - \varphi(x_0) - (p|x - x_0)}{|x - x_0|} \leqq 0 \right\}. \tag{1}$$

(ii) The subdifferential of φ at $x_0 \in \mathbb{R}^N$, denoted by $D^-\varphi(x_0)$, is the set given by $D^-\varphi(x_0) = -D^+(-\varphi)(x_0)$, i.e.,

$$D^-\varphi(x_0) = \left\{ p \in \mathbb{R}^N, \liminf_{x \to x_0} \frac{\varphi(x) - \varphi(x_0) - (p|x - x_0)}{|x - x_0|} \geqq 0 \right\}. \tag{2}$$

Remark 1.1. For $x \in \mathbb{R}^N$, $D^+\varphi(x)$ (resp. $D^-\varphi(x)$) is a closed convex set in $\mathbb{R}^N$.

Remark 1.2. The notion of subdifferential considered in [6] has been introduced independently for different purposes, by E. de Giorgi, A. Marino and M. Tosques [7] and A. Marino and M. Tosques [12].

DEFINITION 1.2. $u \in$ BUC $(\mathbb{R}^N)$ said to be a viscosity solution of the problem (P) if we have:

$$\forall y \in \mathbb{R}^N, \forall p \in D^+u(y), \quad \max\left(H(y, u(y), p), u - Mu\right) \leqq 0, \tag{3}$$

$$\forall y \in \mathbb{R}^N, \forall p \in D^- u(y), \quad \max\,(H(y, u(y), p), u - Mu) \geqq 0. \tag{4}$$

Let us give another equivalent definition which is more practical in particular to show uniqueness results.

PROPOSITION 1.1. *$u \in \mathrm{BUC}\,(\mathbb{R}^N)$ is a viscosity solution of the problem* (P) *if and only if the two following properties hold*:

$$\begin{aligned}&\forall \phi \in C^1(\mathbb{R}^N), \text{ at each local maximum point } x_0 \text{ of } u - \phi, \text{ we have}\\ &\max\,(H(x_0, u(x_0), D\phi(x_0)), u(x_0) - Mu(x_0)) \leqq 0;\end{aligned} \tag{5}$$

$$\begin{aligned}&\forall \phi \in C^1(\mathbb{R}^N), \text{ at each local minimum point } x_0 \text{ of } u - \phi, \text{ we have}\\ &\max\,(H(x_0, u(x_0), D\phi(x_0)), u(x_0) - Mu(x_0)) \geqq 0.\end{aligned} \tag{6}$$

Remark 1.3. The proof is exactly the same as for first-order Hamilton-Jacobi equations; we just have to consider the Hamiltonian:

$$\tilde{H}(x, t, p) = \max\,(H(x, t, p), t - Mu(x)).$$

The two definitions mean that u is a viscosity solution of the obstacle problem, with the implicit obstacle Mu. The following proposition shows how we can use the particular form of M.

PROPOSITION 1.2. *$u \in \mathrm{BUC}\,(\mathbb{R}^N)$ is a viscosity solution of the problem* (P) *if and only if the two following properties hold: for all $\varphi \in C_b^1(\mathbb{R}^N)$*:

$$\begin{aligned}&\text{at each } \underline{\text{global}} \text{ maximum point } x_0 \text{ of } u - \varphi, \text{ we have}\\ &\max\,(H(x_0, u(x_0), D\varphi(x_0)), \varphi(x_0) - M\varphi(x_0)) \leqq 0;\end{aligned} \tag{7}$$

$$\begin{aligned}&\text{at each } \underline{\text{global}} \text{ minimum point } x_0 \text{ of } u - \varphi, \text{ we have}\\ &\max\,(H(x_0, u(x_0), D\varphi(x_0)), \varphi(x_0) - M\varphi(x_0)) \geqq 0.\end{aligned} \tag{8}$$

Remark 1.4. This proposition can be extended to more general operators M. It suffices that M is nondecreasing and that M satisfies

$$\forall \phi \in \mathrm{BUC}\,(\mathbb{R}^N), \forall c \in \mathbb{R}, \quad M(\phi + c) = M(\phi) + c.$$

1.2. Main properties of viscosity solutions of problem (P). Most of the properties of the viscosity solutions of first-order Hamilton-Jacobi equations are still valid for the viscosity solutions of the problem (P). It is an easy consequence of Remark 1.3. Only the stability results need to be considered because of the nonlocal form of M. So, we give a stability result.

PROPOSITION 1.3. *Assume that u_ε is a viscosity solution of*

$$max\,(H_\varepsilon(x, u^\varepsilon, Du^\varepsilon), u^\varepsilon - Mu^\varepsilon) = 0 \quad in\ \mathbb{R}^N$$

and that H_ε converges to H uniformly on compact subsets of $\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^N$ and u^ε converges to $u \in \mathrm{BUC}\,(\mathbb{R}^N)$ uniformly on compact subsets of $\mathbb{R}^N$ when ε goes to zero, with the following condition:

$$\|(u^\varepsilon - u)^-\|_{L^\infty(\mathbb{R}^N)} \to 0 \quad when\ \varepsilon \to 0. \tag{9}$$

Then u is a viscosity solution of the problem (P).

Remark 1.5. With the same assumptions, the result is also valid for the vanishing viscosity method.

Remark 1.6. The condition (9) is quite optimal. One can find in G. Barles [1] a counterexample of the statement in Proposition 1.3 in the case when (9) is not satisfied.

More details and complete proofs can be found in G. Barles [1]; the proof of this result being essentially routine adaptations of M. G. Crandall and P. L. Lions [6].

2. The deterministic impulse control problem. Our purpose is to show that the optimal cost function of a deterministic impulse control problem is always a viscosity solution of a problem (P).

We also give regularity results and results concerning the behavior at infinity for this optimal cost function. All these results are analogous to those obtained in the classical deterministic control problem (see P. L. Lions [10]). These are obtained essentially by using the dynamic programming principle.

2.1. Setting of the problem. Through all this section, $\theta=(\theta_i)_{i\in\mathbb{N}}$ will be a non-decreasing sequence of positive reals satisfying:

$$\theta_n \to +\infty \quad \text{when } n\to+\infty \tag{10}$$

and $\xi=(\xi_i)_{i\in\mathbb{N}}$ will be a sequence of elements of $(\mathbb{R}^+)^N$.

Letting V be a separable metric space, we consider the functions $b_i(1\leqq i\leqq N)$ and f satisfying:

$$\begin{aligned}&\varphi\in C(\mathbb{R}^N x V),\\ &\forall v\in V \quad \varphi(\cdot,v)\in W^{1,\infty}(\mathbb{R}^N) \quad \text{and} \quad \sup_{v\in V}\|\varphi(\cdot,v)\|_{1,\infty}<\infty,\end{aligned} \tag{11}$$

where $\varphi=b_i\ (1\leqq i\leqq N), f$.

Finally, $v(\cdot)$ will be measurable function which takes its value in a compact subset of V. Then the collection $K=(\theta,\xi,v(\cdot))$ will be called the *control*.

Next, we consider a *system* whose *state* is given by the solution $y_x(t)$ of the following problem:

$$\begin{aligned}&\frac{dy_x(t)}{dt}+b(y_x(t),v(t))=0 \quad \text{for } t\in\,]\theta_i,\theta_{i+1}[\quad \text{for all } i\in\mathbb{N},\\ &y_x(0)=x,\\ &y_x(\theta_i+0)=y_x(\theta_i-0)+\xi_i.\end{aligned} \tag{12}$$

The assumptions (11) imply the existence and the uniqueness of a solution $y_x(t)$ of (12).

Then we define the cost function (or pay-off function) for each control K:

$$J(x,K)=\int_0^\infty f(y_x(t),v(t))\,e^{-\lambda t}\,dt+\sum_{i\in\mathbb{N}} c(\xi_i)\,e^{-\lambda\theta_i}, \tag{13}$$

where $\lambda>0$ and c is a continuous function from $(\mathbb{R}^+)^N$ into $\mathbb{R}$ which satisfies

$$\begin{aligned}&c=k+c_0 \quad \text{where } k>0,\\ &c_0(\xi_1+\xi_2)\leqq c_0(\xi_1)+c_0(\xi_2) \quad \text{for all } \xi_1\xi_2\in(\mathbb{R}^+)^N,\\ &c_0(0)=0, \qquad c(\xi_1)\leqq c(\xi_2) \quad \text{if } (\xi_2-\xi_1)\in(\mathbb{R}^+)^N.\end{aligned} \tag{14}$$

The problem to solve is to minimize the cost function over all controls K, that is to find for all $x\in\mathbb{R}^N$:

$$u(x)=\inf_K (J(x,K)). \tag{15}$$

2.2. Dynamic programming principle. The dynamic programming principle is the essential tool to obtain the solution of our problem. We shall give two forms of this

result; the first shows that u is the solution of an optimal stopping time problem, the second being more classical.

Remark 2.1. In the introduction of this part, we took $\theta_0 \leqq \theta_1 \cdots \leqq \theta_n \leqq \cdots$; in fact the assumption (14) implies that we may assume without loss of generality: $\theta_0 < \theta_1 \cdots < \theta_n < \cdots$ in (15).

THEOREM 2.1. *Under assumptions* (11), (14), *we have*

$$u(x) = \inf_{(v(\cdot),\theta_0)} \left(\int_0^{\theta_0} f(y_x(t), v(t))\, e^{-\lambda t}\, dt + Mu(y_x(\theta_0 - 0))\, e^{-\lambda\theta_0} \right). \tag{16}$$

Remark 2.2. The formula (16) means that u is the solution of an optimal stopping time problem as should be expected (for optimal stopping time problems, see for instance A. Bensoussan and J. L. Lions [3]). Remark also that if we take $\theta_0 = 0$, we have: $u(x) \leqq Mu(x)$.

THEOREM 2.2. *Under assumptions* (11), (14) *we have*

$$u(x) = \inf_{K} \left(\int_0^{T} f(y_x(t), v(t))\, e^{-\lambda t}\, dt + \sum_{\theta_i < T} c(\xi_i)\, e^{-\lambda\theta_i} + e^{-\lambda T} u(y_x(T-0)) \right). \tag{17}$$

Remark 2.3. These two results are easy adaptations of classical results and we just recall briefly the ideas of their proofs (cf. also [10]).

Proof of Theorem 2.1. We call $\tilde{u}(x)$ the right-hand side of (16).

Step 1. $u(x) \leqq \tilde{u}(x)$.

$$\forall K \quad u(x) \leqq \int_0^{\infty} f(y_x(t), v(t))\, e^{-\lambda t}\, dt + \sum_{i \in \mathbb{N}} c(\xi_i)\, e^{-\lambda\theta_i}.$$

Then

$$u(x) \leqq \int_0^{\theta_0} f(y_x(t), v(t))\, e^{-\lambda t}\, dt + e^{-\lambda\theta_0} \left[\int_0^{\infty} f(y_x(t+\theta_0), v(t+\theta_0))\, e^{-\lambda t}\, dt + \sum_{i \geqq 1} c(\xi_i)\, e^{-\lambda(\theta_i - \theta_0)} + c(\xi_0) \right].$$

Taking the infimum in the bracket, we obtain easily

$$u(x) \leqq \int_0^{\theta_0} f(y_x(t), v(t))\, e^{-\lambda t}\, dt + e^{-\lambda\theta_0} [u(y_x(\theta_0 - 0) + \xi_0) + c(\xi_0)].$$

As the left-hand side does not depend on K, we conclude easily by taking the infimum in the right-hand side.

Step 2. $\tilde{u}(x) \leqq u(x)$. Let $\varepsilon > 0$; we choose K such that $u(x) + \varepsilon \geqq J(x, K)$, where $K = (\theta, \xi, v)$. The same computation yields

$$u(x) + \varepsilon \geqq \int_0^{\theta_0} f(y_x(t), v(t))\, e^{-\lambda t}\, dt + e^{-\lambda_0} [u(y_x(\theta_0 - 0) + \xi_0) + c(\xi_0)].$$

Then

$$u(x) + \varepsilon \geqq \tilde{u}(x).$$

Since ε is arbitrary, we conclude easily.

Proof of Theorem 2.2. The proof is essentially the same as the proof of Theorem 2.1. We just take T instead of θ_0. In fact, we obtain

$$u(x) = \inf_K \left(\int_0^T f(y_x(t), v(t))\, e^{-\lambda t}\, dt + \sum_{\theta_i < T} c(\xi_i)\, e^{-\lambda \theta_i} + e^{-\lambda T} \min\left(u(y_x(T-0)), Mu(y_x(T-0))\right)\right).$$

The result is then given by the inequality $u(x) \leqq Mu(x)$ in $\mathbb{R}^N$.

2.3. Properties of u. First, we give a result concerning the regularity of u: let

$$\lambda_0 = \sup_{\substack{x \neq x' \\ v \in V}} \left(-\frac{(x-x' | b(x, v) - b(x', v))}{|x-x'|^2} \right).$$

PROPOSITION 2.1. *Under assumptions* (11), (14), *we have*: $u \in \mathrm{BUC}(\mathbb{R}^N)$. *More precisely, we have*:

- *if* $0 < \lambda < \lambda_0$, $u \in C^{0,\delta}(\mathbb{R}^N)$ *with* $\delta = \lambda/\lambda_0$;
- *if* $\lambda = \lambda_0$, $u \in C^{0,\delta}(\mathbb{R}^N)$ *for all* $\delta \in [0, 1[$;
- *if* $\lambda > \lambda_0^+$, $u \in W^{1,\infty}(\mathbb{R}^N)$.

Proof of Proposition 2.1. The proof is exactly the same as in the standard deterministic case (see P. L. Lions [10]).

We use (17) to obtain

$$u(x) - u(x') \leqq \sup_K \left(\int_0^T |f(y_x(t), v(t)) - f(y_{x'}(t), v(t))|\, e^{-\lambda t}\, dt + 2\, e^{-\lambda T} \|u\|_{L^\infty(\mathbb{R}^N)} \right).$$

(Recall that $\|u\|_{L^\infty(\mathbb{R}^N)} \leqq (\|f\|/\lambda) L^\infty(\mathbb{R}^N)$.)

For a given control K, we have (see [10]):

$$|y_x(t) - y_{x'}(t)| \leqq e^{\lambda_0 t} |x - x'|.$$

So we use (11) to obtain

$$u(x) - u(x') \leqq \int_0^T C\, e^{(\lambda_0 - \lambda)t} |x - x'|\, dt + 2 \|u\|_{L^\infty(\mathbb{R}^N)}\, e^{-\lambda T}$$

where $C = \sup_{v \in V} \|f(\cdot, v)\|_{1,\infty}$. If $\lambda > \lambda_0^+$, we let $T \to \infty$.

In the other case, we may assume $|x - x'| < 1$ and we choose T such that $e^{-\lambda_0 T} = |x - x'|$. Then one concludes easily.

Now, we can show the relation between the deterministic impulsive control and the first-order Hamilton-Jacobi quasi-variational inequalities.

THEOREM 2.3. *Under assumptions* (11), (14) *and if u is differentiable at x, we have*

$$\max\left(\sup_{v \in V} (b(x, v) \cdot Du(x) + \lambda u - f(x, v)),\, u(x) - Mu(x) \right) = 0. \tag{18}$$

Remark 2.4. We shall use the notation

$$H(x, t, p) = \sup_{v \in V} (b(x, v)p + \lambda t - f(x, v)). \tag{19}$$

By the well-known Rademacher theorem, we have the following corollary:

COROLLARY 2.1. *Under assumptions* (11), (14) *and if* $\lambda > \lambda_0^+$, *then* $u \in W^{1,\infty}(\mathbb{R}^N)$ *and satisfies*

$$\max\left(\sup_{v\in V}(b(x,v)Du+\lambda u-f(x,v)),\ u-Mu\right)=0 \quad \text{a.e. in } \mathbb{R}^N. \tag{20}$$

Proof of Theorem 2.3. It is again a more or less standard proof (see P. L. Lions [10]).

Step 1. $\max(H(x,u(x),Du(x)),u(x)-Mu(x))\leqq 0$. We take the particular control $v(t)\equiv v\in V$ and $\theta_0=+\infty$. For all T, we have by the dynamic programming principle

$$u(x)\leqq\int_0^T f(y_x(t),v(t))\,e^{-\lambda t}\,dt+e^{-\lambda T}u(y_x(T)).$$

Dividing by T and letting $T\to 0$, we obtain the result as in [10]:

$$H(x,u(x),Du(x))\leqq 0.$$

As we know that $u(x)\leqq Mu(x)$, the proof of the first step is complete.

Step 2. $\max(H(x,u(x),Du(x)),u(x)-Mu(x))\geqq 0$. If $u(x)=Mu(x)$, we have nothing to show.

If $u(x)<Mu(x)$, we must prove that $H(x,u(x),Du(x))\geqq 0$. We need the following lemma:

LEMMA 2.1. *Let* $x\in\mathbb{R}^N$ *such that* $u(x)<Mu(x)$. *Then there exists* $\varepsilon>0$ *such that*

$$u(x)=\inf_{\substack{K\\ \theta_0\geqq\varepsilon}}\left(\int_0^\infty f(y_x(t),v(t))\,e^{-\lambda t}\,dt+\sum_{i\in\mathbb{N}}c(\xi_i)\,e^{-\lambda\theta_i}\right).$$

Proof of Lemma 2.1. Let $K^n=(\theta^n,\xi^n,v^n)$ a sequence such that

$$J(x,K^n)\to u(x).$$

We denote by $\chi(n)=J(x,K^n)-u(x)$,

$$u(x)+\chi(n)=\int_0^\infty f(y_x^n(t),v^n(t))\,e^{-\lambda t}\,dt+\sum_{i\in\mathbb{N}}c(\xi_i^n)\,e^{-\lambda\theta_i^n}.$$

By the same computation as in the proof of Theorem 2.1, we obtain

$$u(x)+\chi(n)\geqq\int_0^{\theta_0^n}f(y_x^n(t),v^n(t))\,e^{-\lambda t}\,dt+e^{-\lambda\theta_0^n}Mu(y_x^n(\xi_0^n-0)).$$

Since $|y_x^n(\theta_0^n-0)-x|\leqq\|b\|_{L^\infty(\mathbb{R}^N\times V)}\theta_0^n$, if there exists a subsequence $\theta_0^{n'}$ which converges to 0, we deduce easily that

$$u(x)\geqq Mu(x).$$

(Recall that $Mu\in\mathrm{BUC}(\mathbb{R}^N)$ because $u\in\mathrm{BUC}(\mathbb{R}^N)$, so Mu is continuous at x.) So we obtain a contradiction which proves the lemma.

Using Lemma 2.1 and (17) with $T<\varepsilon$, we obtain:

$$\forall T<\varepsilon\quad u(x)=\inf_{v(\cdot)}\left(\int_0^T f(y_x(t),v(t))\,e^{-\lambda t}\,dt+e^{-\lambda T}u(y_X(T))\right)$$

y_x is defined by (12) with $\theta_0\geqq\varepsilon$.

We conclude as in the standard case of continuous control only (see [10]).

THEOREM 2.4. *Under assumptions* (11), (14) *and* $\lambda > 0$, $u \in C^{0,\alpha}(\mathbb{R}^N) \cap L^\infty(\mathbb{R}^N)$ *for some* $\alpha \in]0, 1]$ *and satisfies*:

$$\begin{aligned} &u(x) \leqq Mu(x), \\ &\forall v \in V \quad b(x, v)Du + \lambda u - f(x, v) \leqq 0 \quad \text{in } D'(\mathbb{R}^N). \end{aligned} \tag{21}$$

In addition, if $w \in \mathrm{BUC}(\mathbb{R}^N)$ *and satisfies*:

$$\begin{aligned} &w \leqq Mw, \\ &\forall v \in V \quad b(x, v)Dw + \lambda w - f(x, v) \leqq 0 \quad \text{in } D'(\mathbb{R}^N), \end{aligned} \tag{22}$$

then $w \leqq u$.

Remark 2.5. The last property means that u is the maximum subsolution of (P) in $D'(\mathbb{R}^N)$ (for related results see P. L. Lions [10], R. Gonzalez and E. Rofman [8] or P. L. Lions and J. L. Menaldi [11]).

Proof of Theorem 2.4. We choose the particular control $\theta_0 = +\infty$, $v(t) \equiv v \in V$. Then

$$\forall t \geqq 0 \quad \frac{1}{t}(u(x) - u(y_x(t))\, e^{-\lambda t}) \leqq \frac{1}{t}\int_0^t f(y_x(t), v)\, e^{-\lambda t}\, dt.$$

We conclude in the same way as in P. L. Lions [10].

Remark 2.6. For the second part, in order to be clear we use the following remark. We can consider in the definition of u only the case of a finite number of impulses (i.e. a finite number of θ_i or $\theta_{n+1} = +\infty$).

Then

$$u(x) = \inf_{(\theta, \xi, v, n)} \left(\int_0^\infty f(y_x(t), v(t))\, e^{-\lambda t}\, dt + \sum_{i=1}^n c(\xi_i)\, e^{-\lambda \theta_i} \right).$$

LEMMA 2.2. *Under the assumptions of Theorem* 2.4, *if* $w \in L^\infty(\mathbb{R}^N) \cap C^1(\mathbb{R}^N)$ *satisfies*

$$\begin{aligned} &w(x) \leqq Mw(x), \\ &\forall v \in V, \quad b(x, v)Dw(x) + \lambda w(x) \leqq f(x, v) + \delta \quad \text{in } \mathbb{R}^N \end{aligned} \tag{23}$$

for $\delta \geqq 0$, *then* $w \leqq u + \delta/\lambda$.

Proof of Lemma 2.2. Let $\theta = (\theta_1, \cdots, \theta_n)$, $\xi = (\xi_1, \cdots, \xi_n)$, $v(t)$ a control. We assume that $\theta_0 > 0$ Then, since $w \in C^1(\mathbb{R}^N)$:

$$\begin{aligned} \forall i \in \mathbb{N} \quad & w(y_x(\theta_i + 0))\, e^{-\lambda \theta_i} - w(y_x(\theta_{i+1} - 0))\, e^{-\lambda \theta_{i+1}} \\ &= \int_{\theta_i}^{\theta_{i+1}} \{Dw(y_x(s))b(y_x(s), v(s)) + \lambda w(y_x(s))\}\, e^{-\lambda s}\, ds, \end{aligned}$$

$$w(x) - w(y_x(\theta_0 - 0))\, e^{-\lambda \theta_0} = \int_0^{\theta_0} \{Dw(y_x(s))b(y_x(s), v(s)) + \lambda w(y_x(s))\}\, e^{-\lambda s}\, ds.$$

Since $Dw \cdot b + \lambda w \leqq f + \delta$, we obtain

$$\begin{aligned} w(x) - w(y_x(T))\, e^{-\lambda T} \leqq & \int_0^T f(y_x(s), v(s))\, e^{-\lambda s}\, ds \\ & + \sum_{i=1}^n e^{-\lambda \theta_i}(w(y_x(\theta_i - 0)) - w(y_x(\theta_i + 0))) + \int_0^T \delta\, e^{-\lambda s}\, ds. \end{aligned}$$

But $y_x(\theta_i + 0) = y_x(\theta_i - 0) + \xi_i$ and from (23)

$$w(y_x(\theta_i - 0)) - w(y_x(\theta_i + 0)) \leqq c(\xi_i).$$

Therefore, we finally obtain

$$w(x)-w(y_x(T))\,e^{-\lambda T}\leqq\int_0^T f(y_x(s),v(s))\,e^{-\lambda s}\,ds+\sum_{i=1}^{n}c(\xi_i)\,e^{-\lambda\theta_i}+\frac{\delta}{\lambda}.$$

Letting $T\to\infty$, we have the result (because w is bounded).

In order to prove the theorem, it is enough to build $w_\varepsilon\in C^\infty(\mathbb{R}^N)\cap L^\infty(\mathbb{R}^N)$, which converges uniformly to w in $\mathbb{R}^N$ and such that:

$$\begin{aligned}&\forall v\in V\quad b(x,v)Dw_\varepsilon(x)+\lambda w_\varepsilon(x)\leqq f(x,v)+\delta(\varepsilon),\\&w_\varepsilon(x)\leqq Mw_\varepsilon(x)\end{aligned}\tag{24}$$

where $\delta(\varepsilon)\to 0$, as $\varepsilon\to 0^+$.

Let $\rho\in D^+(\mathbb{R}^N)$, $\operatorname{supp}\rho\subset B_1$, $\int_{\mathbb{R}^N}\rho(x)\,dx=1$ and let ρ_ε defined by:

$$\rho_\varepsilon(x)=\frac{1}{\varepsilon^N}\rho\left(\frac{x}{\varepsilon}\right).$$

It is well known that $w_\varepsilon=w*\rho_\varepsilon=\int_{\mathbb{R}^N}w(y)\rho_\varepsilon(x-y)\,dy\in C^\infty(\mathbb{R}^N)\cap L^\infty(\mathbb{R}^N)$, converges uniformly to w (actually $\|w_\varepsilon-w\|_{L^\infty(\mathbb{R}^N)}\leqq\sup_{|h|\leqq\varepsilon}\|w(x+h)-w(x)\|_{L^\infty(\mathbb{R}^N)}$). Moreover w_ε satisfies the first inequality in (24) (see P. L. Lions [10]).

To conclude, we just have to prove $w_\varepsilon\leqq Mw_\varepsilon$. Since $w(x)\leqq k+c_0(\xi)+w(x+\xi)$ and $\rho\geqq 0$:

$$w_\varepsilon(x)\leqq k+c_0(\xi)+w_\varepsilon(x+\xi).$$

Then, $w_\varepsilon\leqq Mw_\varepsilon$ and the proof is complete.

Let us finally give a result concerning the behavior of u at infinity.

PROPOSITION 2.2. *Under assumptions* (11), (14) *and if we assume*

$$c(\xi)\to+\infty\quad \textit{if } |\xi|\to+\infty\tag{25}$$

and

$$f(x,v)\to l\quad \textit{if } |x|\to+\infty \textit{ uniformly with respect to } v,\tag{26}$$

then

$$u(x)\to l/\lambda\quad \textit{if } |x|\to+\infty.\tag{27}$$

Remark 2.7. The boundedness of u implies that in the definition of u we can impose the following additional condition on the control:

$$\forall T>0\quad \sum_{\theta_i\leqq T}c(\xi_i)\leqq C\,e^{+\lambda T}.$$

But from assumption (14), we have

$$c\left(\sum_{\theta_i\leqq T}\xi_i\right)\leqq\sum_{\theta_i\leqq T}c(\xi_i)\leqq C\,e^{+\lambda T}.$$

Then, from (25), we obtain

$$\left|\sum_{\theta_i\leqq T}\xi_i\right|\leqq C(T).\tag{28}$$

Proof of Proposition 2.2. Let $\varepsilon>0$. We want to show that for $|x|$ large enough, we have: $|u(x)-l/\lambda|\leqq\varepsilon$.

Since f is bounded, we can choose T independent of x and of all controls (θ, ξ, v) such that

$$\left| \int_T^\infty f(y_x(s); v(s))\, e^{-\lambda s}\, ds \right| \leqq \frac{\varepsilon}{4}. \tag{29}$$

Now we fix T with the property (29) and such that

$$\left| \int_T^\infty l \cdot e^{-\lambda s}\, ds \right| \leqq \frac{\varepsilon}{4}. \tag{30}$$

We use the following inequalities:

$$\inf_{K_1}\left(\int_0^\infty f(y_x(t), v(t))\, e^{-\lambda t}\, dt \right) \leqq u(x) \leqq \inf_{K_2}\left(\int_0^\infty f(y_x(t), v(t))\, e^{-\lambda t}\, dt \right),$$

where $K_1 = \{K = (v, \theta, \xi) \text{ such that } \xi \text{ satisfies (28)}\}$ and $K_2 = \{K = (v, \theta, \xi)/\theta_0 = +\infty\}$. Then, from (29) and (30),

$$\begin{aligned} &-\frac{\varepsilon}{2} + \inf_{K_1}\left(\int_0^\infty [f(y_x(t), v(t)) - l]\, e^{-\lambda t}\, dt \right) \\ &\leqq u(x) - \frac{l}{\lambda} \leqq \frac{\varepsilon}{2} + \inf_{K_2}\left(\int_0^\infty [f(y_x(t), v(t)) - l]\, e^{-\lambda t}\, dt \right). \end{aligned} \tag{31}$$

Considering Remark 2.7, we have for all $K \in K_1$ and for all $x \in \mathbb{R}^N$:

$$0 \leqq t \leqq T, \qquad |y_x(t) - x| \leqq \|b\|_\infty T + C(T),$$

and then

$$|y_x(t)| \geqq |x| - \|b\|_\infty T - C(T).$$

So, for $|x|$ large enough, we have $|f(y_x(t), v(t)) - l| \leqq \varepsilon\lambda/2$ for $t \in [0, T]$. Finally:

$$\int_0^T |f(y_x(t), v(t)) - l|\, e^{-\lambda t}\, dt \leqq \frac{\varepsilon}{2} \quad \text{for all } K \in K_1 \text{ and } K \in K_2; \tag{32}$$

(31) and (32) give the result.

Remark 2.8. The result of Proposition 2.2 is false without (25). Take in $\mathbb{R}$: $b \equiv 0$, $0 < \lambda < 1$, $c \equiv k < 1$; $f = 1$ if $|x| \geqq 2$, 0 if $|x| \leqq 1$ and $0 \leqq f \leqq 1$ if $|x| \in [1, 2]$. Then $u \to k$ if $x \to -\infty$.

2.4. The viscosity formulation of the dynamic programming principle.

THEOREM 2.5. *Let u be the optimal cost function defined by* (15). *Under assumptions* (11), (14), *then $u \in \mathrm{BUC}(\mathbb{R}^N)$ is a viscosity solution of*

$$\max(H(x, u, Du), u - Mu) = 0 \quad \text{in } \mathbb{R}^N. \tag{33}$$

Proof of Theorem 2.5. The proof is inspired from the corresponding one in P. L. Lions [10].

Let $\phi \in C^1(\mathbb{R}^N)$ and x_0 a local maximum point of $u - \phi$. We fix a control K such that $v(t) \equiv v \in V$ and $\theta_0 = +\infty$. Then for all $T > 0$ we have

$$u(x_0) \leqq \int_0^T f(y_{x_0}(t), v)\, e^{-\lambda t}\, dt + u(y_{x_0}(T))\, e^{-\lambda T}.$$

If T is small enough we have $u(y_{x_0}(T)) \leqq \phi(y_{x_0}(T)) + (u(x_0) - \phi(x_0))$ because

$|y_{x_0}(t)-x_0| \leqq Ct$. So we obtain easily

$$u(x_0)\frac{(1-e^{-\lambda T})}{T} \leqq \frac{1}{T}\int_0^T f(y_{x_0}(t), v)\, e^{-\lambda t}\, dt + e^{-\lambda T}\frac{[\phi(y_{x_0}(T))-\phi(x_0)]}{T}.$$

Letting $T \to 0$, we obtain

$$\forall v \in V \quad b(x_0, v)D\phi(x_0) + \lambda u(x_0) - f(x_0, v) \leqq 0,$$

which ends the first part of the proof.

Let $\phi \in C^1(\mathbb{R}^N)$ and x_0 a local minimum of $u-\phi$; then two cases are possible.

Case 1. $u(x_0) = Mu(x_0)$ and there is nothing to prove.

Case 2. $u(x_0) < Mu(x_0)$. We use Lemma 2.1 to claim, for $0 < T < \varepsilon$,

$$u(x_0) = \inf_{\substack{v(\cdot) \\ \theta_0 > T}} \left(\int_0^T f(y_x(t), v(t))\, e^{-\lambda t}\, dt + u(y_{x_0}(T))\, e^{-\lambda T} \right).$$

By assumption for T small enough $u(y_{x_0}(T)) \geqq \phi(y_{x_0}(T)) + (u(x_0) - \phi(x_0))$. So we have

$$u(x_0)\left(\frac{1-e^{-\lambda T}}{T}\right) \geqq \inf_{v(\cdot)} \left(\frac{1}{T}\int_0^T f(y_{x_0}(t), v(t))\, e^{-\lambda t}\, dt + e^{-\lambda T}\frac{[\phi(y_{x_0}(T))-\phi(x_0)]}{T} \right).$$

But

$$|f(y_{x_0}(t), v(t)) - f(x_0, v(t))| \leqq C|y_{x_0}(t) - x_0| \leqq \tilde{C}t$$

and

$$\phi(y_{x_0}(T)) - \phi(x_0) = \int_0^T - b(y_{x_0}(s), v(s))D\phi(y_x(s))\, ds,$$

$$|b(y_{x_0}(s), v(s))D\phi(y_x(s)) - b(x_0, v(s))D\phi(x_0)| = C(s),$$

where $C(s) \to 0$ if $s \to 0^+$, using the fact that b is Lipschitz and $D\phi$ continuous at x_0.

We deduce easily:

$$u(x_0)\left(\frac{1-e^{-\lambda T}}{T}\right) \geqq \inf_{v(\cdot)} \left(\frac{1}{T}\int_0^T (f(x_0, v(t)) - b(x_0, v(t))D\phi(x_0))\, dt - \varepsilon(T) \right),$$

where $\varepsilon(T) \to 0$ when $T \to 0$.

Now we remark that:

$$\frac{1}{T}\int_0^T (f(x_0, v(t)) - b(x_0, v(t))D\phi(x_0))\, dt \geqq \inf_{v \in V} (f(x_0, v) - b(x_0, v)D\phi(x_0)).$$

Letting $T \to 0$, we obtain the result

$$\sup\, (b(x_0, v)D\phi(x_0) + \lambda u(x_0) - f(x_0, v)) \geqq 0.$$

Remark 2.9. We have shown that u is a viscosity solution of the problem (P). In fact, this result will be completely satisfactory if we prove a uniqueness result for viscosity solutions of (33). This is the goal of the third part.

3. A uniqueness result for viscosity solutions of quasi-variational inequality (33).

THEOREM 3.1. *Under assumptions* (11), (14) *and* $\lambda > 0$, *there exists a unique viscosity solution of the problem* (33) *in* BUC $(\mathbb{R}^N)$.

Remark 3.1. The idea of the proof consists in adapting the method due to B. Hanouzet and J. L. Joly [9] for elliptic Q.V.I. (see also A. Bensoussan [2] or A. Bensoussan and J. L. Lions [4]).

Remark 3.2. The proof uses in an essential way results concerning optimal stopping time problems (in the deterministic case) that we give briefly in the following proposition. For more details and complete proof of this result, see P. L. Lions [10], A. Bensoussan and J. L. Lions [3] or the part II which gives all the ideas necessary to the proof.

Let us define the optimal stopping time problem

$$\frac{dy_x}{dt}(t) + b(y_x(t), v(t)) = 0 \quad \text{for all } t \geqq 0, \tag{34}$$

$$y_x(0) = x,$$

$$J(x, v, \theta) = \int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + \psi(y_x(\theta))\, e^{-\lambda\theta}, \tag{35}$$

$$u(x) = \inf_{(v,\theta)} J(x, v, \theta). \tag{36}$$

Now we have the following result:

PROPOSITION 3.1. *Under assumptions* (11) *and* $\lambda > 0$, *and if* $\psi \in \mathrm{BUC}(\mathbb{R}^N)$, *then* u *given by* (36) *is the unique viscosity solution in* $\mathrm{BUC}(\mathbb{R}^N)$ *of*

$$\max(H(x, u, Du), u - \psi) = 0. \tag{37}$$

In addition, if $\psi \in W^{1,\infty}(\mathbb{R}^N)$ *and* $\lambda > \lambda_0^+$, *then* $u \in W^{1,\infty}(\mathbb{R}^N)$.

Proof of Theorem 3.1. Without loss of generality, we can assume that $f \geqq 0$ (if this is not the case, we add constants to u and f).

We define the operator T, for $w \in \mathrm{BUC}(\mathbb{R}^N)$, by:

$$Tw(x) = \inf_{(v,\theta)} \left(\int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + Mw(y_x(\theta))\, e^{-\lambda\theta} \right), \tag{38}$$

where y_x is defined by (34).

We need the two following lemmas:

LEMMA 3.1. T *maps* $\mathrm{BUC}(\mathbb{R}^N)$ *into* $\mathrm{BUC}(\mathbb{R}^N)$, *is increasing and concave. Furthermore,* Tw *is the unique viscosity solution in* $\mathrm{BUC}(\mathbb{R}^N)$ *of*

$$\max(H(x, z, Dz), z - Mw) = 0 \quad \text{in } \mathbb{R}^N. \tag{39}$$

LEMMA 3.2. *Let* u_0 *defined by*

$$u_0(x) = \inf_{v(\cdot)} \left(\int_0^\infty f(y_x(t), v(t))\, e^{-\lambda t}\, dt \right),$$

with y_x *defined by* (34).

Let $\mu > 0$ *be such that* $\mu \|u_0\|_{L^\infty(\mathbb{R}^N)} < k$ *and* $\mu < 1$. *Let* z *and* $\tilde{z}$ *two positive functions of* $\mathrm{BUC}(\mathbb{R}^N)$ *satisfying*

$$z - \tilde{z} \leqq \gamma z \quad \text{for one } \gamma \in [0, 1]. \tag{40}$$

Then

$$Tz - T\tilde{z} \leqq \gamma(1 - \mu) Tz. \tag{41}$$

Remark 3.3. The formulation of Lemma 3.2 is very much akin to the lemma due to B. Hanouzet and J. L. Joly [9] used in elliptic Q.V.I.

Remark 3.4. Let us just recall that u_0 defined in Lemma 3.2 is the unique viscosity solution in $\mathrm{BUC}(\mathbb{R}^N)$ of: $H(x, u, Du) = 0$ in $\mathbb{R}^N$ (cf. P. L. Lions [10]).

Moreover, by uniqueness result in $\mathbb{R}^N$ (see M. G. Crandall and P. L. Lions [6] or P. L. Lions [10]) or directly, we have $Tw \leqq u_0$ for all $w \in \mathrm{BUC}(\mathbb{R}^N)$. (This inequality can be obtained by (38).)

We first show the theorem using the two lemmas.

If we consider two viscosity solutions u and v of (33), we can assume that they are positive: it suffices to add $\max(\|u\|_\infty, \|v\|_\infty)$ to u and v and so f is changed in $f + \lambda \max(\|u\|_\infty, \|v\|_\infty)$.

Besides, by Lemma 3.1, u and v are fixed points of T. (It is a consequence of the uniqueness for the obstacle problem in $\mathrm{BUC}(\mathbb{R}^N)$ for Mu and Mv).

To conclude, we just have to use the Lemma 3.2: since $v \geqq 0$, $u - v \leqq u$, then:

$$Tu - Tv \leqq (1-\mu)u,$$

but $u = Tu$ and $v = Tv$; we obtain by induction

$$\forall n \in \mathbb{N} \quad u - v \leqq (1-\mu)^n u.$$

Since u is bounded, letting $n \to \infty$, we have $u \leqq v$. Changing u and v, we have the result.

Next, we prove the two lemmas.

Proof of Lemma 3.1. T maps $\mathrm{BUC}(\mathbb{R}^N)$ in $\mathrm{BUC}(\mathbb{R}^N)$ is an easy consequence of the Proposition 3.1 because M maps $\mathrm{BUC}(\mathbb{R}^N)$ in $\mathrm{BUC}(\mathbb{R}^n)$. The fact that T is increasing is obvious. Let us prove the concavity.

Let w_1 and $w_2 \in \mathrm{BUC}(\mathbb{R}^N)$ and $\mu \in [0, 1]$.

$$T(\mu w_1 + (1-\mu)w_2)(x) = \inf_{(v(\cdot),\theta)} \left(\int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + e^{-\lambda\theta} M(\mu w_1 + (1-\mu)w_2)(y_x(\theta)) \right).$$

But M is concave:

$$T(\mu w_1 + (1-\mu)w_2)(x) \geqq \inf_{(v(\cdot),\theta)} \left(\int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + e^{-\lambda\theta} M(\mu M w_1(y_x(\theta)) + (1-\mu) M w_2)(y_x(\theta)) \right).$$

We conclude easily using that:

$$\int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + e^{-\lambda\theta}(M w_1(y_x(\theta)) + (1-\mu) M w_2(y_x(\theta)))$$

$$= \mu \left(\int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + e^{-\lambda\theta} M w_1(y_x(\theta)) \right)$$

$$+ (1-\mu) \left(\int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + M w_2(y_x(\theta)) \right).$$

Then

$$T(\mu w_1 + (1-\mu)w_2) \geqq \mu T w_1 + (1-\mu) T w_2, \qquad T \text{ is concave.}$$

Finally, we prove that Tw is the *unique* viscosity solution of (39) in $\mathrm{BUC}(\mathbb{R}^N)$. In fact, it is an easy consequence of uniqueness results for first-order Hamilton-Jacobi equations in $\mathbb{R}^N$. It suffices to consider the Hamiltonian

$$\tilde{H}(x, t, p) = \max(H(x, t, p), t - Mw(x))$$

(see M. G. Crandall and P. L. Lions [6] or P. L. Lions [10]). This ends the proof of Lemma 3.1].

Proof of Lemma 3.2. According to the monotonicity of T, (38) implies that

$$T\tilde{z} \geqq T((1-\gamma)z).$$

Using the concavity gives $T\tilde{z} \geqq (1-\gamma)Tz + \gamma T(0)$.

We first prove that $\mu u_0 \leqq T(0)$:

$$T(0)(x) = \inf_{(v(\cdot),\theta)} \left(\int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + k\, e^{-\lambda\theta} \right).$$

Now we use that $k \geqq \mu u_0(y_x(\theta))$ and that $\mu f \leqq f$ $(f \geqq 0)$:

$$T(0)(x) \geqq \inf_{(v(\cdot),\theta)} \left(\int_0^\theta \mu f(y_x(t), v(t))\, e^{-\lambda t}\, dt + \mu u_0(y_x(\theta))\, e^{-\lambda\theta} \right).$$

But, for all $\theta > 0$, using the dynamic programming principle for standard deterministic control problems, we have

$$\mu \left[\int_0^\theta f(y_x(t), v(t))\, e^{-\lambda t}\, dt + u_0(y_x(\theta))\, e^{-\theta} \right] \geqq \mu u_0(x).$$

This last inequality gives the result we need.

Now using Remark 3.4 gives $u_0 \geqq Tz$. Then

$$T\tilde{z} \geqq (1-\gamma)Tz + \gamma\mu Tz,$$

and we easily deduce the result:

$$Tz - T\tilde{z} \leqq \gamma(1-\mu)Tz.$$

REFERENCES

[1] G. BARLES, *Inéquations quasi-variationnelles du premier ordre et équations de Hamilton-Jacobi*, Comptes-Rendus Paris (1983); detailed paper to appear.

[2] A. BENSOUSSAN, *Stochastic Control by the Functional Analysis Method*, North-Holland, Amsterdam, 1982.

[3] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.

[4] ———, *Inéquations variationnelles et quasi-variationnelles en contrôle stochastique en contrôle impulsionnel*, Dunod, Paris, 1982.

[5] M. G. CRANDALL, L. C. EVANS AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc. (1983), to appear.

[6] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc. (1983), to appear.

[7] E. DE GIORGI, A. MARINO AND M. TOSQUES, *Problemi di evoluzione in spazi metrici e curve di massima pendenza*, *Rend.* Classe Fis. Math. Nat. Acad. Naz. Lincei. Ser. 8, 68 (1980), pp. 180-187.

[8] R. GONZALEZ AND E. ROFMAN, *An algorithm to obtain the maximum solution of Hamilton-Jacobi equations*, in Optimization Techniques, Stoes, ed., Springer, Berlin, 1978.

[9] B. HANOUZET AND J. L. JOLY, *Convergence uniforme des itérées définissant la solution d'une inéquation quasi-variationnelle abstraite*, Comptes-Rendus, Paris, 286, série A (1978), p. 735.

[10] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, London, 1982.

[11] P. L. LIONS AND J. L. MENALDI, *Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations*, I, II, this Journal, 20 (1982), pp. 58-81, 82-95.

[12] A. MARINO AND M. TOSQUES, *Curves of maximal slope for a certain class of nonregular functions*, preprint.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

006

CONNECTIONS BETWEEN OPTIMAL STOPPING AND SINGULAR STOCHASTIC CONTROL II. REFLECTED FOLLOWER PROBLEMS*

IOANNIS KARATZAS† AND STEVEN E. SHREVE‡

Abstract. The stochastic control problem of following a Brownian path by a process of bounded variation, and subject to a reflecting barrier at the origin (reflected follower), is reduced to a question of optimal stopping with absorption. Direct probabilistic arguments are used to establish the equivalence of the two problems under suitable conditions on the cost functions.

Key words. Brownian motion, optimal stopping, stochastic control, reflection of discontinuous trajectories

1. Introduction. This paper is a sequel to [14] and is concerned with the equivalence between a problem of stochastic control and a related question of optimal stopping for Brownian motion. The control problem has the state process

$$X_t = x + W_t + \xi_t + K_t, \qquad 0 \leqq t \leqq \tau, \tag{1.1}$$

where x is a nonnegative number, $W = \{W_t;\ t \geqq 0\}$ is a Brownian motion and $\xi = \{\xi_t;\ t \geqq 0\}$ is a left-continuous process of bounded variation with $\xi_0 = 0$, expressible as the difference of two nondecreasing processes $\xi^{\pm}(\xi_t = \xi_t^+ - \xi_t^-,\ t \geqq 0)$ and with total variation $\check{\xi}_t = \xi_t^+ + \xi_t^-,\ t \geqq 0$. Given W and ξ, the nondecreasing, left-continuous process $K = \{K_t;\ t \geqq 0\}$ with $K_0 = 0$ is constructed in such a way as to keep the state X nonnegative, i.e.,

$$X_t \geqq 0 \quad \forall 0 \leqq t \leqq \tau. \tag{1.2}$$

The "control process" ξ is then to be chosen as to minimize the expected cost

$$E\left[\int_0^{\tau} h(t, X_t)\, dt + \int_{[0,\tau)} f(t)\, d\check{\xi}_t + g(X_\tau)\right],$$

where $h(t, \cdot)$, $g(\cdot)$ are even, convex functions. We call this problem the *Reflected Follower Stochastic Control Problem* and denote its value function by $V(\tau, x)$. The term "bounded variation follower" is reserved for a situation where the constraint (1.2) is absent (or, equivalently, there is no reflecting process K in the state equation (1.1)); the Monotone Follower Problem of [14] is obtained from the Bounded Variation Follower Problem by setting $\xi^+ \equiv 0$.

Special cases of these problems were studied in [2], [3], [5], [9], [10], [12], [13], [16]. In these works the optimal control process ξ^* turns out to be *singular* (as a function of time, with respect to Lebesgue measure), and can be characterized in terms of two regions in (t, x)-space: an open domain $\mathcal{O}$ of inaction and its complement $\mathcal{O}^c$, the domain of action. If the time-state pair (t, X_t) is in $\mathcal{O}^c$, ξ^* causes the state to jump immediately to the nearest point on the boundary $\partial\mathcal{O}$ demarcating the two regions; it acts thereafter only when the pair (t, X_t) is on $\partial\mathcal{O}$, and pushes only as much as necessary

* Received by the editors August 9, 1983.

† Department of Mathematical Statistics, Columbia University, New York, New York 10027. The research of this author was supported in part by the National Science Foundation under grant NSF MCS-81-03435-A01.

‡ Department of Mathematics, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213. The research of this author was supported by the U.S. Air Force under grant AFOSR 82-0259.

to prevent a crossing of $\partial\mathcal{O}$ into the interior of $\mathcal{O}^c$. The optimal process ξ^* can thus be identified as the local time of the optimal state process X^* on $\partial\mathcal{O}$.

Using the cost functions of the control problem, we now formulate a question of *Optimal Stopping with Absorption* (at the origin) for the Brownian motion W. For any $x \geqq 0$, we consider the first passage time

$$S = S(x) = \inf\{t \geqq 0; x + W_t = 0\}.$$

The question is then to choose a stopping time $\sigma \leqq \tau$ in such a way as to minimize the risk

$$E\left[\int_0^{S\wedge\sigma} h_x(t, x+W_t)\, dt + f(\sigma)1_{\{\sigma<S\wedge\tau\}} + g'(x+W_\tau)1_{\{\sigma=\tau<S\}}\right].$$

Let us denote by $u(\tau, x)$ the optimal risk for this problem. The reader is referred to the works [1], [5], [7], [8], [15], [17] for treatments of the question of optimal stopping for continuous-time processes.

Bather and Chernoff [2] were the first to notice the connection between the problems of Bounded Variation Follower Control and Optimal Stopping with Absorption. These authors posed a specific control problem as above, with $h(t, x) \equiv 0$, $f(t) = 1/(\tau - t)$, $g(x) = \frac{1}{2}x^2$, and argued on heuristic grounds that for $x > 0$ the gradient $U_x(\tau, x)$ of its value function should satisfy the free-boundary problem which characterizes the optimal risk $u(\tau, x)$ for the associated optimal stopping problem with absorption. Their approach was made completely rigorous in [13] by one of the present authors, who studied the problem with $h(t, x) = h(x)$ even and convex, $f(t) = 1$ and $g(x) \equiv 0$, as well as the discounted and long-run-average-cost-per-unit-time variants of it. The Reflected and Bounded Variation Follower Problems were shown to have the same value functions on $\mathbb{R}^+$, and were related to the appropriate stopping problems; again, though, the treatment was mostly analytical, relying heavily on the properites of solutions to variational inequalities and free boundary problems.

In this article we show by direct probabilistic arguments that, under proper conditions on the cost functions f, g and h, the Reflected Follower Control Problem and the Optimal Stopping Problem with Absorption are *equivalent* in the sense that $V_x(\tau, x) \equiv u(\tau, x)$, and that the region of inaction in the control problem is the optimal continuation region for the stopping problem. For a discussion of the merits and benefits of this equivalence the reader is referred to the introduction of article [14].

The full equivalence between the two problems is obtained here under the assumption that the control problem admits an optimal solution. Unlike our previous work [14] on the Monotone Follower Problem, we do not establish here the existence of an optimal process for the Reflected Follower Problem under additional conditions similar to those employed in [14]. For technical and expository reasons, we restrict attention to control processes ξ whose jumps on $[0, \tau)$ are exhausted by an increasing sequence of stopping times.

2. Summary. The Reflected Follower Problem and the Optimal Stopping Problem with Absorption are formulated in § 4. The relationship that they bear to one another is much more delicate than in the case of the Monotone Follower Problem (cf. [14, § 3]). We proceed by first discerning a class of processes which is complete for the control problem (Proposition 4.1). Working with processes in this class we prove that the upper-right and upper-left derivatives of $V(\tau, x)$, the value function for the control problem, are dominated by $u(\tau, x)$, the optimal risk for the stopping problem (Propositions 4.3 and 4.8). In order to obtain inequalities in the opposite direction, we have

to assume that the control problem admits an optimal process ξ^*; it is then possible to establish $u(\tau, x)$ as a lower bound on the lower-left derivative of $V(\tau, x)$ (Proposition 4.7). With the help of convexity, one then shows that the gradient of $V(\tau, x)$ exists and is equal to $u(\tau, x)$. In addition, a stopping time σ^* is defined in terms of ξ^* as the first time the latter process becomes positive; σ^* can be shown to be optimal for the stopping problem with absorption (Theorem 4.9).

Instrumental in establishing our "comparison" results (Propositions 4.3, 4.7 and 4.8) is the principle of following, at some distance, an optimal or nearly optimal path up to a certain stopping time, and then jumping on it.

Section 3 contains a summary of results by Chaleyat-Maurel, El Karoui and Marchal [4] concerning the reflection of discontinuous trajectories, which are used repeatedly in the sequel.

This article can be read independently of its precursor [14].

3. Reflecting a discontinuous process. Let us suppose that $Y(t):\mathbb{R}^+\to\mathbb{R}$ is a continuous function, with $Y(0)\geqq 0$. The *Reflection Problem* associated with this function is to find a pair (X, A) of continuous functions, such that:

(i) $X(t)=Y(t)+A(t)$ for all $t\geqq 0$;

(ii) $X(t)\geqq 0$ for all $t\geqq 0$; and

(iii) A is a nondecreasing function, such that $A(0)=0$ and

$$\int_0^\infty X(s)\,dA(s)=0.$$

In other words, A creates a nonnegative function by "pushing" Y to the right, but the pushing occurs only at the origin. The unique solution to the above problem is provided by

$$A(t)=\max\left[0,\ \max_{0\leqq s\leqq t}\{-Y(s)\}\right] \tag{3.1}$$

and $X(t)=Y(t)+A(t)$, $t\geqq 0$; see Ikeda and Watanabe [11, p. 120].

The analogous problem with a discontinuous function Y was studied by Chaleyat-Maurel, El Karoui and Marchal in [4]. We review here their basic results which are of particular importance in the present paper. For more information the reader is referred to [4].

We denote by $\mathcal{H}$ the class of left-continuous functions $Y(t):\mathbb{R}^+\to\mathbb{R}$ which admit right-hand limits and satisfy $Y(0)\geqq 0$. For such functions, we define

$$\Delta Y(t)\triangleq Y(t+)-Y(t), \qquad S_Y\triangleq\{t\geqq 0;\ |\Delta Y(t)|>0\}.$$

A nondecreasing function K in $\mathcal{H}$ can be written as

$$K(t)=K_c(t)+\sum_{0\leqq s<t}\Delta K(s),$$

with K_c continuous and nondecreasing. We can now pose the *Discontinuous Reflection Problem* (DRP (Y)) associated with a given function $Y\in\mathcal{H}$: to find a pair of functions (X, K) in $\mathcal{H}$ so that

(i) $X(t)=Y(t)+K(t)$ for all $t\geqq 0$;

(ii) $X(t)\geqq 0$ for all $t\geqq 0$;

(iii) K is nondecreasing, with $K(0)=0$ and

$$\int_0^\infty X(t)\,dK_c(t)=0, \qquad \Delta K(t)=2X(t+) \quad \forall t\in S_K.$$

Example. Consider the function

$$Y(t) = \alpha 1_{[0,T]}(t) - \beta 1_{(T,\infty)}(t), \qquad t \geq 0,$$

with $\alpha, \beta, T > 0$. The only "natural" candidate for the reflected process is

$$X(t) = \alpha 1_{[0,T]}(t) + \beta 1_{(T,\infty)}(t), \qquad t \geq 0.$$

Here we have

$$K(t) = 2\beta 1_{(T,\infty)}(t) = 2A(t), \qquad t \geq 0,$$

where A is given by (3.1). Clearly, $K_c \equiv 0$ and $\Delta K(T) = 2X(T+)$.

THEOREM 3.1 (Chaleyat-Maurel et al. [4]). *For any $Y \in \mathcal{H}$ there exists a unique solution (X, K) to the* DRP(Y). *The following are satisfied, with A as in* (3.1):

$$A(t) \leq K(t) \leq A(t) + \sum_{0 \leq s < t} \Delta A(s), \qquad t \geq 0; \tag{3.2}$$

$$X(t+) = |X(t) + \Delta Y(t)| = |Y(t+) + K(t)|, \qquad t \geq 0; \tag{3.3}$$

$$S_K = \{t \geq 0; X(t) + \Delta Y(t) < 0\}; \tag{3.4}$$

$$K \text{ is continuous if and only if } K \equiv A. \tag{3.5}$$

COROLLARY 3.2. *Suppose that $X(t) + \Delta Y(t) \geq 0$ holds for all $t \geq 0$. Then K is continuous, and equal to A, i.e.,*

$$K(t) \equiv A(t) \triangleq \max\left[0, \sup_{0 \leq s \leq t} (-Y(s))\right].$$

Remark. Speaking intuitively, K is the function that "pushes at the origin" to keep X nonnegative. This push is continuous, and is given by the expression for A in (3.1) as long as the jumps of Y do not cause any crossings of the origin (i.e., as long as $X(t) + \Delta Y(t) \geq 0$). The determination of K when there are jumps across the origin can be quite complex. However, if there is a sequence of points $0 = T_0 \leq T_1 \leq \cdots$ such that Y is continuous on each interval (T_n, T_{n+1}), then X and K can be defined inductively by

$$\Delta K(T_n) = 2 \max[0, -X(T_n) - \Delta Y(T_n)], \tag{3.6}$$

$$K(t) - K(T_n+) = \max\left[0, \sup_{T_n < s \leq t} \{-Y(s) - K(T_n+)\}\right], \qquad T_n < t \leq T_{n+1}. \tag{3.7}$$

4. The Reflected Follower Problem. We consider a probability space $(\Omega, \mathcal{F}, P)$ endowed with an increasing family of σ-fields $\{\mathcal{F}_t; t \geq 0\}$ which satisfy the "usual conditions" of right-continuity and completeness with respect to P. A Brownian motion $W = \{W_t; t \geq 0\}$ on this space is assumed. Let $\tau > 0$ be a fixed time-horizon and let $\mathcal{A}(\tau)$ be the class of all $\{\mathcal{F}_t\}$-adapted processes $\xi = \{\xi_t; t \geq 0\}$ which are a.s. nondecreasing, left-continuous, null at the origin and constant on $[\tau, \infty)$. Let $\tilde{\mathcal{A}}(\tau)$ be the subset of processes ξ in $\mathcal{A}(\tau)$ for which there exists an increasing sequence $\{T_n\}_{n=1}^{\infty}$ of $\{\mathcal{F}_t\}$-stopping times which exhaust the jumps of the process ξ, P-almost surely. In the notation of the previous section,

$$S_{\xi(\omega)} \subseteq \bigcup_{n=1}^{\infty} \{T_n(\omega)\}$$

for P-almost every $\omega \in \Omega$. We denote by $\mathcal{B}(\tau)$ the class of processes ξ (of bounded variation) which admit a minimal decomposition of the form $\xi_t = \xi_t^+ - \xi_t^-$, $t \geq 0$ as the

difference of two processes $\xi^{\pm}$ in $\tilde{\mathscr{A}}(\tau)$. We shall use the symbol

$$\check{\xi}_t = \xi_t^+ + \xi_t^-, \qquad t \geqq 0$$

for the total variation of the process $\xi \in \mathscr{B}(\tau)$ on the interval $[0, t]$. $\mathscr{B}(\tau)$ will assume the role of the class of admissible processes in the Reflected Follower Problem.

For each $x \geqq 0$, $\xi \in \mathscr{B}(\tau)$, we can construct in a pathwise fashion the pair of processes (X, K) which solves uniquely the Discontinuous Reflection Problem DRP (Y), associated with the random function $Y(t) \triangleq x + W_t + \xi_t$, $t \geqq 0$. The basic facts about this construction are outlined in the preceding section; the process K is nondecreasing, left-continuous and null at the origin, while the process X given by

$$X_t = x + W_t + \xi_t + K_t, \qquad t \geqq 0$$

is nonnegative. It will be the state process in the Reflected Follower Problem. It is convenient to have this process defined for all $t \geqq 0$, even though the problem is posed only on the interval $[0, \tau]$, and ξ is taken to be constant on $[\tau, \infty)$. The cost functions for this problem are the following:

(4.1i) a nonnegative, continuous function $f(t)$ on $[0, \tau]$, representing a *running cost of controlling action per unit time*;

(4.1ii) a real-valued, continuous and continuously differentiable function $g(x)$ on $\mathbb{R}^+$, such that $g'(x)$ is nondecreasing and $g'(0) = 0$, representing a *terminal cost on the state*; and

(4.1iii) a real-valued, continuous function $h(t, x)$ on $[0, \tau] \times \mathbb{R}^+$, with continuous gradient $h_x(t, x)$ which is nondecreasing in the space variable x and satisfies $h_x(t, 0) \geqq 0$, representing a *running cost per unit time on the state.*

The functions $h_x(t, x)$ and $g'(x)$ will be assumed to satisfy a polynomial growth condition in the space variable:

(4.1iv) $$0 \leqq h_x(t, x) + g'(x) \leqq C(1 + x^m), \qquad \text{on } [0, \tau] \times \mathbb{R}^+$$

for some $m \geqq 1$, $C > 0$. Conditions (4.1) will be assumed to hold throughout this paper.

To each process $\xi \in \mathscr{B}(\tau)$, we associate the expected total cost

$$J(\xi; \tau, x) = E\left[\int_0^{\tau} h(t, X_t)\, dt + \int_{[0,\tau)} f(t)\, d\check{\xi}_t + g(X_\tau)\right]. \tag{4.2}$$

The *Reflected Follower Problem* is to choose the process $\xi \in \mathscr{B}(\tau)$ so as to minimize the expected total cost in (4.2), i.e., to achieve the infimum

$$V(\tau, x) = \inf_{\xi \in \mathscr{B}(\tau)} J(\xi; \tau, x).$$

Associated with this control problem is an *Optimal Stopping Problem with Absorption (at the Origin)* for the Brownian motion process W. With $x \geqq 0$, we consider the first hitting time $S = S(x)$ of the origin

$$S = S(x) = \begin{cases} \inf\{t \geqq 0;\, x + W_t = 0\}, \\ +\infty, \quad \text{if } \{t \geqq 0;\, x + W_t = 0\} = \varnothing. \end{cases}$$

Corresponding to each $\{\mathscr{F}_t\}$-stopping time σ such that $P(0 \leqq \sigma \leqq \tau) = 1$, we associate the risk

$$R(\sigma; \tau, x) = E\left[\int_0^{S \wedge \sigma} h_x(t, x + W_t)\, dt + f(\sigma) 1_{\{\sigma < S \wedge \tau\}} + g'(x + W_\tau) 1_{\{\sigma = \tau < S\}}\right]. \tag{4.3}$$

In an obvious interpretation, h_x is a cost of continuation, f is a fee for premature termination (i.e. stopping before hitting the origin or running out of time), and g' represents a terminal cost. The problem is to select the stopping time σ so as to minimize the risk in (4.3), or equivalently to achieve the infimum

$$u(\tau, x) = \inf_{0 \leqq \sigma \leqq \tau} R(\sigma; \tau, x). \tag{4.4}$$

Remarks. (i) Although the infimum in (4.4) is over all stopping times σ such that $P(0 \leqq \sigma \leqq \tau) = 1$, for $x > 0$ we need only consider such times which also satisfy

$$P(\sigma = S) = 0.$$

To see this, consider a stopping time σ such that $0 \leqq \sigma \leqq \tau$, a.s. P, and define the new random variable

$$\hat{\sigma} = \begin{cases} \sigma & \text{on } \{\sigma \neq S\}, \\ \tau & \text{on } \{\sigma = S\}. \end{cases}$$

Clearly, $\{\hat{\sigma} \leqq \tau\}$ is an event of probability one, and with $0 \leqq t < \tau$:

$$\{\hat{\sigma} \leqq t\} = \{\sigma \leqq t, \sigma \neq S\}.$$

According to standard theory (e.g. Dellacherie [6, p. 53]) both $\{S \leqq \sigma\}$ and $\{S < \sigma\}$ belong to the σ-field $\mathcal{F}_\sigma$, and therefore $\{S = \sigma\}$ and $\{S \neq \sigma\}$ are also in $\mathcal{F}_\sigma$. By definition of the latter, we conclude that $\{\hat{\sigma} \leqq t\} \in \mathcal{F}_t$; for all $t \geqq 0$, so $\hat{\sigma}$ is an $\{\mathcal{F}_t\}$-stopping time. On the other hand: $\{\hat{\sigma} = S\} \subseteqq \{S = \tau\}$, a P-negligible event, so $P(\hat{\sigma} = S) = 0$, and furthermore

$$R(\hat{\sigma}; \tau, x) = R(\sigma; \tau, x),$$

as one can easily verify. This validates our claim.

(ii) It is quite obvious from the above that $u(\tau, 0) = 0$. Besides, since the functions h_x, f and g' are nonnegative, $h_x(t, \cdot)$ and $g'(\cdot)$ are nondecreasing, and $S(x_2) \geqq S(x_1)$; P-almost surely for $x_2 \geqq x_1 \geqq 0$, it is not hard to see that the function $u(\tau, x)$ is nondecreasing on $\mathbb{R}^+$.

We introduce now a particular subclass of $\tilde{\mathcal{A}}(\tau)$ which will be very useful in our study of the Reflected Follower Problem. For any given $x \geqq 0$, we say that a process $\xi^- \in \tilde{\mathcal{A}}(\tau)$ is in *class* $\mathcal{D}(\tau, x)$ if for the solution (X, K) of the Discontinuous Reflection Problem DRP (Y) corresponding to

$$Y(t) = x + W_t - \xi_t^-, \qquad t \geqq 0,$$

we have

$$X_t = x + W_t - \xi_t^- + K_t, \qquad 0 \leqq \Delta\xi_t^- \leqq X_t, \quad t \geqq 0.$$

According to the results in § 3 (Corollary 3.2), the increasing process K is then continuous and is given by

$$K_t = \max\left[0, \sup_{0 \leqq s \leqq t} \{\xi_s^- - (x + W_s)\}\right], \qquad t \geqq 0.$$

Clearly, for any ξ^- in $\mathcal{D}(\tau, x)$ the process $\xi = -\xi^-$ is admissible for the Reflected Follower Problem, with associated expected cost

$$J(-\xi^-; \tau, x) = E\left[\int_0^\tau h(t, X_t)\, dt + \int_{[0,\tau)} f(t)\, d\xi_t^- + g(X_\tau)\right].$$

Remark. The class $\mathscr{D}(\tau, x)$ is nonempty; indeed, with $b>0$ a given number and $x \geqq 0$, one can construct a unique pair (ξ^-, K) of processes in $\tilde{\mathscr{A}}(\tau)$ such that

$$\xi_t^- = \max\left[0, \sup_{0 \leqq s \leqq t \wedge \tau} \{x + W_s + K_s - b\}\right], \qquad t \geqq 0,$$

$$K_t = \max\left[0, \sup_{0 \leqq s \leqq t} \{\xi_s^- - (x + W_s)\}\right], \quad t \geqq 0.$$

Apart from a possible leftward jump of size $(x-b)^+$, the resulting state process: $X_t = x + W_t - \xi_t^- + K_t$, $t \geqq 0$ is, for $0 \leqq t \leqq \tau$, a Brownian motion reflected at the origin and at the point b, and for $t > \tau$ it is a Brownian motion reflected at the origin only. For the construction of the process on $[0, \tau]$ see [13, § 8]. For $0 \leqq t \leqq \tau$, the processes K_t and ξ_t^- are constant multiples of the local time spent during $[0, t]$ by the state process X at $x = 0$ and $x = b$, respectively. Other examples can also be concocted.

PROPOSITION 4.1. *For any given $\tau > 0$ and $x \geqq 0$, we have*

$$V(\tau, x) = \inf_{\xi^- \in \mathscr{D}(\tau, x)} J(-\xi^-; \tau, x).$$

Moreover, if the Reflected Follower Problem started at $x \geqq 0$ has an optimal process in $\mathscr{B}(\tau)$, then there exists a process ξ_*^- in $\mathscr{D}(\tau, x)$ such that $V(\tau, x) = J(-\xi_*^-; \tau, x)$.

Proof. It suffices to show that, given any process $\eta \in \mathscr{B}(\tau)$, one can construct a process $\xi^- \in \mathscr{D}(\tau, x)$ with the property

$$J(-\xi^-; \tau, x) \leqq J(\eta; \tau, x). \tag{4.5}$$

Let $\{T_n\}_{n=1}^{\infty}$ be an increasing sequence of stopping times which exhausts the jumps of η on $[0, \tau)$ and such that $0 \leqq T_n \leqq \tau$; $\forall n \geqq 1$, $\lim_{n\to\infty} T_n = \tau$ and $T_0 = 0$ hold a.s. P. Let also (X, K) be the solution to the DRP (Y), as in § 3, with $Y(t) = x + W_t + \eta(t)$, $t \geqq 0$. We shall construct a process $\xi^- \in \tilde{\mathscr{A}}(\tau)$ such that, if (Z, L) is the solution to the DRP (Q) with $Q(t) = x + W_t - \xi^-(t)$, $t \geqq 0$, we have

$$0 \leqq \Delta\xi^-(t) \leqq Z(t), \qquad t \geqq 0,$$

i.e. $\xi^- \in \mathscr{D}(\tau, x)$, and (4.5) holds.

The construction proceeds by induction. As the induction hypothesis, we assume ξ^-, Z and L have been constructed on $[0, T_n]$ with (cf. (3.6), (3.7))

$$\begin{aligned} &Z(t) = x + W_t - \xi^-(t) + L(t), \qquad 0 \leqq t \leqq T_n, \\ &\Delta L(T_k) = 2 \max\{0, -Z(T_k) + \Delta\xi^-(T_k)\}, \qquad k = 0, \cdots, n-1, \\ &L(t) - L(T_k+) = \max\left[0, \sup_{T_k < s \leqq t} \{-x - W_s + \xi^-(s) - L(T_k+)\}\right], \\ &\qquad\qquad T_k < t \leqq T_{k+1}, \quad k = 0, \cdots, n-1, \end{aligned} \tag{4.6}$$

such that

$$Z(t) \leqq X(t), \qquad 0 \leqq t \leqq T_n, \tag{4.7}$$

$$\Delta\xi^-(t) \leqq \Delta\eta^-(t), \qquad 0 \leqq t < T_n, \tag{4.8}$$

$$\xi_c^-(t) = \eta_c^-(t), \qquad 0 \leqq t \leqq T_n, \tag{4.9}$$

where the subscript c denotes the continuous parts of the processes. When $n = 1$, this hypothesis is trivially valid. We now extend ξ^- to $[0, T_{n+1}]$ in such a way that (4.7)-(4.9)

are preserved. To do this, we define

$$\begin{aligned} &\Delta\xi^-(T_n) = \min\{Z(T_n), \Delta\eta^-(T_n)\}, \\ &\xi^-(t) - \xi^-(T_n+) = \eta^-(t) - \eta^-(T_n+), \qquad T_n < t \leqq T_{n+1}. \end{aligned} \tag{4.10}$$

It is clear that with ξ^- thus extended, $\Delta\xi^-(t) \leqq \Delta\eta^-(t)$, $0 \leqq t < T_{n+1}$, and $\xi_c^-(t) = \eta_c^-(t)$, $0 \leqq t \leqq T_{n+1}$. Furthermore, $\Delta\xi^-(T_n) \leqq Z(T_n)$, so $\Delta L(T_n) = 0$. According to (4.10), we have either $\Delta\xi^-(T_n) = \Delta\eta^-(T_n)$ or else $Z(T_n+) = 0$. It follows from (4.7) that $Z(T_n+) \leqq X(T_n+)$, or equivalently,

$$\eta(T_n+) + K(T_n+) \geqq -\xi^-(T_n+) + L(T_n+).$$

To extend this inequality for $t \in (T_n, T_{n+1}]$, we use (3.7) and (4.9) to write, for $T_n < t \leqq T_{n+1}$,

$$\begin{aligned} \eta(t) + K(t) &= \eta(t) + \max\left[K(T_n+), \sup_{T_n < s \leqq t}\{-x - W_s - \eta(s)\}\right] \\ &= \max\left[(\eta(t) - \eta(T_n+)) + (\eta(T_n+) + K(T_n+)),\right. \\ &\qquad \left.\sup_{T_n < s \leqq t}\{-x - W_s + (\eta^+(t) - \eta^+(s)) - (\eta^-(t) - \eta^-(s))\}\right] \\ &\geqq \max\left[(\eta^+(t) - \eta^+(T_n+)) - (\eta^-(t) - \eta^-(T_n+)) - \xi^-(T_n+) + L(T_n+),\right. \\ &\qquad \left.\sup_{T_n < s \leqq t}\{-x - W_s - (\xi^-(t) - \xi^-(s))\}\right] \\ &\geqq \max\left[-\xi^-(t) + L(T_n+), \sup_{T_n < s \leqq t}\{-x - W_s - \xi^-(t) + \xi^-(s)\}\right] \\ &= -\xi^-(t) + L(t). \end{aligned}$$

It follows that

$$Z(t) = x + W_t - \xi^-(t) + L(t) \geqq x + W_t + \eta(t) + K(t) = X(t), \qquad 0 \leqq t \leqq T_{n+1}.$$

We thus obtain ξ^- on $[0, \tau)$, and since $\xi^-(t) \leqq \eta^-(t) \leqq \eta^-(\tau) < \infty$ for $0 \leqq t < \tau$, we can define:

$$\xi^-(\tau) = \lim_{t \uparrow \tau} \xi^-(t) \quad \text{and} \quad \xi^-(t) = \xi^-(\tau), \qquad t \geqq \tau.$$

Then $\xi^- \in \mathcal{D}(\tau, x)$, and (4.7)-(4.9) as well as the properties of f, g and h imply (4.5). □

We shall need the following result later in this section:

LEMMA 4.2. *With $0 \leqq x_1 < x_2$, we have $\mathcal{D}(\tau, x_1) \subseteq \mathcal{D}(\tau, x_2)$.*

Proof. Let us suppose that $\xi^- \in \mathcal{D}(\tau, x_1)$, i.e., that the solution (X_1, K_1) to the DRP(Y_1), with $Y_1(t) = x_1 + W_t - \xi_t^-$, $t \geqq 0$ satisfies

$$X_1(t) = x_1 + W_t - \xi_t^- + K_1(t), \qquad 0 \leqq \Delta\xi_t^- \leqq X_1(t), \qquad t \geqq 0.$$

Therefore, by Corollary 3.2:

$$K_1(t) = \max\left[0, \sup_{0 \leqq s \leqq t}\{\xi_s^- - (x_1 + W_s)\}\right], \qquad t \geqq 0.$$

Similarly, with $x_2 > x_1$, we consider the solution (X_2, K_2) to the DRP(Y_2), with $Y_2(t) = x_2 + W_t - \xi_t^-$, $t \geqq 0$. From Theorem 3.1 (relation (3.2)) we have

$$K_2(t) \geqq A_2(t) = \max\left[0, \sup_{0 \leqq s \leqq t}\{\xi_s^- - (x_2 + W_s)\}\right].$$

It can then be checked that $x_2+K_2(t)\geqq x_1+K_1(t)$, $t\geqq 0$, so

$$X_2(t)=x_2+W_t-\xi_t^-+K_2(t)\geqq X_1(t),\qquad t\geqq 0$$

and a fortiori, $0\leqq\Delta\xi_t^-\leqq X_2(t)$, $t\geqq 0$.

Therefore, ξ^- is in class $\mathscr{D}(\tau, x_2)$ for all initial positions $x_2>x_1$. □

In the light of Proposition 4.1, let us consider for the purposes of the ensuing discussion a point $x>0$ and control processes $\xi^-\in\mathscr{D}(\tau, x)$, such that the solution (X, K) to the DRP (Q) with $Q(t)=x+W_t-\xi_t^-$, $t\geqq 0$ satisfies

$$X_t=x+W_t-\xi_t^-+K_t,\qquad 0\leqq\Delta\xi_t^-\leqq X_t,\qquad t\geqq 0. \tag{4.11}$$

Under these circumstances, K is continuous and is given by

$$K_t=\max\left[0,\ \sup_{0\leqq s\leqq t}\{\xi_s^--(x+W_s)\}\right]\quad\text{for } t\geqq 0. \tag{4.11$'$}$$

We consider $\delta>0$, and the solution $(X(\delta), K(\delta))$ to the DRP $(Q(\delta))$, with $Q_t(\delta)=x+\delta+W_t-\xi_t^-$, $t\geqq 0$:

$$X_t(\delta)=x+\delta+W_t-\xi_t^-+K_t(\delta),\qquad t\geqq 0. \tag{4.12}$$

As was shown in Lemma 4.2, ξ^- is in class $\mathscr{D}(\tau, x+\delta)$,

$$0\leqq X_t\leqq X_t(\delta),\qquad t\geqq 0$$

and $K(\delta)$ is continuous:

$$K_t(\delta)=\max\left[0,\ \sup_{0\leqq s\leqq t}\{\xi_s^--(x+\delta+W_s)\}\right],\qquad t\geqq 0. \tag{4.12$'$}$$

We shall be interested in the stopping times

$$\begin{aligned}&T=\inf\{t\geqq 0;\ \xi_t^-(x+W_t)\geqq 0\}=\inf\{t\geqq 0;\ x_t=0\},\\ &T^\delta=\inf\{t\geqq 0;\ X_t\geqq X_t(\delta)\}.\end{aligned} \tag{4.13}$$

First, let us note that for $t\geqq 0$,

$$\begin{aligned}X_t(\delta)-X_t&=\delta+K_t(\delta)-K_t\\ &=\max\left[\delta,\ \sup_{0\leqq s\leqq t}\{\xi_s^-(x+W_s)\}\right]-\max\left[0,\ \sup_{0\leqq s\leqq t}\{\xi_s^-(x+W_s)\}\right]\end{aligned}$$

is a continuous function. It follows that

$$T^\delta=\inf\{t\geqq 0;\ \xi_t^-(x+W_t)\geqq\delta\}\geqq T.$$

Let us also observe that, for $t\geqq T^\delta$, $K_t(\delta)+\delta=K_t$, so

$$X_t=X_t(\delta),\qquad t\geqq T^\delta.$$

If we take $\xi^-\equiv 0$ in the above, we obtain the reflected Brownian motion starting at x (respectively, $x+\delta$):

$$\begin{aligned}&Z_t=x+W_t+L_t,&&L_t=\max\left[0,\ \sup_{0\leqq s\leqq t}\{-(x+W_s)\}\right],\\ &Z_t(\delta)=x+\delta+W_t+L_t(\delta),&&L_t(\delta)=\max\left[0,\ \sup_{0\leqq s\leqq t}\{-(x+\delta+W_s)\}\right],\end{aligned} \tag{4.14}$$

and by analogy with the stopping times T, T^δ we define

$$\begin{aligned}&S\triangleq\inf\{t\geqq 0;\ Z_t=0\}=\inf\{t\geqq 0;\ x+W_t=0\},\\ &S^\delta\triangleq\inf\{t\geqq 0;\ Z_t=Z_t(\delta)\}=\inf\{t\geqq 0;\ x+W_t\leqq-\delta\}.\end{aligned} \tag{4.15}$$

We note the relations $0 \leqq X_t \leqq Z_t$, $0 \leqq X_t(\delta) \leqq Z_t(\delta)$, $\forall t \geqq 0$ and $T^\delta \leqq S^\delta$, $T \leqq S$, which are valid P-almost surely. Since S, S^δ are a.s. finite, so are T, T^δ. Besides, $S^\delta \downarrow S$ as $\delta \downarrow 0$, a.s. P.

As in [14], we use the notation

$$\Delta^{\pm} V(\tau, x) \triangleq \overline{\lim_{\delta \to 0\pm}} \frac{V(\tau, x+\delta) - V(\tau, x)}{\delta}, \qquad \Delta_{\pm} V(\tau, x) \triangleq \underline{\lim_{\delta \to 0\pm}} \frac{V(\tau, x+\delta) - V(\tau, x)}{\delta}$$

for the four derivatives of the function $V(\tau, \cdot)$ at x.

PROPOSITION 4.3. *For all $x \geqq 0$, we have*

$$\Delta^+ V(\tau, x) \leqq u(\tau, x).$$

Proof. It is sufficient to show that

$$\overline{\lim_{\delta \downarrow 0}} \frac{V(\tau, x+\delta) - V(\tau, x)}{\delta} \leqq R(\sigma; \tau, x) \tag{4.16}$$

holds for any stopping time σ satisfying: $0 \leqq \sigma \leqq \tau$ and $\sigma \neq S$, a.s. P. We choose such a σ, as well as a process ξ^- which is in class $\mathcal{D}(\tau, x)$—and a fortiori in class $\mathcal{D}(\tau, x+\delta)$ for any fixed $\delta > 0$, by Lemma 4.2.

This gives a reflected state trajectory X as in (4.11). We create a slightly perturbed trajectory emanating from $x+\delta$, which agrees with that of $X(\delta)$ as in (4.12) up to time σ, then jumps on X and agrees with it thereafter:

$$Y_t(\delta) = \begin{cases} X_t(\delta), & 0 \leqq t \leqq \sigma, \\ X_t, & \sigma < t. \end{cases}$$

It is not hard to see that, with

$$\eta_t^-(\delta) = \begin{cases} \xi_t^-, & 0 \leqq t \leqq \sigma, \\ \xi_t^- + (X_\sigma(\delta) - X_\sigma), & \sigma < t \end{cases} \tag{4.17}$$

and

$$\tilde{K}_t(\delta) = \max\left[0, \sup_{0 \leqq s \leqq t} \{\eta_s^-(\delta) - (x + \delta + W_s)\}\right], \qquad t \geqq 0$$

the pair $(Y(\delta), \tilde{K}(\delta))$ is the unique solution to the DRP(Q), with $Q(t) = x + \delta + W_t - \eta_t^-(\delta)$, $t \geqq 0$,

$$Y_t(\delta) = x + \delta + W_t - \eta_t^-(\delta) + \tilde{K}_t(\delta), \qquad 0 \leqq \Delta\eta_t^-(\delta) \leqq Y_t(\delta), \qquad t \geqq 0,$$

i.e., $\eta^-(\delta)$ is in class $\mathcal{D}(\tau, x+\delta)$. The performance of this process is certainly suboptimal for the Reflected Follower Problem starting at $x+\delta$, so

$$\begin{aligned} V(\tau, x+\delta) &\leqq E\left[\int_0^\tau h(t, Y_t(\delta))\, dt + \int_{[0,\tau)} f(t)\, d\eta_t^-(\delta) + g(Y_\tau(\delta))\right] \\ &\leqq E\left[\int_0^{T^\delta \wedge \sigma} h(t, X_t(\delta))\, dt + \int_{T^\delta \wedge \sigma}^{T} h(t, X_t)\, dt \right. \\ &\quad + \int_{[0,\tau)} f(t)\, d\xi_t^- + f(\sigma)(X_\sigma(\delta) - X_\sigma) 1_{\{\sigma < \tau \wedge T^\delta\}} \\ &\quad \left. + g(X_\tau) 1_{\{\sigma < \tau\}} + g(X_\tau) 1_{\{\sigma = \tau \geqq T^\delta\}} + g(X_\tau(\delta)) 1_{\{\sigma = \tau < T^\delta\}}\right]. \end{aligned}$$

Subtracting the performance $J(-\xi^-; \tau, x)$ of the process $\xi^- \in \mathcal{D}(\tau, x)$ from both members of the above inequality, we obtain

$$
\begin{aligned}
&V(\tau, x+\delta) - J(-\xi^-; \tau, x) \\
&\leqq E\Bigg[\int_0^{T^\delta \wedge \sigma} \{h(t, X_t(\delta)) - h(t, X_t)\}\, dt \\
&\qquad + f(\sigma)(X_\sigma(\delta) - X_\sigma) 1_{\{\sigma < T^\delta \wedge \tau\}} + \{g(X_\tau(\delta)) - g(X_\tau)\} 1_{\{\sigma = \tau < T^\delta\}}\Bigg].
\end{aligned} \tag{4.18}
$$

Let us recall the relations

$$
\begin{aligned}
&X_t(\delta) = \delta + X_t, \quad 0 \leqq t \leqq T, \\
&X_t \leqq X_t(\delta) \leqq \delta + X_t, \quad T < t \leqq T^\delta, \\
&X_t(\delta) = X_t, \quad T^\delta \leqq t,
\end{aligned}
$$

and their corollary

$$
0 \leqq X_\sigma(\delta) - X_\sigma \begin{cases} = \delta & \text{on } \{0 \leqq \sigma \leqq T\}, \\ \leqq \delta & \text{on } \{T < \sigma \leqq T^\delta\}, \\ = 0 & \text{on } \{T^\delta \leqq \sigma \leqq \tau\}. \end{cases}
$$

Using these in conjunction with the properties of the functions f and g, we deduce that the following inequalities hold, a.s. P:

$$
\begin{aligned}
&\int_0^{T^\delta \wedge \sigma} \{h(t, X_t(\delta)) - h(t, X_t)\}\, dt \leqq \delta \int_0^{S^\delta \wedge \sigma} h_x(t, Z_t + \delta)\, dt, \\
&f(\sigma)(X_\sigma(\delta) - X_\sigma) 1_{\{\sigma < T^\delta \wedge \tau\}} \leqq \delta f(\sigma) 1_{\{\sigma < S^\delta \wedge \tau\}}, \\
&(g(X_\tau(\delta)) - g(X_\tau)) 1_{\{\sigma = \tau < T^\delta\}} \leqq \delta g'(Z_\tau + \delta) 1_{\{\sigma = \tau < S^\delta\}}.
\end{aligned}
$$

We substitute into (4.18) to obtain a right-hand side which is independent of ξ^-; thus, if we take the supremum of the left-hand side with respect to $\xi^- \in \mathcal{D}(\tau, x)$, we get

$$
\begin{aligned}
\frac{V(\tau, x+\delta) - V(\tau, x)}{\delta} \leqq E\Bigg[&\int_0^{S^\delta \wedge \sigma} h_x(t, Z_t + \delta)\, dt \\
&+ f(\sigma) 1_{\{\sigma < S^\delta \wedge \tau\}} + g'(Z_\tau + \delta) 1_{\{\sigma = \tau < S^\delta\}}\Bigg].
\end{aligned}
$$

Now, since $P(0 \leqq \sigma \leqq \tau, \sigma \neq S) = 1$, we have that, as $\delta \downarrow 0$,

$$
\{\sigma < S^\delta \wedge \tau\} \downarrow \{\sigma < S \wedge \tau\} \mod P, \qquad \{\sigma = \tau < S^\delta\} \downarrow \{\sigma = \tau < S\} \mod P,
$$

and consequently

$$
\int_0^{S^\delta \wedge \sigma} h_x(t, Z_t + \delta)\, dt \xrightarrow[\delta \downarrow 0]{\text{a.s.}} \int_0^{S \wedge \sigma} h_x(t, Z_t)\, dt = \int_0^{S \wedge \sigma} h_x(t, x + W_t)\, dt,
$$

$$
f(\sigma) 1_{\{\sigma < S^\delta \wedge \tau\}} \xrightarrow[\delta \downarrow 0]{\text{a.s.}} f(\sigma) 1_{\{\sigma < S \wedge \tau\}},
$$

$$
g'(Z_\tau + \delta) 1_{\{\sigma = \tau < S^\delta\}} \xrightarrow[\delta \downarrow 0]{\text{a.s.}} g'(Z_\tau) 1_{\{\sigma = \tau < S\}} = g'(x + W_\tau) 1_{\{\sigma = \tau < S\}}.
$$

The result (4.16) follows by the Dominated and Bounded Convergence Theorems. □

By way of preparing the ground for the proof of the opposite inequality, let us consider a control process $\xi^- \in \mathcal{D}(\tau, x)$ for given $x > 0$, along with the corresponding

state process X as in (4.11) and T as in (4.13). We also introduce the stopping time

$$\sigma = \begin{cases} \inf\{0 \leqq t \leqq \tau;\ \xi_t^- > 0\}, \\ \tau, \quad \text{if } \{0 \leqq t \leqq \tau;\ \xi_t^- > 0\} = \varnothing. \end{cases} \tag{4.19}$$

We shall need the processes Z and $Z(-\delta)$ for $0 < \delta < x$, which are Brownian motions starting at x and $x-\delta$, respectively, and are reflected at the origin. Let Z and S be given by (4.14), (4.15), and define

$$Z_t(-\delta) = x - \delta + W_t + L_t(-\delta), \qquad L_t(-\delta) = \max\left[0, \sup_{0 \leqq s \leqq t} \{-(x-\delta) - W_s\}\right]$$

and

$$S(-\delta) \triangleq \inf\{t \geqq 0;\ x - \delta + W_t = 0\} = \inf\{t \geqq 0;\ Z_t(-\delta) = 0\}.$$

We also define the stopping times

$$\sigma^\delta = \begin{cases} \inf\{0 \leqq t \leqq \tau;\ X_t \leqq Z_t(-\delta)\}, \\ \tau, \quad \text{if } \{0 \leqq t \leqq \tau;\ X_t \leqq Z_t(-\delta)\} = \varnothing \end{cases} \tag{4.20}$$

and

$$S^* = \inf\{t \geqq 0;\ Z_t = Z_t(-\delta)\}.$$

Let us observe that $S^* = S$ and $\sigma \wedge T \leqq \sigma^\delta \leqq T \wedge \tau$ hold P-almost surely, and

$$X_t = Z_t \quad \text{on } \{0 \leqq t \leqq \sigma\}, \qquad Z_t = Z_t(-\delta) \quad \text{on } \{t \geqq S\}.$$

We shall also need a number of lemmata which are either evident or easily provable.

LEMMA 4.4. $\{\sigma < S\} = \{\sigma < T\}$, mod P.

LEMMA 4.5. *On* $\{\sigma \geqq T\}$, *we have* $S = T$. *Therefore,* $\sigma \wedge T = \sigma \wedge S$, *a.s.* P.

LEMMA 4.6. $\sigma^\delta \downarrow \sigma \wedge T = \sigma \wedge S$ *as* $\delta \downarrow 0$, *a.s.* P.

For the last result, and for establishing relation (4.22) below, it is instructive to consider separately the three events

$$\Omega_1 = \{\sigma < T,\ \xi_{\sigma+}^- > 0\}, \quad \Omega_2 = \{\sigma < T,\ \xi_{\sigma+}^- = 0\}, \quad \Omega_3 = \{\sigma \geqq T\}.$$

On Ω_1 and with $0 < \delta < \delta_1(\omega) \triangleq \xi_{\sigma+}^-(\omega) \wedge x$, we have

$$\sigma^\delta(\omega) = \sigma(\omega), \quad \xi_{\sigma^\delta(\omega)}^-(\omega) = 0, \quad X_{\sigma^\delta(\omega)}(\omega) - Z_{\sigma^\delta(\omega)}(-\delta, \omega) = \delta.$$

For Ω_2, we introduce the first jump time

$$\bar{\sigma} = \begin{cases} \inf\{0 \leqq t \leqq T;\ \Delta\xi_t^- > 0\}, \\ T, \quad \text{if } \{0 \leqq t \leqq T;\ \Delta\xi_t^- > 0\} = \varnothing \end{cases}$$

of the ξ^- process before time T. Because of our assumption that the jumps of ξ^- are well-ordered (see definition of $\tilde{\mathcal{B}}(\tau)$ at the beginning of this section), we have $\bar{\sigma}(\omega) > \sigma(\omega)$ and

$$\delta_2(\omega) \triangleq [Z_{\bar{\sigma}(\omega)}(\omega) - X_{\bar{\sigma}(\omega)}(\omega)] \wedge x > 0.$$

For $0 < \delta < \delta_2(\omega)$, one has $\sigma^\delta(\omega) \downarrow \sigma(\omega)$, as $\delta \downarrow 0$. Besides, we obtain

$$\xi_{\sigma^\delta(\omega)}^-(\omega) = \delta \quad \text{and} \quad X_{\sigma^\delta(\omega)}(\omega) - Z_{\sigma^\delta(\omega)}(-\delta, \omega) = 0.$$

Finally, on Ω_3 we have for $0 < \delta < \delta_3(\omega) \triangleq x$: $\sigma^\delta(\omega) = S^*(\omega) = S(\omega) = \sigma(\omega) \wedge T(\omega) = \sigma(\omega) \wedge S(\omega)$ and thus $\xi_{\sigma^\delta(\omega)}^-(\omega) = 0$, $X_{\sigma^\delta(\omega)}(\omega) - Z_{\sigma^\delta(\omega)}(-\delta, \omega) = 0$.

The reader might find drawing a picture for each separate case helpful. We shall denote by $\bar{\delta}$ the positive random variable $\sum_{i=1}^3 \delta_i 1_{\Omega_i}$.

PROPOSITION 4.7. *Let us suppose that, for a given $x>0$, there exists an optimal process $\xi^* \in \mathcal{B}(\tau)$ for the Reflected Follower Control Problem. Then*

$$\Delta_- V(\tau, x) \geqq u(\tau, x).$$

Proof. By virtue of Proposition 4.1 we may assume, without loss of generality, that $\xi^* = -\xi^-$, for some $\xi^- \in \mathcal{D}(\tau, x)$. If X is the corresponding reflected state process as in (4.11), we have

$$V(\tau, x) = E\left[\int_0^\tau h(t, X_t)\, dt + \int_{[0,\tau)} f(t)\, d\xi_t^- + g(X_\tau)\right]. \tag{4.21}$$

For any fixed δ in $(0, x)$ we consider the stopping times σ and σ^δ, defined by (4.19) and (4.20), respectively, and we construct the new process

$$Y_t(-\delta) = \begin{cases} Z_t(-\delta), & 0 \leqq t \leqq \sigma^\delta, \\ X_t, & \sigma^\delta < t, \end{cases}$$

which starts at $x-\delta$, follows the reflected Brownian path up to time σ^δ, and then jumps on the X path. It can be verified that the solution to the DRP(Q), with $Q(t) \triangleq x - \delta + W_t - \eta_t^-(-\delta)$, $t \geqq 0$,

$$\eta_t^-(-\delta) = \begin{cases} 0, & 0 \leqq t \leqq \sigma^\delta, \\ \xi_t^- - j(\delta), & \sigma^\delta < t, \end{cases}$$

and $j(\delta) \triangleq \xi^-_{\sigma^\delta +} + X_{\sigma^\delta +} - Z_{\sigma^\delta}(-\delta)$, is precisely $(Y(-\delta), K(-\delta))$. The reflection process

$$K_t(-\delta) = \max\left[0, \sup_{0 \leqq s \leqq t} \{\eta_s^-(-\delta) - (x-\delta) - W_s\}\right], \qquad t \geqq 0$$

is continuous, $Y_t(-\delta) = (x-\delta) + W_t - \eta_t^-(-\delta) + K_t(-\delta)$, $t \geqq 0$, and

$$0 \leqq \Delta\eta_t^-(-\delta) \leqq Y_t(-\delta), \qquad t \geqq 0.$$

In other words, $\eta^-(-\delta)$ is in $\mathcal{D}(\tau, x-\delta)$. Besides, we observe that

$$j(\delta) = \xi^-_{\sigma^\delta} + X_{\sigma^\delta} - Z_{\sigma^\delta}(-\delta) \begin{cases} = \delta 1_{\{\sigma < T\}}, & 0 < \delta < \bar{\delta}(\omega), \\ \geqq 0, & \delta \geqq \bar{\delta}(\omega). \end{cases} \tag{4.22}$$

The first equality is a consequence of the definition of $j(\delta)$ and of the fact that the process $X_t + \xi_t^- = x + W_t + K_t$, $t \geqq 0$ is continuous. The second equality for $0 < \delta < \bar{\delta}(\omega)$ can be verified easily by considering the events Ω_1, Ω_2 and Ω_3 separately; see the discussion following Lemma 4.6. The inequality for $\delta \geqq \bar{\delta}(\omega)$ follows from the definition of σ^δ and the left-continuity of the processes involved.

It is evident now that

$$V(\tau, x-\delta) \leqq E\left[\int_0^\tau h(t, Y_t(-\delta))\, dt + \int_{[0,\tau)} f(t)\, d\eta_t^-(-\delta) + g(Y_\tau(-\delta))\right], \tag{4.23}$$

which yields, in conjunction with (4.21), the inequality

$$\begin{aligned} V(\tau, x) - V(\tau, x-\delta) &\geqq E\left[\int_0^{\sigma^\delta} \{h(t, X_t) - h(t, Z_t(-\delta))\}\, dt + \int_{[\sigma \wedge T, \sigma^\delta)} f(t) d\xi_t^- - f(\sigma^\delta)\xi^-_{\sigma^\delta} 1_{\{\sigma^\delta < \tau\}}\right. \\ &\qquad \left. + j(\delta) f(\sigma^\delta) 1_{\{\sigma^\delta < \tau\}} + \{g(X_\tau) - g(Z_\tau(-\delta))\} 1_{\{\sigma^\delta = \tau\}}\right]. \end{aligned}$$

Furthermore, by taking into account the properties (4.1) of the cost functions h, f and g, we obtain the lower bound

$$\begin{aligned}\frac{V(\tau,x)-V(\tau,x-\delta)}{\delta} \geqq E\Bigg[\int_0^{\sigma^\delta} h_x(t,Z_t(-\delta))\frac{X_t-Z_t(-\delta)}{\delta}\,dt \\ +\frac{j(\delta)}{\delta}1_{\{\delta<\bar{\delta}(\omega)\}}f(\sigma^\delta)1_{\{\sigma^\delta<\tau\}} \\ +g'(Z_\tau(-\delta))\frac{X_\tau-Z_\tau(-\delta)}{\delta}1_{\{\sigma^\delta=\tau\}}\Bigg] \\ +E\Bigg[\{\min_{\sigma\wedge T\leqq t\leqq\sigma^\delta} f(t)-f(\sigma^\delta)\}\frac{\xi_{\sigma^\delta}^-}{\delta}1_{\{\sigma^\delta<\tau\}}\Bigg].\end{aligned} \tag{4.24}$$

For the first term on the right-hand side of (4.24) we recall that $h_x(t,\cdot)$ is increasing and nonnegative on $\mathbb{R}^+$ (condition (4.1iii)). Along with $\sigma^\delta \geqq S\wedge\sigma$, this gives the lower bound

$$\begin{aligned}\int_0^{\sigma^\delta} h_x(t,Z_t(-\delta))\frac{X_t-Z_t(-\delta)}{\delta}\,dt &\geqq \int_0^{S\wedge\sigma\wedge S(-\delta)} h_x(t,Z_t(-\delta))\frac{Z_t-Z_t(-\delta)}{\delta}\,dt \\ &= \int_0^{S(-\delta)\wedge\sigma} h_x(t,x-\delta+W_t)\,dt,\end{aligned}$$

P-almost surely. We obtain from (4.24) in this way:

$$\frac{V(\tau,x)-V(\tau,x-\delta)}{\delta} \geqq R(\sigma;\tau,x)-\sum_{j=1}^4 I_j(\delta), \tag{4.25}$$

where the stopping time σ is given by (4.19) and

$$I_1(\delta) \triangleq E\Bigg[\{f(\sigma^\delta)-\min_{\sigma\wedge T\leqq t\leqq\sigma^\delta} f(t)\}\frac{\xi_{\sigma^\delta}^-}{\delta}1_{\{\sigma^\delta<\tau\}}\Bigg], \tag{4.26}$$

$$I_2(\delta) \triangleq E\Bigg[\int_0^{S\wedge\sigma} h_x(t,x+W_t)\,dt-\int_0^{S(-\delta)\wedge\sigma} h_x(t,x-\delta+W_t)\,dt\Bigg], \tag{4.27}$$

$$I_3(\delta) \triangleq E\left|f(\sigma)1_{\{\sigma<\tau\wedge S\}}-\frac{j(\delta)}{\delta}f(\sigma^\delta)1_{\{\sigma^\delta<\tau\}\cap\{\delta<\bar{\delta}(\omega)\}}\right|, \tag{4.28}$$

$$I_4(\delta) \triangleq E\left|g'(x+W_\tau)1_{\{\sigma=\tau<S\}}-g'(Z_\tau(-\delta))\frac{X_\tau-Z_\tau(-\delta)}{\delta}1_{\{\sigma^\delta=\tau\}}\right|. \tag{4.29}$$

By the Bounded Convergence Theorem, as well as the facts: $0\leqq\xi_{\sigma^\delta}^-\leqq\delta$ and $\sigma^\delta\downarrow\sigma\wedge T$ as $\delta\downarrow 0$, P-almost surely, the term $I_1(\delta)$ is seen to converge to zero as $\delta\downarrow 0$. The term $I_2(\delta)$ in (4.27) also converges to zero as $\delta\downarrow 0$, because of the polynomial growth of $h_x(t,\cdot)$, of the relation

$$S(-\delta)\xrightarrow[\delta\downarrow 0]{\text{a.s.}} S,$$

and of the Dominated Convergence Theorem. On the other hand, the relation $\lim_{\delta\downarrow 0} I_3(\delta)=0$ is a consequence of (4.22), of Lemma 4.6 and of the Bounded Convergence Theorem.

To prove $\lim_{\delta\downarrow 0} I_4(\delta)=0$, we note that since $P\{S=\tau\}=0$, we have $\{\sigma^\delta=\tau\}\downarrow\{\sigma=\tau<S\}$ mod P, and

$$1_{\{\sigma^\delta=\tau\}}-1_{\{\sigma=\tau<S\}}=1_{\{\sigma^\delta=\tau,\sigma\wedge S<\tau\}},\qquad \text{a.s. } P.$$

Therefore,

$$\begin{aligned} &E\left|g'(Z_\tau(-\delta))\frac{X_\tau-Z_\tau(-\delta)}{\delta}1_{\{\sigma^\delta=\tau\}}-g'(x+W_\tau)1_{\{\sigma=\tau<S\}}\right| \\ &\leqq E\left|g'(Z_\tau(-\delta))\frac{X_\tau-Z_\tau(-\delta)}{\delta}1_{\{\sigma^\delta=\tau,\sigma\wedge S<\tau\}}\right| \qquad (4.30)\\ &\quad+E\left\{1_{\{\sigma=\tau<S\}}\left|g'(Z_\tau(-\delta))\frac{X_\tau-Z_\tau(-\delta)}{\delta}-g'(x+W_\tau)\right|\right\}. \end{aligned}$$

On the set $\{\sigma^\delta=\tau\}$, we have

$$0\leqq X_\tau-Z_\tau(-\delta)\leqq Z_\tau-Z_\tau(-\delta)\leqq\delta,$$

so the first term on the right-hand side of (4.30) is bounded above by

$$E|g'(Z_\tau(-\delta))1_{\{\sigma^\delta=\tau,\,\sigma\wedge S<\tau\}}|.$$

Because of the polynomial growth condition (4.1iv) on g' and Lemma 4.6, this approaches zero as $\delta\downarrow 0$. For fixed $\omega\in\{\sigma=\tau<S\}$, we have $x+W_\tau(\omega)>0$, and so for sufficiently small δ, $Z_\tau(-\delta,\omega)=x-\delta+W_\tau(\omega)$ and $X_\tau(\omega)-Z_\tau(-\delta,\omega)=\delta$. Therefore,

$$g'(Z_\tau(-\delta))\frac{X_\tau-Z_\tau(-\delta)}{\delta}\xrightarrow[\delta\downarrow 0]{} g'(x+W_\tau),\quad P\text{-a.s. on }\{\sigma=\tau<S\},$$

and the Dominated Convergence Theorem implies that the second term on the right-hand side of (4.30) converges to zero as δ does; here we employ the fact that, on $\{\sigma=\tau<S\}$: $X_\tau-Z_\tau(-\delta)=Z_\tau-Z_\tau(-\delta)\in[0,\delta]$.

We can now pass to the limit as $\delta\downarrow 0$ in (4.25) to obtain

$$\Delta_- V(\tau,x)\geqq R(\sigma;\tau,x)\geqq u(\tau,x). \qquad (4.31)$$ □

Remark. Based on the polynomial growth of $h_x(t,\cdot)$, $g'(\cdot)$ and on the boundedness of $f(t)$, one can show that the quantities in (4.26)-(4.29) admit, for any $M>0$, a bound of the form

$$\sup_{\substack{0<\delta<x\\0<x\leqq M}}\sum_{j=1}^{4} I_j(\delta)\leqq q(\tau,M)$$

depending on τ, M, the bound on f and the constants C, m in (4.1iv). We thus deduce from (4.25) the useful lower bound

$$V(\tau,x)-V(\tau,x-\delta)\geqq-\delta\cdot q(\tau,M),\qquad 0<\delta\leqq x\leqq M, \qquad (4.32)$$

provided that an optimal process for the Reflected Follower Problem exists for every $x\in(0,M]$.

PROPOSITION 4.8. *For all $x>0$, we have*

$$\Delta^- V(\tau,x)\leqq u(\tau,x). \qquad (4.33)$$

Proof. Let us fix $\delta \in (0, x]$. For any process $\xi^- \in \mathcal{D}(\tau, x-\delta)$ we create the solution $(X(-\delta), K(-\delta))$ to the DRP$(Q(-\delta))$ with $Q_t(-\delta) = x - \delta + W_t - \xi_t^-$, $t \geq 0$:

$$X_t(-\delta) = x - \delta + W_t - \xi_t^- + K_t(-\delta), \qquad t \geq 0,$$

$$K_t(-\delta) = \max\left[0, \sup_{0 \leq s \leq t} \{\xi_s^- - (x - \delta + W_s)\}\right], \qquad t \geq 0.$$

Because ξ^- is also in $\mathcal{D}(\tau, x)$ (Lemma 4.2), the solution (X, K) to the DRP(Q) with $Q_t = x + W_t - \xi_t^-$, $t \geq 0$ is given by (4.11), (4.11)′. Defining the stopping times

$$\Gamma(-\delta) \triangleq \inf\{t \geq 0;\ \xi_t^- - (x - \delta + W_t) = 0\} = \inf\{t \geq 0;\ X_t(-\delta) = 0\},$$

$$\Gamma \triangleq \inf\{t \geq 0;\ X_t(-\delta) \geq X_t\},$$

and recalling the stopping time T from (4.13), it is not hard to see that $\Gamma = T$ and

$$X_t(-\delta) = X_t, \qquad t \geq T$$

hold P-almost surely.

Imitating the proof of Proposition 4.3, we choose an $\{\mathcal{F}_t\}$-stopping time σ with $P(0 \leq \sigma \leq \tau) = 1$ and create a trajectory $\tilde{Y}$ which emanates from x:

$$\tilde{Y}_t \triangleq \begin{cases} X_t, & 0 \leq t \leq \sigma, \\ X_t(-\delta), & t > \sigma, \end{cases}$$

or $\tilde{Y}_t = x + W_t - \eta_t^- + \Lambda_t$, $t \geq 0$, so that $(\tilde{Y}, \Lambda)$ is the unique solution to the DRP$(\tilde{Q})$ with $\tilde{Q}_t = x + W_t - \tilde{\eta}_t$, $t \geq 0$, and

$$\tilde{\eta}_t = \begin{cases} \xi_t^-, & 0 \leq t \leq \sigma, \\ \xi_t^- + (X_\sigma - X_\sigma(-\delta)), & t > \sigma. \end{cases}$$

It can be checked here again that

$$\Lambda_t = \max\left[0, \sup_{0 \leq s \leq t} \{\tilde{\eta}_s - (x + W_s)\}\right], \qquad 0 \leq \Delta\tilde{\eta}_t \leq \tilde{Y}_t, \qquad t \geq 0,$$

in other words that $\tilde{\eta} \in \mathcal{D}(\tau, x)$. Consequently, we have

$$V(\tau, x) \leq E\left[\int_0^\tau h(t, \tilde{Y}_t)\, dt + \int_{[0,\tau)} f(t)\, d\tilde{\eta}_t + g(\tilde{Y}_\tau)\right]$$

and, repeating the argument in the proof of Proposition 4.3 with the obvious changes, we obtain

$$\begin{aligned} &\frac{V(\tau, x) - J(-\xi^-; \tau, x-\delta)}{\delta} \\ &\leq E\left[\int_0^{\Gamma \wedge \sigma} h_x(t, Z_t)\, dt + f(\sigma) 1_{\{\sigma < \Gamma \wedge \tau\}} + g'(Z_\tau) 1_{\{\sigma = \tau < \Gamma\}}\right] \\ &\leq E\left[\int_0^{S \wedge \sigma} h_x(t, x + W_t) + f(\sigma) 1_{\{\sigma < S \wedge \tau\}} + g'(x + W_\tau) 1_{\{\sigma = \tau < S\}}\right] \end{aligned}$$

because $\Gamma = T < S$, a.s. P. By taking the supremum of the left-hand side with respect to $\xi^- \in \mathcal{D}(\tau, x-\delta)$, and then the infimum of the right-hand side with respect to the arbitrary stopping time σ, we obtain

$$V(\tau, x) - V(\tau, x-\delta) \leq \delta \cdot u(\tau, x), \qquad 0 < \delta \leq x. \tag{4.34}$$

The result (4.33) follows. □

We are now in a position to establish the main result of this paper.

THEOREM 4.9. *Let us fix a point $x>0$ and suppose that an optimal process $\xi_*^- \in \mathcal{D}(\tau, x)$ exists for the Reflected Follower Problem at (τ, x). Then the left gradient*

$$V_x^-(\tau, x) \triangleq \lim_{\delta \downarrow 0} \frac{V(\tau, x) - V(\tau, x-\delta)}{\delta}$$

exists and is equal to the optimal risk $u(\tau, x)$ for the Stopping Problem with Absorption. Besides, the stopping time

$$\sigma^* = \begin{cases} \inf\{0 \leqq t \leqq \tau;\ \xi_*^-(t) > 0\}, \\ \tau, \quad \text{if } \{0 \leqq t \leqq \tau;\ \xi_*^-(t) > 0\} = \varnothing, \end{cases} \tag{4.19)'}$$

is optimal for the latter problem, i.e.,

$$V_x^-(\tau, x) = u(\tau, x) = R(\sigma^*; \tau, x). \tag{4.35}$$

Suppose now that the Reflected Follower Problem admits an optimal process at every point $x \geqq 0$. Then the gradient $V_x(\tau, x)$ exists at every point $x>0$, and $V_x(\tau, x) = u(\tau, x)$. Besides,

$$V_x^+(\tau, 0) \triangleq \lim_{\delta \downarrow 0} \frac{V(\tau, \delta) - V(\tau, 0)}{\delta} = 0.$$

Proof. The first assertion is an immediate consequence of Propositions 4.7 and 4.8, in particular of relations (4.31) and (4.33):

$$\Delta_- V(\tau, x) \geqq R(\sigma^*; \tau, x) \geqq u(\tau, x) \geqq \Delta^- V(\tau, x) \geqq \Delta_- V(\tau, x).$$

For the second assertion, let us consider an arbitrary positive number M, and observe that

$$|V(\tau, x) - V(\tau, x-\delta)| \leqq \delta \cdot \max[u(\tau, M), q(\tau, M)], \qquad 0 < \delta \leqq x \leqq M \tag{4.36}$$

follows from (4.32) and (4.34). Now (4.36) implies that the function $V(\tau, \cdot)$ is absolutely continuous on $[0, M]$, and in particular that the gradient

$$v(\tau, x) = V_x(\tau, x)$$

exists for almost all (with respect to Lebesgue measure λ) points x in $[0, M]$. The function $V(\tau, \cdot)$ can thus be written as the integral of its derivative

$$V(\tau, x) = V(\tau, 0) + \int_0^x v(\tau, \xi)\, d\xi, \qquad 0 \leqq x \leqq M.$$

However, under our assumptions and by virtue of the first assertion of this theorem, the left derivative $V_x^-(\tau, x)$ exists *everywhere* in $(0, M]$ and is equal to $u(\tau, x)$. Therefore, $v(\tau, x) = u(\tau, x)$ for λ-a.e. $x \in [0, M]$, and so

$$V(\tau, x) = V(\tau, 0) + \int_0^x u(\tau, \xi)\, d\xi, \qquad 0 \leqq x \leqq M. \tag{4.37}$$

Because $M>0$ is arbitrary, (4.37) actually holds on the whole of $[0, \infty)$. Let us recall that the integrand $u(\tau, \cdot)$ is nondecreasing and verify, in a completely straightforward manner, that for any $0 \leqq y < z$, $0 \leqq \theta \leqq 1$ and with $x = \theta y + (1-\theta) z$, we have

$$V(\tau, x) \leqq \theta V(\tau, y) + (1-\theta) V(\tau, z), \tag{4.38}$$

i.e., $V(\tau, \cdot)$ is convex on $\mathbb{R}^+$.

Now we fix a point $x>0$ and choose two numbers $0<\delta_1 \leqq x$, $\delta_2>0$; with $y=x-\delta_1$, $z=x+\delta_2$ and $\theta=\delta_2/(\delta_1+\delta_2)$ in (4.38), we have

$$\frac{V(\tau,x)-V(\tau,x-\delta_1)}{\delta_1} \leqq \frac{V(\tau,x+\delta_2)-V(\tau,x)}{\delta_2},$$

and upon letting δ_1, δ_2 converge to zero, we obtain

$$\Delta^- V(\tau,x) \leqq \Delta_+ V(\tau,x). \tag{4.39}$$

Propositions 4.3 and 4.7 yield the string of inequalities

$$\Delta_+ V(\tau,x) \leqq \Delta^+ V(\tau,x) \leqq u(\tau,x) \leqq \Delta_- V(\tau,x) \leqq \Delta^- V(\tau,x),$$

which establish, in conjunction with (4.39), the existence of the gradient and the identity

$$V_x(\tau,x)=u(\tau,x)$$

at the arbitrarily chosen point $x>0$. At the origin $x=0$ we have from Proposition 4.3

$$\Delta^+ V(\tau,0) \leqq 0,$$

and from the fact that the function $V(\tau,\cdot)$ is nondecreasing (a consequence of the integral representation (4.37), since $u \geqq 0$):

$$\Delta_+ V(\tau,0) \geqq 0.$$

It follows that the right derivative $V_x^+(\tau,0)$ exists and is equal to zero. □

5. Discussion of the Bounded Variation Follower Problem. Let us suppose that the functions $h(t,\cdot)$, $g(\cdot)$ have been extended evenly on the whole of $\mathbb{R}$, and formulate the "unfolded" (Bounded Variation Follower) control problem: to find a process $\xi \in \mathcal{B}(\tau)$ which minimizes the expected cost $J(\xi;\tau,x)$ as in (4.2), where now the state process X is given by

$$X_t = x + W_t + \xi_t, \qquad 0 \leqq t \leqq \tau, \quad x \in \mathbb{R}.$$

We call $U(\tau,x)$ the value function of this problem.

It is rather obvious that $U(\tau,\cdot)$ is an even, convex function on $\mathbb{R}$; it inherits these properties from the cost functions $h(t,\cdot)$, $g(\cdot)$. You can also easily believe (but less easily prove) that the value functions of the two control problems agree on $[0,\infty)$:

$$U(\tau,x)=V(\tau,x), \qquad x \geqq 0. \tag{5.1}$$

Indeed, (5.1) can be verified by means of the change-of-variable formula for semimartingales, once $V(\tau,x)$ is known to satisfy the relevant variational inequality or free boundary problem on $\mathbb{R}^+ \times [0,\infty)$. This (mostly analytical) approach was taken up in [13]; however, it requires greater smoothness on the data f, g and h than has been assumed here. The authors have attempted to establish (5.1) by direct, purely probabilistic arguments and encountered some unexpectedly thorny issues. They suggest such a derivation as an interesting open problem.

REFERENCES

[1] J. A. BATHER, *Optimal stopping problems for Brownian motion*, Adv. Appl. Prob., 2 (1970), pp. 259–286.

[2] J. A. BATHER AND H. CHERNOFF, *Sequential decisions in the control of a spaceship*, Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability, 3 (1966), pp. 181–207.

[3] V. E. BENEŠ, L. A. SHEPP AND H. S. WITSENHAUSEN, *Some solvable stochastic control problems*, Stochastics, 4 (1980), pp. 134–160.

[4] M. CHALEYAT-MAUREL, N. EL KAROUI AND B. MARCHAL, *Réflexion discontinue et systèmes stochastiques*, Ann. Probab., 8 (1980), pp. 1049-1067.
[5] H. CHERNOFF, *Optimal stochastic control*, Sankhyā, Ser. A, 30 (1968), pp. 221-252.
[6] C. DELLACHERIE, *Capacités et processus stochastiques*, Springer-Verlag, Berlin, 1972.
[7] A. FRIEDMAN, *Stochastic Differential Equations and Applications, Vol. 2*, Academic Press, New York, 1976.
[8] B. I. GRIGELIONIS AND A. N. SHIRYAEV, *On Stefan's problem and optimal stopping rules for Markov processes*, Theory Prob. Appl., 11 (1966), pp. 541-558.
[9] J. M. HARRISON AND M. I. TAKSAR, *Instantaneous control of a Brownian motion*, Math. Oper. Res., 8 (1983), pp. 439-453.
[10] J. M. HARRISON AND A. J. TAYLOR, *Optimal control of a Brownian storage system*, Stoch. Proc. Appl., 6 (1978), pp. 179-194.
[11] N. IKEDA AND S. WATANABE, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.
[12] I. KARATZAS, *The monotone follower problem in stochastic decision theory*, Appl. Math. Optim., 7 (1981), pp. 175-189.
[13] ———, *A class of singular stochastic control problems*, Adv. Appl. Probab., 15 (1983), pp. 225-254.
[14] I. KARATZAS AND S. E. SHREVE, *Connections between optimal stopping and singular stochastic control I. Monotone follower problems*, this Journal, 22 (1984), pp. 856-877.
[15] A. N. SHIRYAEV, *Optimal Stopping Rules*, Springer-Verlag, Berlin, 1978.
[16] S. E. SHREVE, J. P. LEHOCZKY AND D. P. GAVER, *Optimal consumption for general diffusions with absorbing and reflecting barriers*, this Journal, 22 (1984), pp. 55-75.
[17] P. VAN MOERBEKE, *On optimal stopping and free boundary problems*, Arch. Rational Mech. Anal., 60 (1976), pp. 101-148.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

GLOBAL CONTROLLABILITY FOR SMOOTH NONLINEAR SYSTEMS: A GEOMETRIC APPROACH*

DIRK AEYELS†

Abstract. The paper consists of two parts. In Part I a global controllability theory is constructed. Local controllability criteria are "integrated" by means of the qualitative behavior of the dynamics involved—leading to a global result. The relation with global observability is discussed. Part II is concerned with controllability by means of two vectorfields. It is well known that each connected paracompact manifold of class C^k, $2 \leqq k \leqq \infty$, or $k = \omega$, carries a globally controllable set of two C^{k-1}-vectorfields. As an application of Part I we construct control systems of the form $\dot{x} = X + uY$, defined on a compact manifold, with the property that for X a Morse-Smale vectorfield with no periodic orbits, there exists a large class of vectorfields Y such that the system is globally controllable by means of a bang-bang (two-valued) control. It is stressed that the techniques applied imply the existence of a multitude of basically different globally controllable systems with two-valued controls on each smooth connected compact manifold. In particular, it is shown that Y can be taken to be Morse-Smale. The results are extended to noncompact manifolds.

Key words. nonlinear systems, global controllability, bang-bang controllability, Morse-Smale systems

Part I. Global controllability: a geometric approach.

1. Introduction. This article is concerned with the study of global controllability properties of nonlinear control systems defined on smooth manifolds. An extensive literature on the subject of controllability of systems on manifolds is available. It includes the work of Hermann [4]—who apparently initiated the field—followed by contributions of Lobry, Sussmann and Jurdjevic, Sussmann, Brockett, and Krener, among others (for a reference list, see [5]). Most of this work is preoccupied with *local* controllability properties. The question of *global* controllability—when can *any* two distinct points on the manifold be connected following the trajectories of the control system?—has received much less attention. For smooth systems in their general form $\dot{x} = f(x, u)$ the problem is clearly too hard. Hence, the problem statement is narrowed down to finding "reasonable" classes of control systems for which "tractable" global controllability conditions can be derived. Of course, the meaning of the words "reasonable" and "tractable" is to be filled in, and can in fact only be appreciated after a class of systems has been picked and its global controllability theory developed.

In this paper we present a class of control systems and an associated control theory which satisfy the subjective requirements mentioned above. A formal description will be given in the next section. As for now, it suffices to say that they constitute a restricted form of Morse-Smale systems [10]. The motivation for a study of this class is linked to Aeyels [1]. In that paper a global observability condition has been established, based on information on the topology of the dynamics of the control system. This paper again tries to utilize qualitative properties of the control system and to translate them into a global controllability condition. The similarities and points of difference between [1] and this paper, with regard to the results obtained, will be discussed in § 5.

The organization of part I is as follows: in § 2 we define the class of systems we are concerned with; in addition we introduce notations and basic concepts. Section 3

* Received by the editors March 24, 1983, and in revised form April 20, 1984. This research was supported in part by the National Fund for Scientific Research (FKFO grant).
† Department of Systems Dynamics, State University of Gent, Gent, Belgium.

is devoted to establishing controllability results which will be needed later. In § 4 we prove our main result on global controllability. Section 5 contains a discussion of various aspects of the theory.

2. Definitions. Notation. A class of nonlinear control systems. We will study control systems of the form $\dot{x} = X(x) + uY(x)$, $u \in \mathbb{R}$, with X and Y C^∞-vectorfields defined on a smooth compact manifold. The compactness condition will be discussed in the final section. The controls $u(t)$ are piecewise constant and are allowed to assume arbitrary large, but finite values. Let $x(t, x_0)$ with $x(0, x_0) = x_0$ and $y(t, x_0)$ with $y(0, x_0) = y_0$, as functions of t, denote the *integral curves* of X and Y starting at x_0 and y_0 respectively; $x(\cdot, \cdot)$ is unique and is called the *flow* induced by X. A vectorfield X is *complete* if $x(\cdot, \cdot)$ is defined on $R \times M$. If M is compact, its smooth vectorfields are complete. An *orbit* of the vectorfield X is the image of an integral curve. A point p of M is called *equilibrium point* or *critical point* if $x(R, p) = p$. An equilibrium point p of X is *hyperbolic*, if (in local coordinates) the derivative of X, evaluated at p, has no eigenvalues with zero real part. If all eigenvalues have a negative real part, then p is a *sink* or an *attractor*; if all eigenvalues have positive real part, then p is a *source* or a *repellor*; if eigenvalues are present with both positive and negative real part, then p is a *saddlepoint* or *saddle*. A point $p \in M$ is a *nonwandering point* for the flow of X if, for all neighborhoods U of p and all $T > 0$, there is a $t > T$ such that $U \cap x(t, U)$ is nonempty. Equilibrium points are nonwandering. To describe asymptotic behavior, we define the *ω-limit set* (*α-limit set*) of $p \in M$ as: $\{m \in M$ for which there exists a sequence $\{t_n : t_n \to \infty\}$ ($\{t_n : t_n \to -\infty\}$) such that $x(p, t_n) \to m\}$. They are denoted by $\omega(p)$ and $\alpha(p)$ respectively. Define the *stable manifold* of a critical point p by $W^s(p) = \{m \in M : \omega(m) = p\}$; for the *unstable manifold* $W^u(p)$, ω is replaced by α. $W^s(p)$ and $W^u(p)$, for hyperbolic p, are immersed submanifolds.

For the control systems to be studied in the following sections, we assume that both X and Y have a finite number of hyperbolic equilibrium points and that their nonwandering sets equal the union of their equilibrium points. Furthermore, assume for both X and Y, that if p and q are equilibrium points of a vectorfield (say X), then $W^s(p)$ and $W^u(q)$ (with respect to X) are transversal [10]. A control system with these properties will be denoted by $\dot{x} = X_1(x) + uX_2(x)$, indices indicating the restrictions imposed. It can be shown that, generically, gradient systems on compact manifolds satisfy the above properties [16]. The vectorfields X_1 and X_2 are called Morse-Smale (with no periodic orbits).

It is possible to introduce a partial order among the equilibrium points of the vectorfield X_1 as follows: $p \geqq q$ (p and q are equilibrium points of X_1) in case $W^u(p)$ meets $W^s(q)$. It is important to realize that as a consequence of the transversality assumption on stable and unstable manifolds, $\dim W^u(q) < \dim W^u(p)$; thus a source is a maximal element of the partial order, and a sink is a minimal element. It is convenient to picture this qualitative information on X_1 as a diagram with vertices representing the equilibrium points and "arrowed" segments (arrows in the sense from p to q, if $p > q$) connecting vertices related by the partial order. Moreover, the equilibrium points are grouped in subsets, each set containing the equilibrium points having unstable manifolds of the same dimension. Vertices representing equilibrium points of a particular subset constitute a row. The vertices representing the sources are at the top row; the next row contains the vertices corresponding to the equilibrium points having a $(n-1)$-dimensional unstable manifold, etc., down to the bottom row which contains the sinks. Notice that in this graphical representation, all arrows are directed *downwards*.

For an investigation of the vectorfields, having the properties above, with respect to their structural stability properties and their position within the set of all vectorfields, one is referred to [10]. For our purposes it is important to realize that the systems $\dot{x} = X_1 + uX_2$ constitute an *open* class of systems exhibiting "honest" nonlinearities, and that an investigation of such systems might give us hints as to what can be achieved and expected in a broader theory on *nonlinear* controllability.

3. Local controllability. In this section two local controllability results are presented; they are essential in the proof of a theorem on global controllability, to be derived in the next section. By *local controllability at a point p* we mean that there is a neighborhood of p, such that for any two points a and b in that neighborhood, a piecewise constant control $u(t)$ with a finite number of switchings exists, which controls a to b, i.e. $x_{u(t)}(T, a) = b$ for some finite time T, where $x_{u(t)}(\cdot, \cdot)$ represents the flow of the control system $\dot{x} = X_1(x) + u(t)X_2(x)$, $x \in M$. A control system has the *accessibility property to (from) a point p*, if the set of points that can be steered towards p (that can be reached from p) by means of piecewise constant controls, has a nonempty interior.

THEOREM 3.1. *Let $\mathscr{L}\{X_1, X_2\}$ denote the Lie algebra generated by $\{X_1, X_2\}$ and suppose that* $\dim \mathscr{L}\{X_1, X_2\}(x_0) = n$, *where n is the dimension of the state manifold. Then x_0 is contained in the closure of the interior of the reachable set of $\dot{x} = X_1 + uX_2$ from x_0. It follows that the system has the accessibility property to (from) x_0, even with bang-bang control* [12].

Note that if $X_2(x_0) = 0$ and if $b = X_1(x_0)$, $A = DX_2(x_0)$, then rank $[b, Ab, \cdots, A^{n-1}b] = n \Rightarrow \dim \mathscr{L}\{X_1, X_2\}(x_0) = n$. We will refer to this assumption by saying that the rank condition is satisfied.

The following theorem is the well-known result of Lee and Markus [9].

THEOREM 3.2. *The system $\dot{x} = X_1(x) + uX_2(x)$ is locally controllable at a point x_e, with x_e an equilibrium point of $X_1(x)$, if $[b, Ab, \cdots, A^{n-1}b]$ has full rank equal to n. Here $b := X_2(x_e)$, $A := DX_1(x_e)$ and n is the dimension of the state manifold.*

Remark. Local controllability has been established for piecewise constant controls with values in some interval $[-m, +m]$. By the work of Krener [8], the result remains true when the admissible controls are restricted to bang-bang controls with a positive and a negative value.

4. The global controllability theorem. In this section we prove a global controllability criterion for systems of the form $\dot{x} = X_1(x) + uX_2(x)$, $u \in R$, $x \in M$ and M compact. It is a standard assumption that the rank condition at each equilibrium point p of X_1 is satisfied (i.e. rank $[b, Ab, \cdots, A^{n-1}b] = n$, with $b := X_2(p)$ and $A := DX_1(p)$). Consider the diagram associated with X_1 as described in § 2. The top line contains the vertices representing the sources $\mathrm{so}_1, \mathrm{so}_2, \cdots$ of X_1; the saddlepoints $\mathrm{sa}_1^k, \mathrm{sa}_2^k, \cdots$, with unstable manifold of dimension k occupy the $(n+1-k)$th row; the sinks $\mathrm{si}_1, \mathrm{si}_2, \cdots$ are at the bottom row; so may also be denoted as sa^n and si as sa^0.

If p is an equilibrium point of X_1 and $x_0 \in W^s(p)$, then $\lim_{t \to \infty} x_1(t, x_0) = p$. This means that the trajectory through x_0 of the vectorfield X_1 is in an arbitrary small neighborhood of p after some time. Similarly, if $x_0 \in W^u(q)$, with q an equilibrium point of X_1, then $\lim_{t \to -\infty} x_1(t, x_0) = q$, i.e. in any neighborhood of q (no matter how small), there is a point y such that the set $\{x_1(t, y), t \geqq 0\}$ contains x_0. Both these properties will be used frequently in the sequel. For easy reference, they are designated by "LIM" (for limit behavior).

For global controllability, an obvious *necessary* condition is that *each sink* of X_1 can be controlled to *each source* of X_1—since one must be able to connect *any* two

points along the trajectories of the control system. We claim that this condition is also *sufficient.* Indeed, first assign to each equilibrium point s of X_1 an open ball $\mathcal{O}(s)$ given by Theorem 3.2; recall that any two points of $\mathcal{O}(s)$ can be connected along the trajectories of the control system $\dot{x} = X_1 + uX_2$. Let a^* and b^* be two arbitrary points on the state manifold M. Then, by the assumptions on the structure of X_1, $a^* \in W^s(\mathrm{sa}_j^k)$ for some saddle sa_j^k and $b^* \in W^u(\mathrm{sa}_l^m)$ for some saddle sa_l^m. Indicate on the diagram (consisting of vertices and arrows) representing X_1, a path, following arrows, which connects sa_j^k with some sink si_r. Such a path exists by the defining properties of X_1 (if sa_j^k is a sink itself, then the path degenerates into a vertex sa_j^k). Also indicate a path in the direction of the arrows connecting some source so_t with sa_l^m (if sa_l^m is a source, the path degenerates into the vertex sa_l^m). Recall that it has been assumed that all sinks can be steered to all sources, in particular, si_r can be connected with so_t. We now show how to control a^* to b^*. First, starting in a^* apply the control $u = 0$; after some time t_1, $a_1 := x_1(t_1, a^*) \in \mathcal{O}(\mathrm{sa}_j^k)$, by LIM. Next, by local controllability in $\mathcal{O}(\mathrm{sa}_j^k)$ and by LIM, a_1 can be steered to a point $b_1 \in \mathcal{O}(\mathrm{sa}_j^k)$ such that b_1 is on the right "track", as traced out above. More specifically, the path mapped out above has an arrow leaving sa_j^k and pointing towards another equilibrium point of X_1. The point b_1 is chosen such that, when setting again $u = 0$, after some time b_1 settles in the neighborhood of that equilibrium point. It should be clear that by repeating this procedure a finite number of times a^* can be controlled to the neighborhood $\mathcal{O}(\mathrm{si}_r)$ and therefore to si_r. Now si_r is steered to so_t (possible by assumption). But so_t can be connected with b^* by applying a control law similar to the one connecting a^* with si_r. We have shown that if the rank condition is satisfied at each equilibrium point of X_1, and if each sink of X_1 can be controlled to each source of X_1, then the system $\dot{x} = X_1 + uX_2$ is globally controllable. *What is needed now, is a control strategy that shows how to attain an arbitrary source of X_1—starting in an arbitrary sink of X_1—along the trajectories of the system $\dot{x} = X_1 + uX_2$.* Here, the dynamics of X_2 will play a prominent part. Notice that there are simple (nontrivial) examples, showing the necessity of a global controllability condition *beyond* a combination of local accessibility (or controllability) conditions.

We now propose such a condition and illustrate its main features by means of a working example. We also show how to establish global controllablility for this example; by doing so we will to some degree have covered the general case.

Consider the unit sphere $S^2 \subset \mathbb{R}^3$ centered at the origin. Let X_1 and X_2 be smooth vectorfields on S^2. The vectorfield X_1 has two equilibrium points: a source in the north-pole $\mathrm{NP} = (0, 0, 1)$ and a sink in the south-pole $\mathrm{SP} = (0, 0, -1)$. The other orbits of the flow are the meridian lines. As for X_2, let it also have two equilibrium points; a sink x_e, say somewhere in Europe, and a source in the USA. The equilibrium points of X_2 are separate from the equilibrium points of X_1. It follows from the previous paragraph, that if NP can be attained from SP, then the system on S^2 is globally controllable. We proceed to show how this can be accomplished. First associate with NP and SP its local controllability neighborhoods $\mathcal{O}(\mathrm{NP})$ and $\mathcal{O}(\mathrm{SP})$ as given by Theorem 3.2. Then attach to x_e its accessibility set $C(\to x_e)$ given by Theorem 3.1 (assume the rank condition is satisfied at x_e). *Notice that the (X_2)-trajectory through* SP *as $t \to \infty$, enters any neighborhood of x_e.* What is required for global controllability is that this (X_2)-trajectory through SP approaches x_e the "right" way, i.e. the set $\{x_2(t, \mathrm{SP}), t > 0\}$ intersects $\mathring{C}(\to x_e)$. This is the global controllability condition for this particular case. By this condition and by continuity of solutions of differential equations with respect to data there is a k_1^*, positive and large enough such that for some t_1 the trajectory corresponding to $X_1 + k_1^* X_2$ sends SP in $\mathring{C}(\to x_e)$ at t_1; from there on it can be steered towards x_e by a bang-bang control denoted by u_1^* (Theorem 3.1). We denote

the control sequence steering SP to x_e by $\{k_1^*, u_1^*\}$. Apply this control to all points in $\mathcal{O}(\mathrm{SP})$. This results in a neighborhood $\mathcal{E}(x_e)$ of x_e. Since NP belongs to the unstable manifold of x_e, corresponding to $-X_2$, then, by LIM and by continuity of solutions of differential equations with respect to data, there is a control $-|k_2^*|$, negative and large enough, which steers an open set $\mathcal{O}_{\mathrm{lift}} \subset \mathcal{E}(x_e)$ into $\mathcal{O}(\mathrm{NP})$. Since $\mathcal{O}(\mathrm{SP})$ can be steered onto $\mathcal{E}(x_e)$ (by definition of $\mathcal{E}(x_e)$), there exists an open set $\mathcal{O}_{\mathrm{start}} \subset \mathcal{O}(\mathrm{SP})$ which by $\{k_1^*, u_1^*\}$ is mapped onto $\mathcal{O}_{\mathrm{lift}}$. The control law realizing the connection of SP with NP now follows immediately. First send SP to some point in $\mathcal{O}_{\mathrm{start}}$ by means of a bang-bang control (possible by Theorem 3.2). Then apply $\{k_1^*, u_1^*\}$ to reach a point in $\mathcal{O}_{\mathrm{lift}}$. Apply now $-|k_2^*|$ to end up in $\mathcal{O}(\mathrm{NP})$ from where NP can be reached, by local control at NP (Theorem 3.2). This concludes the proof of global controllability for our system living on S^2.

It is important to realize that it has not been assumed that the X_2-orbit through NP intersects $\mathring{C}(x_e \rightarrow)$. Also notice that not the whole reachability set $C(\rightarrow x_e)$ must be known. Local information (in the neighborhood of x_e) on the reachability set would suffice, since for some T, $\{x_2(t, \mathrm{SP}), t > T\}$ belongs to a neighborhood of x_e anyway (the smaller the neighborhood, the larger T). For more details, see Remark 5.1.

In order to give a proof in the general case, a few remarks are useful. It is noticed that in the working example the equilibrium points SP and NP of X_1 and x_e of X_2 could have been of a different type, without affecting the proof. Essential for controlling SP to NP is that:

1. The rank condition at x_e (equilibrium point of X_2) is satisfied, i.e. there exists a set $C(\rightarrow x_e)$ with interior (see Theorem 3.1).

2. Both equilibrium points NP and SP of X_1 belong to $W^s(x_e) \cup W^u(x_e)$ (manifolds considered with respect to X_2).

3. The orbit of X_2 through SP, and $\mathring{C}(\rightarrow x_e)$ have nonempty intersection.

If 1, 2, and 3 are valid, we say, by way of definition, that SP is controllable to NP, by *turning at* x_e. For the general system $\dot{x} = X_1 + uX_2$, we will say that sa^k (equilibrium point of X_1) is controllable to sa^l (equilibrium point of X_1), by turning at p (equilibrium point of X_2) if, mutatis mutandis 1., 2., and 3. are satisfied. The possibility to control sa^k to sa^l by turning at p will be designated on the "vertices-arrows" diagram of X_1, by a new *arrow with label "p"*.

Assume that for the system $\dot{x} = X_1 + uX_2$, on the diagram representing X_1, there exists a path from each sink to each source, consisting of *arrows and labeled arrows, then we say that the diagram of* X_1 *is complete* (*with respect to* X_2), or, *has been completed by* X_2. Given completeness of the diagram of X_1 with respect to X_2, a repeated application of the techniques described above leads to our main theorem.

THEOREM 4.1. *The system* $\dot{x} = X_1 + uX_2$, *defined on a compact manifold* M, $u \in \mathbb{R}$, *is globally controllable with piecewise constant control if*:

1. *The rank condition at each equilibrium point of* X_1 *is satisfied.*

2. *The diagram of* X_1 *is complete with respect to* X_2.

COROLLARY 4.2. *The system* $\dot{x} = X_1 + uX_2$ *defined on a compact manifold* M, $u \in \mathbb{R}$ *is globally controllable by bang-bang control, under the conditions specified in Theorem* 4.1.

Proof. The arguments instrumental in the proof of global controllability of the system $\dot{x} = X_1 + uX_2$ have been based on the following control procedures:

1) local controllability in the neighborhood of the equilibrium points of X_1 (as a consequence of the rank condition assumption at the equilibrium points of X_1) and local accessibility at some of the equilibrium points of X_2 (as a consequence of the rank condition assumption at the equilibrium points of X_2);

2) letting $u = 0$ so that the dynamical properties of X_1 are brought to bear;

3) letting $u = k$ or $-k$, with k large enough to let the dynamical properties of X_2 or $-X_2$ have their impact. We will show that there are two numbers l and $-l$ with $l \in \mathbb{R}$ and large enough so that the control laws referred to in 1), 2) and 3) can be replaced by control laws only taking the values l and $-l$.

As for the local controllability arguments near equilibrium points of X_1 and X_2 care has been taken to only use bang-bang controls. The control values have not been specified, since not relevant, except for the necessity of a positive as well as a negative control value. Therefore take any number l_1; all local controllability arguments can be made to work with control laws having values l_1 and $-l_1$.

By the rank condition at equilibrium points of X_2, there exist accessibility sets from where the equilibrium points can be reached by piecewise constant controls. These sets are dependent on the control values allowed: it follows from [8] that the accessibility set corresponding to controls with values s or $-s$ *contains* the accessibility set corresponding to controls with values r and $-r$ when $|s| > |r|$. The same remarks are valid for controllability sets at equilibrium points of X_1. Therefore, whenever an equilibrium point of X_1 is being controlled (by using a large control, say k or $-k$) to the accessibility set of an equilibrium point of X_2 and then to the equilibrium point of X_2 itself (by local controllability with controls l_1 and $-l_1$), this could have been realized by using controls k and $-k$ (if $k > l_1$) or by controls l_1 and $-l_1$ (if $k < l_1$).

The proof of global controllability also requires the use of high-valued controls at other occasions. It should be clear from the remarks above that by taking the highest (denoted by l) of all control values encountered in the proof, one can carry out the local controllability arguments and the global controllability arguments (as far as restricted to 3) by using controls with values l and $-l$.

There is still one problem left (see 2): do control values l and $-l$ suffice to "simulate" a control identical to zero? That it does, is a consequence of a well-known theorem [17, p. 11] which says that—since $2X_1 = (X_1 + lX_2) + (X_1 - lX_2)$—the orbit corresponding to X_1 can be *approximated* as closely as required by switching between $X_1 + lX_2$ and $X_1 - lX_2$. It is further remarked that, in establishing global controllability of $\dot{x} = X_1 + uX_2$, whenever u is set equal to zero and thus the trajectories of X_1 are followed, approximate trajectories would do as well.

We have now shown that in controlling $m \in M$ to $n \in M$, there exists a number $l \in \mathbb{R}$ such that m can be controlled to n with control values k and $-k$, $|k| > |l|$. By continuity of solutions of differential equations with respect to initial conditions and by local bang-bang controllability at equilibrium points of X_1, there exist neighborhoods of m and n that can be controlled to each other. (In controlling m to n, the last step consists in controlling a point in a neighborhood $\mathcal{O}$ of some equilibrium point of X_1 to n; thus one can also reach arbitrary points in a suitable neighborhood of n, by local bang-bang controllability at the equilibrium point of X_1 under consideration.) By compactness of $M \times M$ it follows that there is a largest K such that $\dot{x} = X_1 + uX_2$ is globally bang-bang controllable with control values k and $-k$ with $|k| > K$.

5. Discussion.

5.1. Here we would like to discuss the controllability condition to some extent. Point 2 in the statement of Theorem 4.1 is a sufficient condition that guarantees the possibility of controlling the sinks of X_1 to the sources of X_1. Other procedures can be imagined. Consider the following one as a possible replacement of point 2 of Theorem 4.1 (we will again restrict to the system on S^2 defined earlier). Let $x_0 \in M$ be such that $X_1(x_0) \neq 0$. Then the sink si(of X_1) is controllable to the source so (of X_1) if

i) we have local accessibility at x_0 (i.e. $C(\rightarrow x_0)$ and $C(x_0\rightarrow)$ have nonempty interior);
ii) the orbit of X_2 through si intersects $\mathring{C}(\rightarrow x_0)$ (interior of the accessibility set towards x_0);
iii) the orbit of X_2 through so intersects $\mathring{C}(x_0\rightarrow)$ (interior of the accessibility set from x_0).

This is in fact the "obvious" global controllability condition. Controlling si to so would then be done as follows: control si to $\mathring{C}(\rightarrow x_0)$ (by ii) and then to x_0 and $\mathring{C}(x_0\rightarrow)$ from where, one can reach so (by iii).

The main point of § 4 is that letting x_0 be an equilibrium point of X_2 yields some substantial benefits. Indeed, condition iii) turns out to be superfluous (as has been shown in § 4). This is important: not only does this provide a weaker sufficient condition but also the nonpresence of condition iii) allows extending the ideas of § 4 to classes of systems on "non"-compact manifolds as will be illustrated in § 5.4. In addition, conditions ii) and iii) are strong when taken together: they seem unlikely to be satisfied in general as they relate the trajectories of a vectorfield (X_2) through two unrelated points si and so. In other words, only for particular points x_0, the relation expressed by ii) and iii) has a chance to be satisfied. Checking it would require a whole lot of information on the flow of X_2 on the (*global*) reachability set of x_0 (let alone the problem of picking candidates for x_0). On the other hand, in the condition described in § 4, local information (in the neighborhood of x_0) on the reachability set $C(\rightarrow x_0)$, with x_0 an equilibrium of X_2, is enough, since after some time, the trajectory of X_2 through SP is in the neighborhood of x_0 anyway.

5.2. Here we try to give an evaluation of the position of Theorem 4.1 with respect to other global controllability results for systems with a drift term (for symmetric systems the standard Lie-algebraic condition can be found in [13]). It is first remarked that the controllability literature contains few results on global controllability. There is the result of Jurdjevic and Sussmann [7] for systems on Lie groups, which in fact reduces to conditions for local controllability in the neighborhood of the identity element of the Lie group. There is also an example in Brockett [3] where global controllability is derived from conditions on the drift term, like the periodicity of all its trajectories or denseness of its "Poisson-stable" points. In addition we mention the approaches of Hunt [14] and R. M. Hirschorn [15].

It seems natural, that in constructing global controllability for the system $\dot{x} = X + uY$ (X and Y are C^∞-vectorfields with no additional restrictions) some interaction of the vectorfields X and Y is necessary. For linear systems $\dot{x} = Ax + Bu$ this interaction is expressed by rank $[B, AB, \cdots, A^{n-1}B] = n$. In this paper it is shown that for systems of the form $\dot{x} = X_1 + uX_2$ an interaction at the local level (a number of rank conditions) and an interaction at the global level (the completeness condition) imply controllability. It is stressed that the interaction expressed by the completeness condition requires a minimal amount of information on the flows of X_1 and X_2.

5.3. Linear systems $\dot{x} = Ax + Bu$, $y = Cx$, $x \in \mathbb{R}^n$ exhibit a duality between controllability and observability. The system (A, B) is controllable if rank $[B, AB, \cdots, A^{n-1}B] = n$; the system (A, C) is observable if rank $[C', A', C', \cdots, (A')^{n-1}C'] = n$. This shows that the system $\dot{x} = Ax + Bu$, $y = Cx$ with $A = A'$ and $B = C'$ is controllable if and only if it is observable. In order to write this ("symmetric") system as a gradient system, we introduce the function $V(x, u) = 1/2x'Ax + x'bu$, with $A = A'$, u scalar and b an n-vector (we have restricted the discussion to systems with scalar input and output). Let $\text{grad}_x V$ and $\text{grad}_u V$ denote

the gradient of V with respect to x and u respectively. Then the system $\dot{x} = \operatorname{grad}_x V$, $y = \operatorname{grad}_u V$ is controllable if and only if it is observable. One might be tempted to jump to similar conclusions for systems derived from nonlinear functions $W(x, u) = f(x) + ug(x)$, $x \in M$, $u \in \mathbb{R}$. Is $\dot{x} = \operatorname{grad}_x W (= \operatorname{grad}_x f + u \operatorname{grad}_x g)$, $y = \operatorname{grad}_u W$ $(= g(x))$—gradients taken with respect to some Riemannian metric—observable if and only if it is controllable? We will try to answer this question immediately, but not before mentioning that it contains some typical nonlinear aspects, like dependence of observability on the inputs; it can be argued that—because of the existence of universal inputs—these aspects play a secondary role and are therefore discarded.

It is noticed that as far as local controllability and local observability are concerned, in the neighborhoods of equilibrium points of $\operatorname{grad}_x f$, some sort of duality can be maintained. Indeed, local controllability and observability are satisfied if a rank condition and a dual rank condition are satisfied, by the present paper and by [1]. *Global duality cannot be extrapolated.* In [1], it has been shown that global observability on compact manifolds can be achieved without a global observability condition and hence the duality breaks down, since for global controllability a global condition is necessary (see Theorems 4.1 and 5.1).

5.4. We have illustrated and proved a global controllability result for control systems $\dot{x} = X_1 + uX_2$ defined on compact manifolds. Let us quickly recapitulate the control procedure for the system considered earlier, defined on S^2. In order to control "a" towards "b", both "a" and "b" arbitrary points on S^2, we have first controlled "a" towards SP (eq. point of X_1), then into a neighborhood of x_e (eq. point of X_2), from there to NP and finally to "b". It is this particular control procedure that has been generalized towards more general systems on compact manifolds. However, it should be remarked that for the system on S^2 we could have "stopped" at x_e, from where "b" could have been reached, i.e. we did not have to go to a neighborhood of NP to finally reach "b". This remark makes it possible to extend our theory to classes of systems on noncompact manifolds. At the moment, we will suffice with the following example. Consider the system $\dot{x} = Ax + u(Bx - b)$ with $x \in \mathbb{R}^2$, B nonsingular, $u \in \mathbb{R}$. Assume that A be a sink and that B has complex conjugate eigenvalues (real part $\neq 0$). Furthermore assume we have local controllability at the origin and local accessibility at $B^{-1}b$. We claim that this system is globally controllable. First remark that by the eigenvalue property of B, the origin can be controlled to the accessibility region of $B^{-1}b$ and thus also to $B^{-1}b$. To show that any point m can be controlled to any point n, repeat the proof outlined for the system on S^2, except for "stopping" at $B^{-1}b$ (as explained above): the role of SP and x_e is played respectively by the origin and $B^{-1}b$ (there is no equivalent of NP in the example under discussion).

Actually the drift term Ax does not have to be a sink. Let Bx be as above. Without loss of generality assume that the eigenvalues of B have negative real part. Assume that the drift term $f(x)$ has an equilibrium point at the origin and is such that local controllability at the origin and local accessibility at $B^{-1}b$ is satisfied for the system $\dot{x} = f(x) + u(Bx - b)$. We claim that this system is globally controllable with no further assumptions on $f(x)$. Indeed, for the same reason as above the origin can be controlled to $B^{-1}b$. Also $B^{-1}b$ can be controlled to the origin (for the negative system the origin can be controlled to $B^{-1}b$, therefore $B^{-1}b$ can be controlled to the origin for the original system). In order to control an arbitrary point m to an arbitrary point n the following control procedure is applied. First control m to a point m_1 close to $B^{-1}b$ (by taking u large and positive). From there m_1 is controlled to a point m_2 in a neighborhood of the origin. By local controllability at the origin it is controlled to a

point m_3 in the neighborhood of the origin such that the control sending the origin to $B^{-1}b$ sends m_3 to a point m_4 in the neighborhood of $B^{-1}b$ which by u large and negative is then finally sent to n.

Except for an extension of the theory to systems defined on noncompact manifolds, it should also be clear that including (hyperbolic) *periodic orbits* in X_1 and X_2 next to equilibrium points (and thus extending the analysis to Morse-Smale vectorfields) adds no fundamental difficulties [1].

5.5. The next point to investigate might consist of studying controllability of $\dot{x} = X_1 + uX_2$, $x \in M$ compact, X_1 and X_2 Morse-Smale, with *bounds on the control.* It is clear that for this problem, information on the dependence of the phase portrait of $X_1 + uX_2$ as a function of u (within its bounds) is helpful. On the other hand, systems of the form $\dot{x} = X_1 + \sum_{i=1}^{k} u_i X_i$, $x \in M$ compact, all X_i Morse-Smale, with more than two controls, all of them *unbounded,* are controllable if only a number of *local* conditions are satisfied. Indeed, such a system is globally controllable if dim $\mathscr{L}\{X_2, \cdots, X_k\}(X_e) = n$ at each equilibrium of $X_2, \cdots, X_k$. Therefore in this case, bounds on the controls are needed in order to have a nontrivial problem.

Part II. An application: controllability by means of two vectorfields.

1. Introduction. Part II is concerned with the number of smooth vectorfields, defined on a smooth connected paracompact manifold, necessary to achieve global controllability. A set of vectorfields is globally controllable, if for any two points m_1 and m_2 on the manifold, there exists a trajectory controlling m_1 to m_2. We will not dwell on the technical definitions and terminology involved and refer the reader to [1] and [20] for more specifics.

The problem has quite an interesting history. From the results of [21], it follows that on each smooth connected paracompact manifold, a set of *four* vectorfields exists constituting a globally controllable system; the arguments used in the proof require the vectorfields to have a differentiability degree which is related to the dimension of the manifold. From the work of Sussmann [23], it follows that this differentiability requirement can be weakened. Sussman then proved in [22], by means of different methods, how to bring the number down to *three.* Finally the problem was settled in [20] where it has been shown that on each smooth connected paracompact manifold a globally controllable set of *two* smooth vectorfields exists. The result consists of a careful construction which—although not claimed by the authors—seems to be unique in some sense; i.e., it produces a controllable set of two vectorfields but seems not to leave much choice as to the vectorfields involved.

We want to present an alternative approach to the result of Levitt and Sussmann, which shows that on each smooth connected paracompact manifold there exists a *large class* of systems $\dot{x} = X + uY$, globally controllable by bang-bang control; this obviously implies controllability by means of two vectorfields. It is remarked that stronger differentiability requirements and more switchings are needed than in [20], but in return, we are able to construct a *variety of basically different systems,* controllable by bang-bang control. The construction is based on the theory developed in part I, which will be referred to when necessary.

The organization of Part II is as follows. In § 2 we associate with each C^∞-vectorfield $X = -\text{grad}\, f$, (f belonging to an open and dense set of $C^\infty(M, \mathbb{R})$), defined on a compact manifold, a C^∞-vectorfield Y such that $\dot{x} = X + uY$ is globally controllable. Then we show that this system is *globally controllable with two-valued controls.* In § 3 we show that for systems $\dot{x} = X_1 + uX_2$, X_1 Morse-Smale with no periodic orbits,

there exist Morse-Smale vectorfields X_2 such that the system above is bang-bang controllable. In § 4 the results are extended to noncompact manifolds.

2. Global controllability on compact manifolds. Assume $X = -\text{grad} f$ (gradient taken with respect to some Riemannian metric) with f a C^∞-Morse function [18] defined on a compact manifold M and such that the Hessian of f at its critical points is cyclic [24]. Such functions exist and they are in fact open and dense in $C^\infty(M, \mathbb{R})$. Indeed, pick a Morse function which is slightly altered if necessary in the neighborhood of the critical points so as to have a cyclic Hessian. The alteration might consist in adding the product of a quadratic form with small coefficients and cyclic Hessian together with a smoothing function. This proves the density; openness is obvious.

Under the assumptions above, X can always be C^1-approximated by a new vectorfield (also denoted by X), such that X is Morse-Smale [16] having cyclic derivatives at its equilibrium points. We will now construct a C^∞-vectorfield Y such that $\dot{x} = X + uY$, u scalar, is globally controllable. It is remarked that the following construction of Y is just one of the different possibilities that can be imagined. For an alternative, see the following section.

It is proved in § 4 (Part I) that a necessary and sufficient condition for global controllability is that all sinks of X must be controllable to all sources of X. This will be accomplished by constructing Y such that it has an equilibrium point p, which is a *turning point* (Part I) connecting sinks of X with sources of X. The vectorfield Y is—roughly stated—a compound of a number of vectorfields. First, a smooth vectorfield is defined in the neighborhood of p. Then a number of "tracks" are mapped out from all sinks and sources of X towards p, and on a neighborhood of each track a smooth vectorfield is defined. The rest of the manifold carries the zero-vectorfield. The neighborhoods above overlap partially. As a consequence, at some points of M, more than one vector is defined. These vectorfields considered together, define a *discontinuous* vectorfield Y_{disc} (after adopting a reasonable selection rule at the points of M with two vectors defined). If, in the construction of Y_{disc} care has been taken that the rank conditions and the completeness condition (part I) are satisfied, then the system $\dot{x} = X + uY_{\text{disc}}$ is globally controllable. Of course a *smooth* vectorfield Y is needed and this will be realized by means of a partition of unity argument. A technical problem in the construction of the vectorfields constituting Y_{disc} is therefore that the *completeness condition must persist* when smoothing these vectorfields into a vectorfield Y.

The formal construction of Y goes as follows. Pick a point p different from the critical points of X. Consider an open neighborhood $U(p)$ of p, which does not cover critical points of X. On $U(p)$ a C^∞-function g with one minimum at p and with no other critical points is defined. It is also assumed that p is a nondegenerate critical point of g. Consider the vectorfield on $U(p)$ defined by $Y|_{U(p)} = -\text{grad}\, g$ (gradient is taken with respect to some metric). Assume that the rank condition at p is satisfied. (If it is not, one can always make sure that it is, by locally adding to g a well-chosen small quadratic form with one minimum at p, see the first paragraph of this section).

Let $l \in \mathbb{R}$ be such that $g^{-1}(l)$ is nonempty. Let $g^{-1}(< l)$ be the set points of M, having a g-value smaller than l. Consider a neighborhood $V(p)$ of p properly contained in $g^{-1}(< l)$. By the local rank condition at p there exists an open subset Acc (p) (not necessarily being a neighborhood of p) of $V(p)$ such that any point in Acc (p) can be controlled to p along the trajectories of $-\text{grad} f|_{U(p)} + uY|_{U(p)}$ without leaving $V(p)$. Let $y(\cdot, \cdot)$ be the flow corresponding with $Y|_{U(p)}$. Consider the set $\{y(t, x)\}$: $t < 0$, $x \in \text{Acc}\,(p)$ and the expression is defined}. Take the intersection G of this set with $g^{-1}(l)$. By construction of $Y|_{U(p)}$, this intersection is nonempty and open with respect

to $g^{-1}(l)$. For each sink (si) of X consider an imbedding $\text{imb}_{(si)}: (0,1) \to M$ such that:

1) (si) belong to the image of the imbedding and no other equilibrium points do; for different sinks the corresponding images of the imbeddings are pairwise disjoint.

2) the image of $\text{imb}_{(si)}$ intersects the interior of G and has no points in common with $V(p)$;

3) its velocity field, together with X, satisfies the rank condition at (si); (possible by cyclicity of X).

4) its velocity field equals the field $Y|_{U(p)}$ for all points on the image of $\text{imb}_{(si)}$ in $g^{-1}(\leqq l)$.

The velocity field on $\text{imb}_{(si)}$ is now extended to a smooth vectorfield on an open set $U(\text{imb}_{(si)})$ image of the imbedding; this open set does not cover any new equilibrium points—other than (si)—of X and has no points in common with $V(p)$. The extension of the velocity field is actually carried out by first considering ε, such that $\text{imb}_{(si)}[\varepsilon, 1-\varepsilon]$ contains (si) and points of $g^{-1}(< l)$ and then extending to a neighborhood $U(\text{imb}_{(si)})$ of $\text{imb}_{(si)}[\varepsilon, 1-\varepsilon]$ by means of a partition of unity argument [19, p. 29]. This construction is repeated for each sink $(si)_j$ *and each source* $(so)_j$ of X, except that for the sources G could have been replaced by $g^{-1}(l)$.

Consider an open set O which has no points in common with the images of the imbeddings above, neither with $V(p)$. But O is chosen such that $O \cup (\cup_j (U(\text{imb}_{(si)_j}))) \cup (\cup_j (U(\text{imb}_{(so)_j}))) \cup g^{-1}(< l)$ covers M. Let O carry the zero vectorfield; the other open sets of the covering carry the vectorfields defined above.

We will now smooth these locally defined vectorfields into a smooth vectorfield Y defined on M. This will be accomplished by a partition of unity argument. The construction of the locally defined vectorfields has been carried out (points 2 and 4) such that this smoothing process will not destroy the crucial properties leading to p being a turning point. Consider a smooth partition of unity subordinate to the covering defined above. Define $Y(x)$ as the sum of each locally defined vectorfield evaluated at x, multiplied with a "weight" according to the partition of unity. The vectorfield Y is smooth, satisfies the rank conditions at the sinks and sources of X and at p, and by the construction above completes the diagram of X. Therefore the system $\dot{x} = X + uY$ is globally controllable on M, by Theorem 4.1 of Part I.

The control system $\dot{x} = X + uY$ constructed above is actually globally controllable by means of *two* controls, i.e. there exist real numbers $l \in \mathbb{R}$ such that any two points of M can be connected along the trajectories of the vectorfields $X + lY$ and $X - lY$. This follows from the proof of Corollary 4.2 of Part I. Indeed, although X and Y are not Morse-Smale, the arguments used in the proof of Corollary 4.2 carry over and thus global controllability by bang-bang control is implied by global controllability.

3. Morse–Smale systems are globally bang–bang controllable by Morse–Smale systems.

THEOREM 3.1. *Let X_1 be a Morse-Smale vectorfield (defined on a compact manifold) with no periodic orbits and with at least one equilibrium point q (q not a source) which is cyclic (i.e. the derivative of X_1 at q is cyclic). There exists a Morse-Smale vectorfield X_2 on M such that $\dot{x} = X_1 + uX_2$ is bang-bang controllable.*

Proof. Let p be a point on the stable manifold of q with respect to X_1. Let p be close enough to q such that there exists an open ball $U(p)$ around p which covers q but does not cover any other critical points of X_1. The point p will turn out to be the *only sink* of the Morse-Smale system X_2 on M, to be constructed. It is remarked that the construction that follows is just one of many alternatives possible.

It is well known [20], [25] that there exists a C^∞-Morse function f on M which has one minimum at p, which at different critical points of f, takes different values and for which $M_a = \{m \in M, f(m) \leqq a\}$ is compact, for every real a. In addition its critical points are different from the equilibrium points of X_1. Let $c_0 = p, c_1, c_2, \cdots$ (finite in number) be the critical points of f arranged so that $f(p) < f(c_1) < f(c_2) < \cdots$. Consider the vectorfield $-\text{grad } f$. This vectorfield will now be changed into X_2. Consider neighborhoods $U(c_i)$ of c_i, $i \neq 0$, such that $f(U(c_i)) < f(c_{i+1})$ and such that X_1 restricted to $U(c_i)$ does not leave the stable manifold of c_i corresponding to $-\text{grad } f$ invariant. This can be accomplished as follows. Take $U(c_i)$ small enough such that in local coordinates X_1 restricted to $U(c_i)$ is a rectilinear vectorfield ($X_i(c_i) \neq 0$!); if X_1 would leave the stable manifold invariant, change locally such that the stable "eigenspace" of $D(-\text{grad } f)$ of the local representation of $-\text{grad } f$ at c_i is no longer invariant. This altered vectorfield will still be denoted by $-\text{grad } f$.

A second, and more crucial alteration will now be made on the restriction of $-\text{grad } f$ to $U(p)$. We assume that $U(p)$ is small enough so that $-\text{grad } f$ has only one critical point p(a sink) in $U(p)$. Assume that the Hessian of f at p is cyclic and that the rank condition for X_1 and $-D \text{ grad } f$ is satisfied at p (see § 2 if it is not). Therefore there exists an accessibility region $\text{Acc}(\to p)$ towards p. Take an imbedding of the open unit interval into $U(p)$ which connects q to $\text{Acc}(\to p)$ as explained in the previous section. Define $U(\text{imb}_{(q)}) \subset U(p)$ and $V(p)$ as in the previous section where $U(\text{imb}_{(q)})$ now carries a vectorfield $-\text{grad } g$ which is constructed so that it is tangent to the image of the imbedding and so that the rank condition at q is satisfied. The latter is possible by cyclicity of X_1 at q_1.

Now consider an open set O which has no points in common with the image of the imbedding, neither with $V(p)$ and such that $O \cup U(\text{imb}_{(q)})$ equals $U(p)$. Let $-\text{grad } f$ be defined on O and let $-\text{grad } g$ be defined on $U(\text{imb}_{(q)})$. Smoothing the functions f and g by means of a smooth partition of unity subordinate to the cover of $U(p)$ and taking minus the gradient results in our vectorfield X_2, when restricted to $U(p)$. It is remarked that the partition of unity cannot introduce new equilibrium points in X_2 if we assume that the vectorfield $-\text{grad } g$ on the image of the imbedding crosses the level surfaces of f in the "lower f-value direction". Outside $U(p)$, X_2 is defined by $-\text{grad } f$. It is remarked that X_2 is a gradient vectorfield and that gradient vectorfields can always be C^1 approximated by Morse-Smale vectorfields which leave the vectorfield unchanged in the neighborhoods of critical points [16]. Therefore X_2 (keeping notation) might be assumed to be Morse-Smale.

We will now prove that the system $\dot{x} = X_1 + uX_2$ is bang-bang controllable. Let $C_1(q)$ be an open neighborhood of q which is locally controllable. Notice that p can be controlled to q ($u = 0$ and local controllability at q). Let $O_1(p)$ be a neighborhood of p small enough such that the control mapping p to q maps $O_1(p)$ into $C_1(q)$. By construction of the system $\dot{x} = X_1 + uX_2$ there is a control which steers q to p. Let $C(q) \subset C_1(q)$ be such that this control maps $C(q)$ onto $O(p) \subset O_1(p)$. We claim that any two points a, b of $O(p)$ can be controlled to each other. Indeed first "a" is controlled into $a_1 \in C(q)$ by the control sequence steering p to q; then a_1 is controlled towards $a_2 \in C(q)$ by local control at q, with a_2 such that the control sequence controlling q to p, controls a_2 to b.

Take two arbitrary points m and n on M not in $O(p)$. Assume we know how to control m towards a point "a" in $O(p)$ along the trajectories of the system $\dot{x} = X_1 + uX_2$. A similar control procedure controls n to some point $b \in O(p)$ along the trajectories of the negative system; therefore there is a control sequence that controls some point $b \in O(p)$ towards n, along the trajectories of the original system. But "a" can be

controlled to b (previous paragraph), thus showing how to control m to $O(p)$ ends the proof of global controllability.

If m belongs to the stable manifold of p with respect to the vectorfield X_2, then a control K, positive and large enough controls m to $O(p)$. In the following, we can assume that m is not a critical point of X_2 (if it is, apply $u=0$). Assume m belongs to the stable manifold of a critical point c_i different from p. Take a control, positive and large enough such that m is controlled to some point m_1 in $U(c_i)$. Apply $u=0$ for some time such that m_1 is controlled to a point m_2 which is off the stable manifold of c_i and such that $f(m_2)<f(c_{i+1})$. Then m_2 belongs to the stable manifold of a critical point c_j with $j<i$. Repeating this procedure, one finally arrives in $O(p)$. This concludes the proof of global controllability.

To show global *bang-bang* controllability we recall that as is shown in Corollary 4.2 of Part I, in controlling m to n we could have used control values k and $-k$ (the trajectory $u=0$ is approximated by switching between k and $-k$ where k is the largest control value applied). By continuity of solutions of differential equations with respect to initial conditions and by local bang-bang controllability at q there exist *neighborhoods of m and n* that can be controlled to each other by bang-bang control with values $+k$ and $-k$. Repeating this argument for all couples $(m, n)\in M\times M$ and by compactness of $M\times M$ there exists a finite number K such that $\dot{x}=X_1+uX_2$ is controllable with control values l, $-l$ with $l\geqq K$.

4. Bang–bang controllability on noncompact manifolds. We offer an alternative to the paper by Levitt and Sussmann [20] on controllability by means of two vectorfields. We will construct two vectorfields X_1 and X_2 such that $\dot{x}=X_1+uX_2$ is bang-bang controllable. Again the vectorfields that we will be considering are one of many choices possible. Our construction is inspired by the theory of the previous section.

Let $X_2=-\text{grad}\, f$, f Morse, be a vectorfield with one sink p. This time, since M is noncompact, X_2 might have an infinite number of equilibrium points. Both X_1 and X_2 have the same characteristics on $U(p)$ (for notation see previous section) as in the previous section. This time however X_1 is smoothed out equal to zero, outside of $U(p)$ except for small isolated neighborhoods of the critical points of X_2 where X_1 is a rectified parallel vectorfield smoothed off to zero, that maps points on the stable manifold of X_2 out of the stable manifolds.

It follows from § 3 that $\dot{x}=X_1+uX_2$ is globally controllable by bang-bang control. In controlling m to n, the trajectories that connect m with n are in part located inside $U(p)$ and in part outside of $U(p)$. In controlling m to n one can use control values $+1$ and -1 (or any other values, one positive and one negative) as long as one is outside of $U(p)$, since $X_1=0$ (except for the neighborhoods of critical points of X_2). Once arrived in $U(p)$, since $U(p)$ is a subset of a compact set, there exist control values k and $-k$ that implement the required control strategy as explained before. Bang-bang global controllability is an immediate consequence.

5. Remarks.

5.1. The rank conditions require X_1 and X_2 to be of sufficiently high differentiability class, a condition not necessary in the theorem of Levitt and Sussmann [20]. Notice that by [20], X_1 and X_2 can be taken to be real analytic.

5.2. The methods in [20] set a uniform bound on the number of switchings—depending on the dimension of the manifold. We are unable to do so because X_1 is approximated by X_1+lX_2 and X_1-lX_2, with a number of switchings depending on X_1 and X_2 themselves.

Acknowledgment. The author acknowledges several interesting comments made by an anonymous referee.

REFERENCES

[1] D. AEYELS, *Global observabiility of Morse-Smale vectorfields*, J. Differential Equations, 45 (1982), pp. 1-15.

[2] ———, *Generic observability of differentiable systems*, this Journal, 19 (1981), pp. 595-603.

[3] R. BROCKETT, *Nonlinear systems and differential geometry*, Proc. IEEE, 64 (1976), pp. 61-72.

[4] R. HERMANN, *On the accessibility problem in control theory*, International Symposium, Nonlinear Differential Equations and Nonlinear Mechanics, Academic Press, New York, 1963, pp. 325-332.

[5] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 728-740.

[6] M. HIRSCH AND S. SMALE, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic Press, New York, 1974.

[7] V. JURDJEVIC AND H. J. SUSSMANN, *Control systems on Lie groups*, J. Differential Equations, 12 (1972), pp. 313-329.

[8] A. J. KRENER, *A generalization of Chow's theorem and the bang-bang theorem to nonlinear control problems*, this Journal, 12 (1974), pp. 43-52.

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

[10] S. SMALE, *Differentiable dynamical systems*, Bull. Amer. Math. Soc., 73 (1967), pp. 747-817.

[11] H. J. SUSSMANN, *Some properties of vector fields that are not altered by small perturbations*, J. Differential Equations, 20 (1976), pp. 292-315.

[12] H. J. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.

[13] C. LOBRY, *Controllabilité des systèmes non linéaires*, SIAM J. Control, 8 (1970), pp. 573-605.

[14] L. R. HUNT, *Controllability of general nonlinear systems*, Math. Systems Theory, 12 (1979), pp. 361-370.

[15] R. M. HIRSCHORN, *Global controllability of nonlinear systems*, this Journal, 14 (1976), pp. 700-711.

[16] S. SMALE, *On gradient dynamical systems*, Annals of Mathematics, 74 (1961), pp. 199-206.

[17] J. DIEUDONNÉ, *Treatise on Analysis, Vol.* IV, Academic Press, New York, 1974.

[18] M. GOLUBITSKY AND V. GUILLEMIN, *Stable Mappings and Their Singularities*, Graduate Texts in Mathematics, vol. 14, Springer, New York, 1974.

[19] F. WARNER, *Foundations of Differentiable Manifolds and Lie Groups*, Scott, Foresman and Company, Glenview, IL, 1971.

[20] N. LEVITT AND H. SUSSMANN, *On controllability by means of two vectorfields*, this Journal, 13 (1975), pp. 1271-1281.

[21] C. LOBRY, *Une propriété générique des couples de champs de vecteurs*, Czechoslovak Math. J., 22 (1972), pp. 230-237.

[22] H. SUSSMANN, *On the number of directions needed to achieve controllability*, this Journal, 13 (1975), pp. 414-419.

[23] ———, *Some properties of vectorfields that are not altered by small perturbations*, J. Differential Equations, 20 (1976), pp. 292-315.

[24] M. WONHAM, *Linear Multivariable Control*, Lecture Notes in Economics and Mathematical Systems 101, Springer, New York, 1974.

[25] J. MILNOR, *Lectures on the h-cobordism Theorem*, Princeton Univ. Press, Princeton, NJ, 1965.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

008

MEASURABLE SELECTION THEOREMS FOR MINIMAX STOCHASTIC OPTIMIZATION PROBLEMS*

ANDRZEJ S. NOWAK†

Abstract. This paper resolves the measurability questions which arise in the analysis of various minimax stochastic optimization models posed in Borel spaces. For a large class of such models, we provide a set of sufficient conditions which ensure that the questions have positive answers. These conditions are essentially weaker than those described in the existing literature. Moreover, they allow us to conclude a fundamental theorem yielding some new results on the existence of universally measurable selections of extrema for both zero-sum stochastic games and minimax stochastic optimal control problems. In particular, a random version of the well-known minimax theorem of Fan is in this way established. The fundamental theorem yields also a descriptive set theoretic fact concerning the projections of Borel sets. The paper indicates also some counterexamples to possibly more general problems.

Key words. universally measurable selections, zero-sum stochastic games, minimax stochastic optimal control, projections of Borel sets, uniformization of analytic and coanalytic sets

1. Introduction. The aim of this paper is to study some measurability issues which naturally arise in various minimax stochastic optimization problems in Borel spaces. The basic tools of this paper are the fundamental facts concerning Borel and analytic sets in complete separable metric space and some known uniformization (selection) theorems. Also some results concerning probability measures on Borel spaces are employed. For a detailed discussion of most of the facts (with proofs) that we need the reader is referred to the excellent book by Bertsekas and Shreve (see [4, Chap. 7 and Appendix B]).

The notation and basic definitions of this paper are as follows. We use R to denote the real line and R^* to denote the extended real line, i.e., $R^* = R \cup \{-\infty, \infty\}$. The set of positive integers is denoted by N. If X and Y are sets, then proj_X is the projection mapping from $X \times Y$ to X. Let X be a topological space. If there exists a Polish (i.e., complete separable metric) space Y and a Borel subset B of Y such that X is homeomorphic to B, then X is said to be a *Borel space.* It follows that a Borel space X is metrizable and separable. A separable metric space X is said to be an *analytic space* if X is the continuous image of a Polish space. It is known that every Borel space is analytic. An *analytic set* in a Borel space X is a subset of X which is an analytic space when endowed with the relative topology. It is known that the finite or countable union, intersection, and product of analytic (Borel) sets is analytic (Borel) [4, Chapter 7]. The complement of an analytic set relative to a Borel space is called a *coanalytic set.*

Let X be a Borel space. We denote by $\mathcal{B}_X$ the Borel σ-algebra of X. In addition to $\mathcal{B}_X$, we are interested in two more σ-algebras in X. The *universal σ-algebra,* denoted by $\mathcal{U}_X$, is the σ-algebra of all universally measurable subsets of X. The *limit σ-algebra,* denoted by $\mathcal{L}_X$, is the smallest σ-algebra containing the Borel subsets of X and closed under operation (A) (Suslin operation). It is known that $\mathcal{B}_X \subset \mathcal{L}_X \subset \mathcal{U}_X$ (see [4]).

Let X and Y be Borel spaces. A function $f: X \to Y$ is called *limit* (*universally*) *measurable* if $f^{-1}(B) \in \mathcal{L}_X$ ($f^{-1}(B) \in \mathcal{U}_X$) for every $B \in \mathcal{B}_Y$. Clearly, if f is limit measurable then it is universally measurable. An extended real-valued function $f: X \to R^*$, is called *upper semianalytic* (u.s.a.) if the set $\{x \in X: f(x) \geqq c\}$ (equivalently, the set

* Received by the editors March 6, 1984.

† Institute of Mathematics, Technical University of Wrocław, Wyspiańskiego 27, 50-370 Wrocław, Poland.

$\{x \in X: f(x) > c\}$) is analytic for each real number c. (A function $f: X \to R^*$ is called *lower semianalytic* if $-f$ is upper semianalytic.) It is known that every Borel measurable function is u.s.a., and every u.s.a. function is limit measurable.

We are now prepared to describe the fundamental result of this paper. Let S, X, and Y be Borel spaces. Let A be an analytic subset of $S \times X$, and let B be a Borel subset of $S \times Y$. Denote by $A(s)$ the s-section of A, that is, $A(s) = \{x: (s, x) \in A\}$, $s \in S$. Let $B(s)$ denote the s-section of B, $s \in S$. Assume that $A(s)$ and $B(s)$ are nonempty for every $s \in S$. Define the following set

$$C = \{(s, x, y): s \in S, x \in A(s), \text{ and } y \in B(s)\}. \tag{1}$$

In § 4 we shall prove the following lemma.

LEMMA 1.1. *The set C is an analytic subset of $S \times X \times Y$. If in addition the set A is Borel, then so is C.*

Let $u: C \to R$ be an u.s.a. function on C. We define two functions $r: A \to R^*$ and $w: B \to R^*$ as follows:

$$r(s, x) = \inf_{y \in B(s)} u(s, x, y) \quad \text{and} \quad w(s, y) = \sup_{x \in A(s)} u(s, x, y), \tag{2}$$

where $(s, x) \in A$, and $(s, y) \in B$. Further, we define two functions $v_*, v^*: S \to R^*$ by

$$v_*(s) = \sup_{x \in A(s)} r(s, x) \quad \text{and} \quad v^*(s) = \inf_{y \in B(s)} w(s, y), \qquad s \in S. \tag{3}$$

The functions (2) and (3) arise in the analysis of some minimax stochastic optimization problems such as stochastic games or minimax stochastic control problems (see, e.g., [13] and [4]). From the point of view of the applications they would satisfy some measurability conditions. Our purpose is to give sufficient conditions (in the weakest possible form) that ensure the desired measurability properties of the functions (2) and (3).

We now arrive at our fundamental result:

THEOREM 1.1. *Assume that $B(s)$ is σ-compact for each $s \in S$, and u is the limit of a nondecreasing sequence $\{u_n\}$ of upper semianalytic functions on C such that, for each $(s, x) \in A$ and $n \in N$, $u_n(s, x, \cdot)$ is continuous on $B(s)$ endowed with the relative topology. Then:*

(a) *The function r is upper semianalytic.*

(b) *The function w is $(\mathcal{L}_S \otimes \mathcal{B}_Y)$-measurable.*

(c) *Both the functions v_* and v^* are upper semianalytic.*

In § 4 we prove that Theorem 1.1 implies the following fact.

THEOREM 1.2. *Assume that A is a Borel subset of $S \times X$, $B(s)$ is σ-compact for each $s \in S$, and u is a Borel measurable function on C such that, for each $(s, x) \in A$, $u(s, x, \cdot)$ is lower semicontinuous on $B(s)$ endowed with the relative topology. Then the conclusions of Theorem* 1.1 *hold.*

Remarks. (1) The proof of Theorem 1.1 is postponed to § 4. It utilizes some measure and game theoretic arguments. For example, to resolve some measurability issues a general minimax theorem of Fan [7, Thm. 2] is engaged. In this respect, the methods used here are similar to those of [12] and [13] where some particular cases of Theorem 1.1 were studied.

(2) The assumptions of Theorem 1.1 are inspired by the papers of Brown and Purves [6] and Shreve and Bertsekas [17]. Their measurable selection theorems are

crucial in our development. The conclusions of Theorem 1.1 (Theorem 1.2) may fail in general. For a detailed discussion of this fact see § 3.

(3) Suppose the assumptions of Theorem 1.2 are satisfied. If in addition the set $A(s)$ is σ-compact for each $s \in S$, and, for each $(s, y) \in B$, $u(s, \cdot, y)$ is upper semicontinuous on $A(s)$ endowed with the relative topology, then all functions defined in (2) and (3) are Borel measurable. This follows from Brown and Purves [6, Corollary 1]. (Recall that the set C is Borel in this case.)

In order to give some applications of Theorem 1.1 we need the following important fact.

LEMMA 1.2. *Let S and Y be Borel spaces, and E be a nonempty set from the product σ-algebra $\mathcal{L}_S \otimes \mathcal{B}_Y$. Then the projection* $\operatorname{proj}_S(E)$ *of E from $S \times Y$ on S belongs to $\mathcal{L}_S$. Moreover, there exists a limit measurable function $g: S \to Y$ such that $g(s) \in E(s) = \{y: (s, y) \in E\}$ for every $s \in \operatorname{proj}_S(E)$.*

Proof. This is a corollary to Leese [11, Thm. 5.5].

2. Measurable selections for minimax stochastic optimization problems. In this section we present some possible applications of the fundamental theorem from § 1. In the first place we give an application to game theory. Consider a *two-person zero-sum game model* (S, X, Y, A, B, u) where:

(a) S is a Borel space, called the *state space.*

(b) X and Y are Borel spaces of *actions* for players 1 and 2, respectively.

(c) A and B are the *constraint sets.* It is assumed that A is an analytic subset of $S \times X$, and B is a Borel subset of $S \times Y$. Moreover, all the s-sections $A(s)$ and $B(s)$ of A and B, respectively, are assumed to be nonempty. Let C be the set defined by (1).

(d) $u: C \to R$ is an upper semianalytic *pay-off function* for player 1.

Players 1 and 2 observe the state $s \in S$ and then choose actions $x \in A(s)$ and $y \in B(s)$, respectively. As a consequence of the actions chosen by the players, player 2 pays player 1 $u(s, x, y)$ units of money. Player 1 tries to maximize his income and player 2 tries to minimize his loss.

A *strategy* for player 1 is a universally measurable function $f: S \to X$ such that $f(s) \in A(s)$ for each $s \in S$. Strategies for player 2 are defined analogously.

Let v_* and v^* be the functions defined by (3). The function v_* (v^*) is called the *lower* (resp. *upper*) *value function* of the game. If $v_* = v^*$, this common function is called the *value function* of the game and will be denoted by v.

Define the following sets;

$$O_1 = \{s \in S: v_*(s) = \inf_{y \in B(s)} u(s, x, y) = r(s, x_s) \text{ for some } x_s \in A(s)\}$$

and

$$O_2 = \{s \in S: v^*(s) = \sup_{x \in A(s)} u(s, x, y) = w(s, y_s) \text{ for some } y_s \in B(s)\}.$$

A strategy f^* of player 1 is called *ε-optimal* ($\varepsilon > 0$) for him when

$$\begin{aligned} &r(s, f^*(s)) = v_*(s) && \text{if } s \in O_1, \\ &r(s, f^*(s)) \geqq v_*(s) - \varepsilon && \text{if } s \notin O_1,\ v_*(s) < \infty, \\ &r(s, f^*(s)) \geqq \frac{1}{\varepsilon} && \text{if } s \notin O_1,\ v_*(s) = \infty. \end{aligned}$$

A strategy g^* of player 2 is called *ε-optimal* ($\varepsilon > 0$) for him when

$$w(s, g^*(s)) = v^*(s) \qquad \text{if } s \in O_2,$$

$$w(s, g^*(s)) \leqq v^*(s) + \varepsilon \qquad \text{if } s \notin O_2,\ v^*(s) > -\infty,$$

$$w(s, g^*(s)) \leqq -\frac{1}{\varepsilon} \qquad \text{if } s \notin O_2,\ v^*(s) = -\infty.$$

The game model above is abstracted from the theory of sequential stochastic zero-sum games (see [13] and its references). Besides the universal measurability of the value functions v_* and v^* the central problem arising in the analysis of such games is the existence of ε-optimal universally measurable strategies for both players. The following theorem is an answer to the mentioned questions.

THEOREM 2.1. *Suppose the assumptions of Theorem* 1.1 *are satisfied. Then*:

(a) *The lower value function v_* is upper semianalytic and, for each $\varepsilon > 0$, player* 1 *has an ε-optimal limit measurable strategy.*

(b) *The upper value function v^* is upper semianalytic and, for each $\varepsilon > 0$, player* 2 *has an ε-optimal limit measurable strategy.*

Proof. (a) The function v_* is u.s.a. by Theorem 1.1(c). The existence of an ε-optimal limit measurable strategy for player 1 follows from Theorem 1.1(a) and a selection theorem by Shreve and Bertsekas (see [4, Prop. 7.50 and Appendix B] or [17, pp. 968, 971]).

(b) The function v^* is u.s.a. by Theorem 1.1(c). To prove the last statement of the theorem we define

$$I = \{s \in S\colon v^*(s) = -\infty\},$$

$$E = \{(s, y) \in B\colon v^*(s) = w(s, y)\},$$

$$F = \left\{(s, y) \in B\colon w(s, y) \leqq -\frac{1}{\varepsilon}\right\} \cap I \times Y,$$

$$G = \{(s, y) \in B\colon w(s, y) \leqq v^*(s) + \varepsilon\} - (E \cup F).$$

Note that $I \in \mathscr{L}_S$, and $\operatorname{proj}_S(E) = O_1$. From Theorem 1.1(b), it follows that all the sets E, F, and G are in $\mathscr{L}_S \otimes \mathscr{B}_Y$. Therefore applying Lemma 1.2 to E, F, and G, we complete the proof.

From Theorem 2.1 we can deduce a random version of the well-known minimax theorem of Fan [7]. To do this we must introduce some definitions.

A function $u: C \to R$ is called *concavelike* on $A(s)$ for each $s \in S$ if for any $s \in S$, $x_1, x_2 \in A(s)$ and $\lambda \in [0, 1]$ there exists $x_0 \in A(s)$ such that

$$u(s, x_0, y) \geqq \lambda u(s, x_1, y) + (1 - \lambda) u(s, x_2, y) \quad \text{for all } y \in B(s).$$

A function $u: C \to R$ is called *convexlike* on $B(s)$ for each $s \in S$ if for any $s \in S$, y_1, $y_2 \in B(s)$ and $\lambda \in [0, 1]$, there exists $y_0 \in B(s)$ such that

$$u(s, x, y_0) \leqq \lambda u(s, x, y_1) + (1 - \lambda) u(s, x, y_2) \quad \text{for all } x \in A(s).$$

Here is a random version of Fan's minimax theorem.

THEOREM 2.2. *Suppose that $B(s)$ is compact for each $s \in S$ and the remaining assumptions of Theorem* 1.1 *are satisfied. If in addition u is concavelike on $A(s)$ and convexlike on $B(s)$ for each $s \in S$, then the game has a value function v and v is upper semianalytic. Moreover, for each $\varepsilon > 0$, both players have ε-optimal limit measurable strategies.*

Remarks. (1) Under the compactness assumption of Theorem 2.2 the set O_2 is equal to S, so player 2 has an optimal limit measurable strategy. The conclusions of Theorems 2.1 and 2.2 remain valid if the assumptions of Theorem 1.1 are replaced by the ones of Theorem 1.2.

(2) Theorem 2.1 generalizes [12, Props. 3.1, 3.2] while Theorem 2.2 extends [12, Thms. 4.1, 4.2].

(3) The mathematical facts stated in Theorem 2.1 may have another interpretation. Namely, they may be applied to some *minimax* (or *maximin*) *stochastic control systems* related to those of [4, Chaps. 2–5]. Suppose u is a *one-stage cost function* of a minimax stochastic control system where S is the *state space*, Y is a *control space*, and X is a *disturbance space*. Then v^* is the *optimal one-stage cost function*. From Theorem 2.1, it follows that v^* is u.s.a., and the controller has an *ε-optimal one-stage policy*. (The terminology used here is taken from [4].)

As a next application of the fundamental theorem from § 1 we provide a random version of Fan's inequality [8]. This inequality has many applications to game and control theory (see, e.g., [2]). In what follows we fix that $X = Y$ and $A(s) = B(s)$ for each $s \in S$. The set C is defined by (1). We assume now that X is a subset of a linear topological Hausdorff space and is endowed with the relative topology. Let u be a real-valued function on C. We say that $u(s, \cdot, y)$ is *quasiconcave* for each $(s, y) \in B$ if the set $\{x \in A(s)\colon u(s, x, y) > c\}$ is convex for each $(s, y) \in B$ and $c \in R$.

Here is a random version of Fan's inequality.

THEOREM 2.3. *Besides the assumptions above suppose that X is a Borel space and A is a Borel subset of $S \times X$ such that $A(s)$ is a compact convex set for each $s \in S$. Assume further that $u : C \to R$ is a Borel measurable function such that $u(s, \cdot, y)$ is quasiconcave for each $(s, y) \in B$ and $u(s, x, \cdot)$ is lower semicontinuous on $B(s)$ for each $(s, x) \in A$. If in addition*

$$\sup_{y \in B(s)} u(s, y, y) \leqq 0 \quad \textit{for each } s \in S,$$

then there exists a limit measurable function $g : S \to Y$ such that

$$g(s) \in B(s) \quad \textit{and} \quad \sup_{x \in A(s)} u(s, x, g(s)) \leqq 0 \quad \textit{for each } s \in S.$$

Proof. Define $E = \{(s, y) \in B\colon \sup_{x \in A(s)} u(s, x, y) \leqq 0\}$. By Fan's inequality [2, Thm. 7.1.3], the s-section $E(s)$ of E is nonempty for each $s \in S$. By Theorem 1.2, the set E belongs to $\mathscr{L}_S \otimes \mathscr{B}_Y$. Thus the result follows now from Lemma 1.2.

3. Descriptive set theoretic aspects of the fundamental theorem. In this section we shall discuss Theorem 1.2 from the point of view of descriptive set theory. In the first place we shall demonstrate that Theorem 1.2 implies a new result on projections of Borel sets.

THEOREM 3.1. *Let A be a Borel subset of $S \times X$ and B be a Borel subset of $S \times Y$ such that each s-section $B(s)$ of B is σ-compact. Let C be a subset of $S \times X \times Y$ defined by* (1). *Assume that D is a Borel subset of C such that $D(s, x) = \{y\colon (s, x, y) \in D\}$ is open in $B(s)$ endowed with the relative topology. Then the projection* $\mathrm{proj}_{S \times Y}(D)$ *of D from $S \times X \times Y$ on $S \times Y$ belongs to $\mathscr{L}_S \otimes \mathscr{B}_Y$.*

Proof. Let $u = \chi_D$ be the indicator function of D. Then u satisfies the assumptions of Theorem 1.2. By Theorem 1.2 the following set

$$E = \{(s, y) \in B\colon \sup_{x \in A(s)} u(s, x, y) \leqq 0\} = \{(s, y) \in B\colon w(s, y) \leqq 0\}$$

belongs to $\mathscr{L}_S \otimes \mathscr{B}_Y$. Note that

$$
\begin{aligned}
E &= B - \{(s, y) \in B: \sup_{x \in A(s)} u(s, x, y) > 0\} \\
&= B - \operatorname{proj}_{S \times Y} (\{(s, x, y) \in C: u(s, x, y) > 0\}) \qquad (4) \\
&= B - \operatorname{proj}_{S \times Y} (D).
\end{aligned}
$$

Since both B and E belong to $\mathscr{L}_S \otimes \mathscr{B}_Y$, so $\operatorname{proj}_{S \times Y} (D)$ has the same property.

Remark. The conclusions (a) and (c) of Theorem 1.1 (Theorem 1.2) need not hold in general. This fact was noticed for the first time by Rieder in [14, Example 4.1]. In his example (based on the axiom of constructibility) Rieder has shown that both v_* and v^* need not be universally measurable. The same example shows that the conclusion (a) of Theorem 1.1 may fail in general. For a comment on a related topic see also [4, p. 302]. (The example of Rieder is also reported in [12, Example 3.1].)

We close this section with an example showing that also the conclusion (b) of Theorem 1.1 may fail in general.

Counterexample. Assume that $S = X = Y = R$, and $A(s) = B(s)$ for each $s \in S$. According to Aumann [3], it is consistent with the usual axioms of set theory to assume that there exists a function $g: R \to R$ whose graph is a coanalytic subset of R^2, but g is not Lebesgue measurable. (To obtain this fact the axiom of constructibility is assumed.)

Let $F = S \times Y - \text{graph}\,(g)$. Then F is an analytic set in $S \times Y$, and, by [4, Prop. 7.39], there exists a Borel set $D \subset S \times X \times Y$ such that $\operatorname{proj}_{S \times Y} (D) = F$. Let $u = \chi_D$ be the indicator function of D. Then u is a Borel measurable function on $S \times X \times Y$. From (4) we get

$$
E = \{(s, y): \sup_{x \in X} u(s, x, y) \leqq 0\} = S \times Y - F = \text{graph}\,(g).
$$

Suppose that $E \in \mathscr{L}_S \otimes \mathscr{B}_Y$. Then by Lemma 1.2 the function g is limit measurable, which leads to a contradiction. Therefore $E \notin \mathscr{L}_S \otimes \mathscr{B}_Y$.

4. Proof of the fundamental theorem. We start with a proof of the properties of the set C defined in § 1.

Proof of Lemma 1.1. We define two sets

$$
\begin{aligned}
A^* &= \{(s, x, y): s \in S, x \in A(s), y \in Y\}, \\
B^* &= \{(s, x, y): s \in S, x \in X, y \in B(s)\}.
\end{aligned}
$$

Note that $A^* = \operatorname{proj}_{S \times X}^{-1} (A)$ and $B^* = \operatorname{proj}_{S \times Y}^{-1} (B)$ and $C = A^* \cap B^*$. Since the projection mapping is continuous, so the set B^* is Borel, and by [4, Prop. 7.40] the set A^* is analytic. Thus C is analytic. Of course if in addition A is Borel, then so is C.

We now mention some auxiliary measure theoretic facts that we shall be using. Let T be a Borel space. We denote by P_T the set of all probability measures on $\mathscr{B}_T$ and assume that P_T is endowed with the weak topology. By [4, Corollary 7.25.1], P_T is a Borel space too. If in addition T is compact, then so is P_T [4, Prop. 7.22].

For any metric d on T, let $U_d(T)$ be the space of bounded real-valued functions on T which are uniformly continuous with respect to d.

The following lemma follows from [4, Prop. 7.19].

LEMMA 4.1. *There exists a metric d on T consistent with its topology and a countable dense subset* $\{f_n\}$ *of the unit ball of* $U_d(T)$ *such that the function* $m: P_T \times P_T \to R$ *defined*

by

$$m(p, q) = \sum_{n=1}^{\infty} 2^{-n} \left| \int f_n(x)p(dx) - \int f_n(x)q(dx) \right|$$

is a metric on P_T equivalent to the weak topology of P_T.

From (32) and (35) of [5] we can deduce the following fact.

LEMMA 4.2. *Let $f: S \times X \times Y \to R$ be a bounded u.s.a. function. Then the function $\bar{f}: S \times P_X \times P_Y \to R$ defined by*

$$\bar{f}(s, p, q) = \int\int f(s, x, y)p(dx)q(dy), \qquad p \in P_X, q \in P_Y, s \in S, \tag{5}$$

is u.s.a. on the Borel space $S \times P_X \times P_Y$.

COROLLARY 4.1. *Define the following sets:*

$$\bar{A} = \{(s, p): s \in S, \text{ and } p \in P_{A(s)}\}, \qquad \bar{B} = \{(s, q): s \in S, \text{ and } q \in P_{B(s)}\},$$

and

$$\bar{C} = \{(s, p, q): s \in S, p \in P_{A(s)}, \text{ and } q \in P_{B(s)}\}.$$

Assume that A is analytic and B is Borel. Then the sets $\bar{A}$ and $\bar{C}$ are analytic while the set $\bar{B}$ is Borel.

Proof. Let $f = \chi_C$ be the indicator function of C. By Lemma 1.1, f is u.s.a. on $S \times X \times Y$. Note that $\bar{C} = \{(s, p, q): \bar{f}(s, p, q) \geqq 1\}$. By Lemma 4.2, $\bar{f}$ is u.s.a., which implies that $\bar{C}$ is analytic. The proof of the fact that $\bar{A}$ is analytic is similar. Now we prove that $\bar{B}$ is Borel. Clearly, $\bar{B}$ is analytic. Let $f_1 = -\chi_B$. The function f_1 is Borel measurable and hence u.s.a. By Lemma 4.2, $\bar{f}_1$ is u.s.a., which implies that the set $D = \{(s, q): \bar{f}_1(s, q) \leqq -1\}$ is coanalytic. But $D = \bar{B}$, so $\bar{B}$ is also coanalytic. Thus from Suslin's theorem [10] it follows that the set $\bar{B}$ is Borel.

COROLLARY 4.2. *Let $f: C \to R$ be a bounded u.s.a. function. Then the function $\bar{f}$ defined on $\bar{C}$ by* (5) *is u.s.a.*

Proof. This follows from Lemma 4.2, Corollary 4.1, and the fact that the intersection of two analytic sets is analytic.

LEMMA 4.3. *Let S and Y be Borel spaces, and $h: S \times Y \to R$ be a function.*

(a) *If $h(\cdot, y)$ is limit measurable for each $y \in Y$, and $h(s, \cdot)$ is continuous for each $s \in S$, then h is $(\mathcal{L}_S \otimes \mathcal{B}_Y)$-measurable.*

(b) *If $h(\cdot, y)$ is u.s.a. for each $y \in Y$, and $h(s, \cdot)$ is continuous for each $s \in S$, then h is u.s.a.*

Proof. Part (a) follows directly from [9, Thm. 6.1]. A proof of (b) can be given by a straightforward translation of that of part (a), using the fact that the countable union and intersection of analytic sets is analytic.

The following lemma is a special case of Theorem 1.1.

LEMMA 4.4. *Assume that $B(s)$ is compact for each $s \in S$, and u is a bounded function on C. Let the remaining assumptions of Theorem* 1.1 *be satisfied. Then:*

(a) *The function r is u.s.a.*

(b) *The function w is $(\mathcal{L}_S \otimes \mathcal{B}_Y)$-measurable.*

(c) *Both the functions v_* and v^* are u.s.a.*

Proof. We recall that u is assumed to be the limit of a nondecreasing sequence $\{u_n\}$ of u.s.a. functions on C such that, for each $n \in N$ and $(s, x) \in A$, $u_n(s, x, \cdot)$ is continuous on $B(s)$ endowed with the relative topology. By [4, Lemma 7.30], u is u.s.a. Moreover, the assumption above implies that, for each $(s, x) \in A$, $u(s, x, \cdot)$ is lower semicontinuous on $B(s)$.

(a) By the proof of [12, Prop. 3.2], the function defined by

$$f_n(s, x) = \inf_{y \in B(s)} u_n(s, x, y), \qquad (s, x) \in A,$$

is u.s.a. Using [16, Prop. 10.1], we can show that

$$\lim_n f_n(s, x) = \inf_{y \in B(s)} \lim_n u_n(s, x, y) = r(s, x), \qquad (s, x) \in A.$$

(Recall that $B(s)$ is assumed to be compact.) By [4, Lemma 7.30], the function r is u.s.a.

(b) To prove this statement we introduce some measure theoretic tools. We embed $X(Y)$ in the space P_X (P_Y) of probability measures on $\mathcal{B}_X$ ($\mathcal{B}_Y$) endowed with the weak topology. The embedding of Y in P_Y means that every $y \in Y$ is recognized as a probability measure δ_y concentrated at point y and it is a homeomorphism [4, Corollary 7.21.1].

Define

$$\bar{u}(s, p, q) = \int\int u(s, x, y) p(dx) q(dy), \qquad (s, p, q) \in \bar{C},$$

(6)

$$\bar{w}(s, q) = \sup_{p \in P_{A(s)}} \bar{u}(s, p, q), \qquad (s, q) \in \bar{B}.$$

We put $\bar{u}(s, p, y) = \bar{u}(s, p, \delta_y)$ and $\bar{w}(s, y) = \bar{w}(s, \delta_y)$, where $(s, p, \delta_y) \in \bar{C}$. Let $\bar{u}_k$, $k \in N$, be the functions defined by (6) where u is replaced by u_k. Clearly, every function $\bar{u}_k(s, p, \cdot)$ is continuous on $P_{B(s)}$. By Corollary 4.2, both $\bar{u}_k$ and $\bar{u}$ are u.s.a. for each $k \in N$. By the monotone convergence theorem $\bar{u}_k \nearrow \bar{u}$. This implies that $\bar{u}(s, p, \cdot)$ is lower semicontinuous (l.s.c.) on $P_{B(s)}$ for each $(s, p) \in \bar{A}$, and consequently so is $\bar{w}(s, \cdot)$ for each $s \in S$.

Define

$$f_n(s, q) = \inf_{t \in P_{B(s)}} [\bar{w}(s, t) + nm(t, q)], \qquad (s, q) \in S \times P_Y,$$

where m is a metric on P_Y defined in Lemma 4.1. Since $\bar{w}(s, \cdot)$ is l.s.c., so, by the proof of the theorem of Baire (see [1, p. 390] or [4, p. 147]), we get $f_n \nearrow \bar{w}$ on $\bar{B}$. Consequently, $f_n \nearrow w$ on B, because the function $\bar{w}$ restricted to B is equal to w.

We shall prove that $f_n(\cdot, y)$ is limit measurable on S and $f_n(s, \cdot)$ is continuous on Y for every $s \in S$, $y \in Y$, and $n \in N$. Then by Lemma 4.3(a), f_n is $(\mathcal{L}_S \otimes \mathcal{B}_Y)$-measurable for each $n \in N$, which implies that so is w. The continuity of $f_n(s, \cdot)$ for each $s \in S$ and $n \in N$ follows from the proof of Lemma 7.7 in [4]. Let y be an arbitrary element of Y. In order to show the measurability of $f_n(\cdot, y)$, we define

$$g_n(s, p, t, y) = \bar{u}(s, p, t) + nm(t, \delta_y), \tag{7}$$

where $(s, p, t) \in \bar{C}$, $y \in Y$, and $n \in N$,

$$g'_n(s, p, y) = \inf_{t \in P_{B(s)}} g_n(s, p, t, y), \tag{8}$$

where $(s, p) \in \bar{A}$, $y \in Y$, and $n \in N$.

Clearly,

$$f_n(s, y) = \inf_{t \in P_{B(s)}} \sup_{p \in P_{A(s)}} g_n(s, p, t, y), \qquad (s, y) \in S \times Y. \tag{9}$$

Thus, applying the minimax theorem of Fan [7, Thm. 2], (8) and (9), we get

$$f_n(s, y) = \sup_{p \in P_{A(s)}} \inf_{t \in P_{B(s)}} g_n(s, p, t, y) = \sup_{p \in P_{A(s)}} g'_n(s, p, y), \qquad (s, y) \in S \times Y. \tag{10}$$

We now show that $g_n(\cdot, y)$ satisfies the assumptions of our lemma. By Corollary 4.1, the sets $\bar{A}$, $\bar{C}$ are analytic, and $\bar{B}$ is Borel. Note that $\bar{B}(s) = P_{B(s)}$ for each $s \in S$, so each $\bar{B}(s)$ is compact. By Corollary 4.2, $g_n(\cdot, y)$ is u.s.a. Moreover, there exists a nondecreasing sequence $\{g_{nk}\}$ of u.s.a. functions on $\bar{C}$ such that $g_{nk}(s, p, \cdot, y)$ is continuous on $\bar{B}(s)$ for each $(s, p) \in \bar{A}$, $k \in N$, and $g_{nk}(\cdot, y) \nearrow g_n(\cdot, y)$ as $k \to \infty$. As a matter of fact, g_{nk} is the function (7) where $\bar{u}$ is replaced by $\bar{u}_k$. From the proof of part (a) of our lemma we infer that the function $g'_n(\cdot, y)$ defined by (8) is u.s.a. on $\bar{A}$. This, (10), and [4, Prop. 7.47] or [17, p. 968] imply that $f_n(\cdot, y)$ is u.s.a. on S. Thus, we have shown that $f_n(\cdot, y)$ is limit measurable for each $y \in Y$, and $f_n(s, \cdot)$ is continuous for each $s \in S$, which terminates the proof of (b).

(c) The function v_* is u.s.a. by (a) and [17, p. 968] or [4, Proposition 7.47]. It remains to show that v^* is u.s.a. Recall that

$$v^*(s) = \inf_{y \in B(s)} w(s, y), \qquad s \in S.$$

But w is the limit of the nondecreasing sequence $\{f_n\}$ from the proof of part (b). Recall that every $f_n(s, \cdot)$ is continuous and $f_n(\cdot, y)$ is u.s.a. By Lemma 4.3(b), every f_n is u.s.a. Thus the fact that v^* is u.s.a. follows from the proof of part (a) of the lemma.

To prove Theorem 1.1 we shall also use the following fact which follows from the main result of Saint-Raymond from [15].

LEMMA 4.5. *Let S and Y be Borel spaces and let B be a Borel subset of $S \times Y$ such that each s-section $B(s)$ of B is nonempty and σ-compact in Y. Then there exists a sequence $\{B_n\}(n \in N)$ of Borel subsets of $S \times Y$, each of which with nonempty compact s-sections $B_n(s)$, such that*

$$B_n \subset B_{n+1} \quad \text{for each } n \in N \quad \text{and} \quad \bigcup_{n \in N} B_n = B.$$

Remark. Saint-Raymond has shown the above fact under a stronger assumption that both S and Y are compact metric spaces. However, by Urysohn's theorem, the Borel spaces S and Y may be homeomorphically embedded in compact metric spaces, say S^* and Y^*, so that B may be recognized as a subset of $S^* \times Y^*$. Note that such an embedding of $S \times Y$ in $S^* \times Y^*$ is a Borel homeomorphism preserving the σ-compactness of the s-sections of B. Thus the result of Saint-Raymond is valid in the more general case above.

Let $\{B_n\}$ be a sequence of Borel sets from Lemma 4.5. For any $u: C \to R$ and $n \in N$, we put $h_n = \chi_{B_n} u$ and

$$r'_n(s, x) = \inf_{y \in B_n(s)} h_n(s, x, y), (s, x) \in A, \qquad w'_n(s, y) = \sup_{x \in A(s)} h_n(s, x, y),$$

$(s, y) \in B$, and

$$v'_n(s) = \inf_{y \in B_n(s)} w_n(s, y), \qquad s \in S.$$

LEMMA 4.6. *Let u be a bounded nonnegative function and let the remaining assumptions of Theorem 1.1 be satisfied. Then for each $n \in N$ the function h_n is u.s.a. Moreover:*

(a) *The function r'_n is u.s.a.*

(b) *The function w'_n is $(\mathcal{L}_S \otimes \mathcal{B}_Y)$-measurable.*

(c) *The function v'_n is u.s.a.*

Proof. This is a corollary to Lemma 4.4.

LEMMA 4.7. *Assume that u is nonnegative. Then $r'_n \nearrow r$ on A and $w'_n \nearrow w$ on B as $n \to \infty$.*

Proof. The proof is straightforward.

Finally we give the following simple lemma.

LEMMA 4.8. *Let $f_n, f: T \to R (n \in N)$ be functions.*

(a) *If $f_n \nearrow f$ on T, then $\sup_{t\in T} f_n(t) \nearrow \sup_{t\in T} f(t)$.*

(b) *If $f_n \searrow f$ on T, then $\inf_{t\in T} f_n(t) \searrow \inf_{t\in T} f(t)$.*

(c) *If $f_n = \max\{f, -n\}$, then $\sup_{t\in T} f_n(t) \searrow \sup_{t\in T} f(t)$.*

(d) *If $f_n = \min\{f, n\}$, then $\inf_{t\in T} f_n(t) \nearrow \inf_{t\in T} f(t)$.*

Proof of Theorem 1.1. First of all we note that the conclusions (a) and (b) of Theorem 1.1 hold for any bounded nonnegative function u. This follows from Lemmas 4.6, 4.7 and from the fact that the limit of any sequence of u.s.a. (product measurable) functions on a Borel space (product of Borel spaces) is u.s.a. (product measurable) (see [4, Lemma 7.30]). It remains to prove (a) and (b) for unbounded functions.

Let u be a function satisfying the assumptions of Theorem 1.1. For each $n \in N$, we define

$$u_n = \max\{u, -n\} \quad \text{and} \quad u^n = \min\{u, n\}.$$

It is easy to check that u_n and u^n satisfy the assumptions of Theorem 1.1 for each $n \in N$. Define r_n (r^n) and w_n (w^n) by (2) where u is replaced by u_n (u^n).

(a) Suppose u is unbounded and nonnegative. By Lemma 4.8(d), we have $r^n \nearrow r$ on A. Since r^n is u.s.a. on A for each $n \in N$, so is r by [4, Lemma 7.30]. The result we have already obtained can be (in an obvious way) extended to the case of all functions u which are bounded below. Suppose now u is unbounded. By Lemma 4.8(b), we have $r_n \searrow r$ on A. We have already shown that each r_n is u.s.a. (u_n is bounded below). Thus r is u.s.a. by [4, Lemma 7.30].

(b) The proof of (b) for an unbounded function u proceeds similar lines as that of (a). (We use parts (a) and (c) of Lemma 4.8 instead of (b) and (d).)

(c) The fact that v_* is u.s.a. follows from part (a) of Theorem 1.1 and [4, Prop. 7.47] or [17, p. 968]. It remains to show that v^* is u.s.a. It is easy to check that $v'_n \to v^*$ as $n \to \infty$. For each $n \in N$ and a bounded nonnegative u the function v'_n is u.s.a. by Lemma 4.6(c). Thus v^* is u.s.a. for each bounded nonnegative function u. Suppose now that u is unbounded and nonnegative. It is easy to see that $w^n = \min\{w, n\}$ for each $n \in N$. By Lemma 4.8(d), we have

$$\inf_{y\in B(s)} w^n(s, y) \nearrow \inf_{y\in B(s)} w(s, y) = v^*(s), \qquad s \in S.$$

This and [4, Lemma 7.30] imply that v^* is u.s.a. for every nonnegative function u. Clearly, this result can be extended to the case of any function u which is bounded below. Now using this fact and Lemma 4.8(c), (b), we can prove that v^* is u.s.a. for any (unbounded) function u.

Proof of Theorem 1.2. We have to prove that u is the limit of a nondecreasing sequence $\{u_n\}$ of Borel measurable functions on the Borel set C such that, for each $(s, x) \in A$ and $n \in N$, $u_n(s, x, \cdot)$ is continuous on $B(s)$. This can be done in a similar way as the proof of [16, Prop. 11.6]. Note that without loss of generality we can assume that u is bounded. (If u is unbounded then we can apply the arguments given below to the function $\operatorname{arctg}(u)$ instead of u, which also satisfies the assumptions of Theorem 1.2.) For each $n \in N$, let u_n be defined by

$$u_n(s, x, y) = \inf_{t\in B(s)} [u(s, x, t) + nd(t, y)], \qquad (s, x) \in A, y \in Y,$$

where d is a metric on Y consistent with its topology. By Corollary 1 of Brown and Purves [6], the function u_n is Borel measurable for each $n \in N$. It is easy to check that

$u_n(s, x, \cdot)$ is continuous on Y for each $(s, x) \in A$ and $n \in N$. By the proof of the theorem of Baire (see [1, p. 390] or [4, p. 147]), we have $u_n \nearrow u$ as $n \to \infty$, which terminates the proof.

REFERENCES

[1] R. B. ASH, *Real Analysis and Probability*, Academic Press, New York, 1972.

[2] J. P. AUBIN, *Mathematical Methods of Game and Economic Theory*, North-Holland, Amsterdam, 1979.

[3] R. J. AUMANN, *Measurable utility and the measurable choice theorem*, La Decision Actes Colloq. International du Centre National de la Recherche Scientifique, Aix-en-Provence, 1967, Paris, 1969, pp. 15-26.

[4] D. P. BERTSEKAS AND S. E. SHREVE, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, New York, 1978.

[5] D. BLACKWELL, D. FREEDMAN AND M. ORKIN, *The optimal reward operator in dynamic programming*, Ann. Probab., 2 (1974), pp. 926-941.

[6] L. D. BROWN AND R. PURVES, *Measurable selections of extrema*, Ann. Statist., 1 (1973), pp. 902-912.

[7] K. FAN, *Minimax theorems*, Proc. Nat. Acad. Sci. USA, 39 (1953), pp. 42-47.

[8] ———, *A minimax inequality and applications*, in Inequalities, Vol. III, Academic Press, New York, 1972.

[9] C. J. HIMMELBERG, *Measurable relations*, Fund. Math., 87 (1975), pp. 53-72.

[10] K. KURATOWSKI, *Topology* I, Academic Press, New York, 1966.

[11] S. J. LEESE, *Measurable selections and the uniformization of Souslin sets*, Amer. J. Math., 100 (1978), pp. 19-41.

[12] A. S. NOWAK, *Minimax selection theorems*, J. Math. Anal. Appl., 103 (1984), pp. 106-116.

[13] ———, *Universally measurable strategies in zero-sum stochastic games*, Ann. Probab., 13 (1985), No. 1, in press.

[14] U. RIEDER, *On semi-continuous dynamic games*, Preprint, University of Karlsruhe, 1978.

[15] J. SAINT-RAYMOND, *Boréliens à coupes K_σ*, Bull. Soc. Math. France, 104 (1976), pp. 389-400.

[16] M. SCHÄL, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Z. Wahrsch. Verw. Gebiete, 32 (1975), pp. 179-196.

[17] S. E. SHREVE AND D. P. BERTSEKAS, *Alternative theoretical frameworks for finite horizon discrete-time stochastic optimal control*, this Journal, 16 (1978), pp. 953-978.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 3, May 1985

009

ALGORITHM MODELS FOR NONDIFFERENTIABLE OPTIMIZATION*

E. POLAK† AND D. Q. MAYNE‡

Abstract. It is shown that a number of seemingly unrelated nondifferentiable optimization algorithms are special cases of two simple algorithm models: one for constrained and one for unconstrained optimization. In both of these models, the direction finding procedures use parametrized families of maps which are locally uniformly u.s.c. with respect to the generalized gradients of the functions defining the problem. The selection of the parameter is determined by a rule which is analogous to the one used in methods of feasible directions.

Key words. nondifferentiable optimization, algorithm theory, algorithm models, semi-infinite optimization

1. Introduction. A formal extension of a differentiable optimization algorithm to the nondifferentiable case consists of replacing gradient vectors $\nabla f(x)$, used by the algorithm in solving a differentiable problem, by the vectors $h(x) = \operatorname{argmin}\{\|h\| \mid h \in \partial f(x)\}$, when applied to a nondifferentiable problem, with $\partial f(x)$ denoting the Clarke generalized gradient of $f(x)$; see [C1]. Such formal extensions cannot be shown to converge to stationary points. The reason for this is that while gradients are usually locally uniformly continuous, generalized gradients usually are not even locally uniformly upper-semi-continuous (u.s.c.).

An examination of the nondifferentiable optimization literature, see e.g. [B2], [C3], [G1], [G2], [L2]-[L4], [M1], [M3], [P1]-[P8], shows that in order to overcome this lack of local uniform upper-semi-continuity, the search direction procedures of nondifferentiable optimization algorithms invariably replace gradients not by generalized gradients, but by better behaved supersets which are obtained in a variety of ways. These supersets reflect the local behavior of the functions in question. When only local Lipschitz continuity is assumed, the supersets consist of bundles of generalized gradients which are generated by exploring a neighborhood about the current iterate; see e.g. [B2], [G1], [P3]. When the problem functions are convex, subgradient bundles are used as supersets, see e.g. [L2]-[L4]. When the problem functions are semi-smooth, a special line exploration method can be used to eliminate the need for acquiring a bundle of generalized gradients; see e.g. [M1], [M3], [P3], [P4]. When the problem functions are in some sense piece-wise differentiable and allow one to determine whether one is at a differentiable point or not, the need for constructing generalized gradient bundles disappears altogether since much simpler supersets can generally be used, as we see from [C3], [G2], [M3], [P1], [P5], [P6].

In [P7], we find a theory dealing with the extension of differentiable optimization algorithms to the nondifferentiable case. This theory requires the use of bundles of generalized gradients, computed in an ε ball about the current iterate, with the value of $\varepsilon > 0$ controlled by a mechanism analogous to the one used in the Polak method of feasible directions [P9] and in phase I-phase II methods such as those in [P2]. The theory in [P7] does not contribute to the understanding or the construction of algorithms, such as those in [C3], [G2], [M3], [P1], [P5], [P6], that do not use generalized gradient

* Received by the editors April 8, 1983, and in revised form February 9, 1984. This research was sponsored by the National Science Foundation under grants ECS-79-13148 and CEE-8105790, the Joint Services Electronics Program under contract F49620-79-C-0178, and the UK Science Research Council.

† College of Engineering, Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94720.

‡ Department of Electrical Engineering, Imperial College, London SW7 2BT, England.

bundles, and it leads to implementable algorithms only when all the problem functions are semi-smooth.

It has generally been thought that the cumbersome algorithms, which fit within the framework established in [P7], have nothing to do with the highly specialized algorithms in [C3], [G2], [P5], [P6], which exploit the properties of such functions as $f(x) = \max \{\phi(x, t) \mid t \in T\}$, with T a closed interval, or $f(x) = \max$ eigenvalue $(Q(x))$ with $Q(x)$ a differentiable, complex valued Hermitian matrix. It is shown in this paper that this impression is wrong by showing that both classes of nondifferentiable optimization algorithms can be seen as special cases of two simple algorithm models: one for constrained and one for unconstrained optimization. These algorithm models make use of generalized gradient supersets which are "almost" lower semicontinuous (a global version of this concept was first used in [P8]). In particular, it is shown in this paper that both the generalized gradient bundles used in [P7] and the supersets used in the algorithms in [C3], [G2], [M3], [P1], [P5], [P6] have this "almost" l.s.c. property. The algorithms in [C3], [G2], [M3], [P1], [P5], [P6] solve problems involving functions of the form $f(x) = \phi(g(x))$, where $g: R^n \to R^m$ is continuously differentiable and $\phi: R^m \to R^1$ is locally Lipschitz. It is shown that the supersets used by these algorithms are the generalized gradients of perturbation functions. Some rules for the construction of appropriate perturbation functions are given.

It is to be hoped that as a result of the work reported in this paper, both the exposition of nondifferentiable optimization algorithms and the invention of new ones will be considerably simplified.

2. Unconstrained optimization. In this section we shall consider algorithm models for solving problems of the form:

$$\min_{x \in \mathbb{R}^n} f(x) \tag{2.1}$$

where $f: \mathbb{R}^n \to \mathbb{R}^1$ is locally Lipschitz continuous. Extensions of our results to normed spaces are quite straightforward and hence will be left to the interested reader.

We recall that a locally Lipschitz function $f(\cdot)$ is differentiable almost everywhere, and that one can define for it a generalized gradient $\partial f(x)$ [C1], by

$$\partial f(x) = \operatorname{co} \{\lim_{i \to \infty} \nabla f(x + v_i)\} \tag{2.2a}$$

where the $v_i \to 0$ as $i \to \infty$ are such that $\nabla f(x + v_i)$ exists and co denotes the convex hull of the set in question. It is shown in [C1] that the map $\partial f(\cdot)$ is u.s.c. in the sense of Berge [B1] and bounded on bounded sets.

When $f(\cdot)$ is only locally Lipschitz, the ordinary directional derivative

$$df(x, h) \triangleq \lim_{\lambda \searrow 0} \frac{f(x + \lambda h) - f(x)}{\lambda} \tag{2.2b}$$

may not exist. Instead, see [C1], one defines the *generalized* directional derivative of f at x, in the direction h by

$$d_0 f(x, h) \triangleq \overline{\lim_{\substack{\lambda \searrow 0 \\ y \to 0}}} \frac{f(x + y + \lambda h) - f(x + y)}{\lambda}. \tag{2.2c}$$

It was shown in [C1] that

$$d_0 f(x, h) = \max_{\xi \in \partial f(x)} \langle \xi, h \rangle. \tag{2.2d}$$

As we recall, given an $x_i \in \mathbb{R}^n$ the Armijo gradient method [A1], [P9], for differentiable optimization in $\mathbb{R}^n$, first computes the *steepest descent* direction

$$h(x_i) \triangleq \arg\min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}\|h\|^2 + df(x_i, h)\} = -\nabla f(x_i); \tag{2.3a}$$

next, with $\alpha, \beta \in (0, 1)$, it computes the step size

$$\lambda_i \triangleq \max\{\beta^k \mid k \in \mathbb{N}_+, f(x_i + \beta^k h(x_i)) - f(x_i) \leqq -\beta^k \alpha \|h(x_i)\|^2\}, \tag{2.3b}$$

where $\mathbb{N}_+ = \{0, 1, 2, \cdots\}$; then it updates according to

$$x_{i+1} = x_i + \lambda_i h(x_i). \tag{2.3c}$$

The simplest idea for extending this method (as well as others) to the nondifferentiable case, consists of replacing (2.3a) by

$$\begin{aligned} h(x_i) &\triangleq \arg\min_{h \in \mathbb{R}^n} \{\tfrac{1}{2}\|h\|^2 + d_0 f(x_i, h)\} \\ &= -\arg\min\{\tfrac{1}{2}\|h\|^2 \mid h \in \partial f(x_i)\}, \end{aligned} \tag{2.3d}$$

while leaving (2.3b), (2.3c) unaltered.

Unfortunately, because $\partial f(\cdot)$ is *not* continuous, such extensions fail to be convergent. Consequently, many unconstrained optimization algorithms compute a search direction $h(x_i)$ at x_i by solving an auxiliary problem of the form (2.3d), but with $d_0 f(x_i, h)$ replaced by a kind of ε-generalized derivative $d_\varepsilon f(x_i, h)$, with $\varepsilon > 0$, defined by

$$d_\varepsilon f(x, h) \triangleq \max_{\xi \in G_\varepsilon f(x)} \langle \xi, h \rangle, \tag{2.3e}$$

where for every $\varepsilon \geqq 0$, and $x \in \mathbb{R}^n$, $\partial f(x) \subset G_\varepsilon f(x)$, and the sets $G_\varepsilon f(x)$ are compact, convex and "almost" l.s.c., thus making up for the lack of continuity in $\partial f(\cdot)$. We note that because $\partial f(x) \subset G_\varepsilon f(x)$ for all $\varepsilon \geqq 0$, we always have $d_0 f(x, h) \leqq d_\varepsilon f(x, h)$. When this substitution is made (2.3d) becomes

$$\begin{aligned} h(x_i) &\triangleq \arg\min\{\tfrac{1}{2}\|h\|^2 + d_\varepsilon f(x_i, h)\} \\ &= -\arg\min\{\tfrac{1}{2}\|h\|^2 \mid h \in G_\varepsilon f(x_i)\}. \end{aligned} \tag{2.3f}$$

In addition, a mechanism must be introduced for driving ε to zero. The Polak method of feasible directions [P9] provides an idea for this purpose.

The commonly utilized properties of the maps $G_\varepsilon f(x)$ can be summarized as follows.

DEFINITIONS 2.1. We shall say that $\{G_\varepsilon f(\cdot)\}_{\varepsilon \geqq 0}$, $G_\varepsilon f: \mathbb{R}^n \to 2^{\mathbb{R}^n}$, is a family of *convergent direction finding* (c.d.f.) maps for the locally Lipschitz function $f: \mathbb{R}^n \to \mathbb{R}^1$ if

(i) for all $x \in \mathbb{R}^n$, $\partial f(x) = G_0 f(x)$;

(ii) for all $x \in \mathbb{R}^n$, $\varepsilon < \varepsilon' \Rightarrow G_\varepsilon f(x) \subset G_{\varepsilon'} f(x)$;

(iii) for any $\varepsilon \geqq 0$, $G_\varepsilon f(x)$ is convex and bounded on bounded sets;

(iv) $G_\varepsilon f(x)$ is u.s.c. in (ε, x), in the sense of Berge [B1] at $(0, \hat{x})$, for all $\hat{x} \in \mathbb{R}^n$;

(v) given any $\hat{x} \in \mathbb{R}^n$, $\hat{\varepsilon} > 0$ and $\hat{\delta} > 0$, there exists a $\hat{\rho} > 0$ such that for any $x \in B(\hat{x}, \hat{\rho}) \triangleq \{x \mid \|x - \hat{x}\| \leqq \hat{\rho}\}$ and any $\hat{\eta} \in \partial f(\hat{x})$, there exists an $\eta \in G_{\hat{\varepsilon}} f(x)$ such that $\|\eta - \hat{\eta}\| \leqq \hat{\delta}$. □

Part (v) of Definition 2.1 will be used in proofs in conjunction with the Lebourg mean value theorem [L1]. For this purpose the following equivalent statement is useful.

LEMMA 2.1. *Condition* (v) *of Definition* 2.1 *holds if and only if given any* $\hat{x} \in \mathbb{R}^n$, $\hat{\varepsilon} > 0$ *and* $\hat{\delta} > 0$, *there exists a* $\hat{\rho} > 0$ *such that for any* $x', x'' \in B(\hat{x}, \hat{\rho})$ *and any* $\eta' \in \partial f(x')$, *there exists an* $\eta'' \in G_{\hat{\varepsilon}} f(x'')$ *such that* $\|\eta'' - \eta'\| \leqq \hat{\delta}$.

Proof. $\Leftarrow$ Let $x' = \hat{x}$ and $x = x''$ in the above statement; then we see that (v) in Definition 2.1 is satisfied.

$\Rightarrow$ Suppose (v) in Definition 2.1 holds. Then given $\hat{x}$, $\hat{\varepsilon} > 0$ and $\hat{\delta} > 0$, there exists a $\hat{\rho} > 0$ such that

(i) for any $\hat{\eta} \in \partial f(\hat{x})$ and $x'' \in B(\hat{x}, \hat{\rho})$, there exists an $\eta'' \in G_{\hat{\varepsilon}} f(x'')$ such that

$$\|\hat{\eta} - \eta''\| \leqq \hat{\delta}/2;$$

(ii) for any $x' \in B(\hat{x}, \hat{\rho})$ and $\eta' \in \partial f(x')$, by u.s.c. of $\partial f(\cdot)$, there exists an $\hat{\eta} \in \partial f(\hat{x})$ such that

$$\|\hat{\eta} - \eta'\| \leqq \delta/2.$$

Hence, for any $x', x'' \in B(\hat{x}, \hat{\rho})$ and $\eta' \in \partial f(x')$ there exist $\hat{\eta} \in \partial f(\hat{x})$ and $\eta'' \in G_{\varepsilon} f(x'')$ such that

$$\|\eta' - \eta''\| = \|(\eta' - \hat{\eta}) + (\hat{\eta} - \eta'')\| \leqq \hat{\delta}. \qquad \square$$

We note that the property in Lemma 2.1 was referred to in [P8] as "upper-semi-continuity of $G_{\varepsilon} f(\cdot)$ with respect to $\partial f(\cdot)$" and was found very useful in establishing optimality conditions for minimizing sequences.

The simplest known example of a family of c.d.f. maps (see [P7]) for a function $f(\cdot)$ are the maps $\partial_{\varepsilon} f(x)$ defined by

$$\partial_{\varepsilon} f(x) \triangleq \operatorname*{co}_{x' \in B(x, \varepsilon)} \{\partial f(x')\}. \tag{2.4}$$

It is obvious by inspection that they satisfy the properties (i)-(v) in Definition 2.1.

Let $\nu \in (0, 1)$, $\varepsilon_0 > 0$, $\delta > 0$ be given. We define the set $\mathscr{E}$ by

$$\mathscr{E} \triangleq \{0\} \cup \{\varepsilon_0, \nu\varepsilon_0, \nu^2\varepsilon_0, \cdots\} \tag{2.5}$$

and, given a family of c.d.f. maps $\{G_{\varepsilon} f(\cdot)\}$ for $f(\cdot)$, we define the maps $h_{\varepsilon}: \mathbb{R}^n \times \mathbb{R}^1 \to \mathbb{R}^n$ and $\varepsilon: \mathbb{R}^n \to \mathscr{E}$ as follows:

$$h_{\varepsilon}(x) \triangleq -\arg\min \{\tfrac{1}{2}\|v\|^2 \mid v \in G_{\varepsilon} f(x)\}, \tag{2.6}$$

$$\varepsilon(x) \triangleq \max \{\varepsilon \in \mathscr{E} \mid \|h_{\varepsilon}(x)\|^2 \geqq \delta\varepsilon\}. \tag{2.7}$$

The map $\varepsilon(\cdot)$ has the following important property which is crucial to the success of our algorithms.

PROPOSITION 2.1. *For every* $\hat{x} \in \mathbb{R}^n$ *such that* $0 \notin \partial f(\hat{x})$, *there exists a* $\hat{\rho} > 0$ *and* $\hat{\varepsilon} > 0$, $\hat{\varepsilon} \in \mathscr{E}$, *such that* $\varepsilon(x) \geqq \hat{\varepsilon} > 0$ *for all* $x \in B(\hat{x}, \hat{\rho})$.

Proof. Since $G_{\varepsilon} f(x)$ is u.s.c. in (ε, x) at $(0, \hat{x})$ for any $\hat{x} \in \mathbb{R}^n$, it follows that $\|h_{\varepsilon}(x)\|^2$ is l.s.c. in (ε, x) at $(0, \hat{x})$. Since $0 \notin G_0 f(\hat{x})$, $\|h_0(\hat{x})\| = \varepsilon^* > 0$. By l.s.c. of $\|h_{\varepsilon}(x)\|^2$, it follows that there exist a $\hat{\rho} > 0$ and an $\hat{\varepsilon} \in (0, \varepsilon^*/2]$, $\hat{\varepsilon} \in \mathscr{E}$, such that $\|h_{\hat{\varepsilon}}(x)\|^2 \geqq \varepsilon^*/2 \geqq \hat{\varepsilon}$ for all $x \in B(\hat{x}, \hat{\rho})$. Hence $\varepsilon(x) \geqq \hat{\varepsilon}$ for all $x \in B(\hat{x}, \hat{\rho})$. $\square$

We now proceed to state an algorithm model.

ALGORITHM MODEL 2.1.

Parameters: $\delta > 0$, $\varepsilon_0 > 0$ (for $\varepsilon(x)$); $\alpha, \beta \in (0, 1)$ (for Armijo step size rule). A family $\{G_{\varepsilon} f(\cdot)\}_{\varepsilon \geqq 0}$ of c.d.f. maps.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$.

Step 1: Compute $\varepsilon(x_i)$ and $h_i \triangleq h_{\varepsilon(x_i)}(x_i)$.
If $\varepsilon(x_i)=0$, stop.

Step 2: Compute the largest $\lambda_i=\beta^{k_i}$, $k_i\in\mathbb{N}_+$ such that

$$f(x_i+\lambda_i h_i)-f(x_i)\leqq -\lambda_i\alpha\delta\varepsilon(x_i). \tag{2.8}$$

Step 3: Set $x_{i+1}=x_i+\lambda_i h_i$, set $i=i+1$ and go to Step 1.

THEOREM 2.1. *Let* $\{x_i\}$ *be a sequence constructed by Algorithm Model* 2.1.

a) *If* $\{x_i\}$ *is finite, with last element* x_k, *then* $0\in\partial f(x_k)$.

b) *If* $\{x_i\}$ *if infinite, then for any accumulation point* $\hat{x}$ *of* $\{x_i\}$, $0\in\partial f(\hat{x})$ *holds.*

Proof. a) Since $\varepsilon(x_k)=0$ if and only if $0\in\partial f(x_k)$, this part of the theorem is clearly true.

b) Suppose $\{x_i\}$ is infinite and that $x_i\to^K\hat{x}$, with $K\subset\{0,1,2,\cdots\}$ infinite and $0\notin\partial f(\hat{x})$. Then by Proposition 2.1 there exists an i_0 such that $\varepsilon(x_i)\geqq\nu\varepsilon(\hat{x})>0$ for all $i\in K$, $i\geqq i_0$. Since the sets $G_{\varepsilon_0}f(x_i)$ are bounded on bounded sets and $G_{\varepsilon(x_i)}f(x_i)\subset G_{\varepsilon_0}f(x_i)$ by (ii) of Definition 2.1, it follows that there exists a $b\in(0,\infty)$ such that $\nu\varepsilon(\hat{x})\delta\leqq\|h_i\|^2\leqq b$ for all $i\in K$, $i\geqq i_0$. Next, by the mean value theorem of Lebourg [L1], for $\lambda\geqq 0$,

$$f(x_i+\lambda h_i)-f(x_i)=\lambda\langle h_i,\xi_{i\lambda}\rangle \tag{2.9}$$

with $\xi_{i\lambda}\in\partial f(x_i+s\lambda h_i)$ and $s\in(0,1)$. Referring to Lemma 2.1, let $\hat{\delta}=(1-\alpha)[\nu\varepsilon(\hat{x})\delta]^{1/2}$. Then there exists a $\hat{\rho}>0$ such that for all x', $x''\in B(\hat{x},\hat{\rho})$, given any $\eta'\in\partial f(x')$, there exists an $\eta''\in G_{\nu\varepsilon(\hat{x})}f(x'')$ such that $\|\eta''-\eta'\|\leqq(1-\alpha)[\nu\varepsilon(\hat{x})\delta]^{1/2}$. Now let $\hat{\lambda}=\beta^{\hat{k}}\leqq\hat{\rho}/2b$, so that if $x_i\in B(\hat{x},\hat{\rho}/2)$, then $(x_i+s\hat{\lambda}h_i)\in B(\hat{x},\hat{\rho})$ for all $s\in(0,1)$. Then there exists an $i_1\geqq i_0$, such that for all $i\in K$, $i\geqq i_1$,

$$f(x_i+\beta^{\hat{k}}h_i)-f(x_i)=\beta^{\hat{k}}\langle h_i,\xi_{i\hat{\lambda}}\rangle=\beta^{\hat{k}}[\langle h_i,\bar{\xi}_{i\hat{\lambda}}\rangle+\langle h_i,\xi_{i\hat{\lambda}}-\bar{\xi}_{i\hat{\lambda}}\rangle] \tag{2.10}$$

with $\bar{\xi}_{i\hat{\lambda}}\in G_{\nu\varepsilon(\hat{x})}f(x_i)\subset G_{\varepsilon(x_i)}f(x_i)$ such that $\|\xi_{i\hat{\lambda}}-\bar{\xi}_{i\hat{\lambda}}\|\leqq(1-\alpha)(\nu\varepsilon(\hat{x})\delta)^{1/2}\leqq(1-\alpha)(\varepsilon(x_i)\delta)^{1/2}\leqq(1-\alpha)\|h_i\|$. Since $\langle h_i,\bar{\xi}_{i\hat{\lambda}}\rangle\leqq-\|h_i\|^2$ by construction of h_i, (2.10) now yields that

$$\begin{aligned} f(x_i+\beta^{\hat{k}}h_i)-f(x_i)&\leqq\beta^{\hat{k}}[-\|h_i\|^2+\|h_i\|(1-\alpha)(\varepsilon(x_i)\delta)^{1/2}] \\ &\leqq-\beta^{\hat{k}}\alpha\|h_i\|^2\leqq-\beta^{\hat{k}}\alpha\delta\varepsilon(x_i). \end{aligned} \tag{2.11}$$

Hence for all $i\in K$, $i\geqq i_1$, we must have $\lambda_i\geqq\beta^{\hat{k}}$ and therefore for all $i\in K$, $i\geqq i_1$

$$f(x_{i+1})-f(x_i)\leqq-\beta^{\hat{k}}\alpha\delta\nu\varepsilon(\hat{x}). \tag{2.12}$$

Since $\{f(x_i)\}_{i=0}^{\infty}$ is a monotonic decreasing sequence by construction, (2.12) implies that $f(x_i)\to-\infty$ as $i\to\infty$. But by continuity of $f(\cdot)$, and the monotonicity of $\{f(x_i)\}_{i=0}^{\infty}$, we must have that $f(x_i)\to f(\hat{x})$ as $i\to\infty$, and hence we have a contradiction, which completes our proof. □

As we have pointed out earlier, the maps $G_\varepsilon f(x)\triangleq\partial_\varepsilon f(x)$ defined in (2.4) are c.d.f. maps. Unfortunately, (see [M1], [P7]), only when $f(\cdot)$ is convex or, more generally, semi-smooth do we know how to construct an adequate approximation to $\arg\min\{\|h\|\,|\,h\in\partial_\varepsilon f(x)\}$; consequently implementable algorithms based on $\partial_\varepsilon f(x)$ have been proposed only for these cases.

We now turn to a special class of locally Lipschitz functions $f(\cdot)$ for which it is easy to determine whether any given point x is a point of differentiability or not. For such functions, it is possible to construct much nicer c.d.f. maps than $\partial_\varepsilon f(x)$. An examination of the literature shows that this construction involves the use of the generalized gradients of locally Lipschitz perturbation functions $\hat{f}_v(\cdot)$.

The class of functions we are about to consider have the form

$$f(x) = \phi(g(x)) \tag{2.13}$$

where $g: \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable and $\phi: \mathbb{R}^m \to \mathbb{R}^1$ is locally Lipschitz. We note that by the chain rule [C1]

$$\partial f(x) \subset G\hat{f}_0(x) \triangleq \left\{ \xi \,\middle|\, \xi = \frac{\partial g(x)^T}{\partial x} y,\ y \in \partial\phi(z),\ z = g(x) \right\}. \tag{2.14}$$

We now introduce the family of perturbation functions $\{\hat{f}_v(\cdot)\}_{v\in\mathbb{R}^m}$, $\hat{f}_v: \mathbb{R}^n \to \mathbb{R}^1$ defined by

$$\hat{f}_v(x) \triangleq \phi(g(x) + v). \tag{2.15}$$

We note that $\hat{f}_v(\cdot)$ is locally Lipschitz continuous and that by the chain rule [C1], $\partial\hat{f}_v(x) \subset G\hat{f}_v(x)$, where

$$G\hat{f}_v(x) \triangleq \left\{ \xi \,\middle|\, \xi = \frac{\partial g(x)^T}{\partial x} y,\ y \in \partial\phi(z),\ z = g(x) + v \right\}. \tag{2.16}$$

We will show that there are a number of functions $f(\cdot)$, of the form (2.13), for which, given x, $\varepsilon \geqq 0$, it is possible to define vectors $v_\varepsilon(x)$ such that $\|v_\varepsilon(x)\| \leqq K\varepsilon$ (with K fixed) and $G_\varepsilon f(x) \triangleq G\hat{f}_{v_\varepsilon(x)}(x)$ are c.d.f. maps. Clearly, we will need the following hypothesis.

Assumption 2.1. For all $x \in \mathbb{R}^n$, $G\hat{f}_0(x) = \partial f(x)$. □

The various known rules for constructing the vectors $v_\varepsilon(x)$ can be traced as being derived from those for the function

$$f(x) \triangleq \max_{j\in\mathbf{m}} g^j(x) \tag{2.17a}$$

where $\mathbf{m} \triangleq \{1, 2, \cdots, m\}$ and $g^j: \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable. For any $x \in \mathbb{R}^n$ and $\varepsilon \geqq 0$, let

$$I_\varepsilon(x) \triangleq \{j \in \mathbf{m} \,|\, f(x) - g^j(x) \leqq 2\varepsilon\}. \tag{2.17b}$$

Then

$$\partial f(x) = \operatorname*{co}_{j\in I_0(x)} \{\nabla g^j(x)\}, \tag{2.17c}$$

and Assumption 2.1 holds. Since for any $v \in \mathbb{R}^m$, $\hat{f}_v(x) = \max_{j\in\mathbf{m}} (g^j(x) + v^j)$, if we define $v_\varepsilon^j(x) \triangleq (f(x) - g^j(x) - \varepsilon)$ for all $j \in I_\varepsilon(x)$ and set $v_\varepsilon^j(x) = 0$ otherwise, we find that

$$G\hat{f}_{v_\varepsilon(x)}(x) = \operatorname*{co}_{j\in I_\varepsilon(x)} \{\nabla g^j(x)\} \tag{2.17d}$$

and that for all $v \in \mathbb{R}^m$ such that $\|v\|_\infty \leqq \varepsilon$, $G\hat{f}_v(x) \subset G\hat{f}_{v_\varepsilon(x)}(x)$. We now consider the class of functions of the form (2.13) for which a similar fact holds. We shall give some additional examples later.

PROPOSITION 2.2. *Let $f: \mathbb{R}^n \to \mathbb{R}^1$ be of the form* (2.13), *satisfying Assumption* 2.1, *and let $\|\cdot\|$ be some norm on $\mathbb{R}^m$. Suppose that for all $x \in \mathbb{R}^n$ and $\varepsilon > 0$ there exists a $v_\varepsilon(x) \in \mathbb{R}^m$ such that $\|v_\varepsilon(x)\| \leqq \varepsilon$ and for all $v \in \mathbb{R}^n$ satisfying $\|v\| \leqq \varepsilon$ we have*

$$G\hat{f}_v(x) \subset G\hat{f}_{v_\varepsilon(x)}(x), \tag{2.18}$$

where $G\hat{f}_v(\cdot)$ was defined in (2.16). *Then $G_\varepsilon f(x) = G\hat{f}_{v_\varepsilon(x)}(x)$ defines a family of c.d.f. maps, with $v_0(x) = 0$ by definition.*

Proof. (i). We refer to Definition 2.1. Because of Assumption 2.1, it is clear that $\partial f(x) = G\hat{f}_0(x) = G_0 f(x)$ for all $x \in \mathbb{R}^n$. (ii) By construction, see (2.18), it follows that if $\varepsilon' > \varepsilon \geqq 0$ then for any $x \in \mathbb{R}^n$, $G_\varepsilon f(x) \subset G_{\varepsilon'} f(x)$. (iii) For any $\varepsilon \geqq 0$, $G_\varepsilon f(x)$ is obviously convex, and bounded on bounded sets. (iv) We now show that $G_\varepsilon f(x)$ is u.s.c. in (ε, x) at $(0, \hat{x})$ for any $\hat{x} \in \mathbb{R}^n$. Let $\hat{x} \in \mathbb{R}^n$ and $\hat{\delta} > 0$ be given and let $\hat{z} \triangleq g(\hat{x})$. Since $\partial\phi(\cdot)$ is u.s.c., there exists a $\rho > 0$ such that $\partial\phi(z) \subset N_{\hat{\delta}}(\partial\phi(\hat{z}))$ for all $z \in \mathbb{R}^m$ such that $\|z - \hat{z}\| \leqq \rho$, with $N_{\hat{\delta}}(\partial\phi(\hat{z}))$ a $\hat{\delta}$-neighborhood of $\partial\phi(\hat{z})$. Since $g(\cdot)$ is continuously differentiable and $\|v_\varepsilon(x)\| \leqq \varepsilon$ for any $x \in \mathbb{R}^n$, it follows that there exists a $\hat{\rho} > 0$ such that with $\hat{\varepsilon} = \frac{1}{2}\hat{\rho} > 0$

$$\|g(x) + v_\varepsilon(x) - g(\hat{x})\| \leqq \hat{\rho} \tag{2.19}$$

for all $x \in B(\hat{x}, \hat{\rho})$ and $\varepsilon \in [0, \hat{\varepsilon}]$. Consequently, $G_\varepsilon f(x)$ is u.s.c. in (ε, x) at $(0, \hat{x})$. We now show that property (v) of Definition 2.1 holds. Let $\hat{x} \in \mathbb{R}^n$, $\hat{\varepsilon} > 0$ and $\hat{\delta} > 0$ be given. Then, since $g(\cdot)$ is continuous, there exists a $\rho^* > 0$ such that for any $x', x'' \in B(\hat{x}, \rho^*)$, $\|g(x') - g(x'')\| \leqq \hat{\varepsilon}$. Let $\eta' \in \partial f(x') = G_0 f(x')$, then $\eta' = (\partial g(x')^T/\partial x)y'$ for some $y' \in \partial\phi(g(x'))$. Now, by definition of $v_{\hat{\varepsilon}}(\cdot)$

$$G\hat{f}_v(x'') \subset G_{\hat{\varepsilon}} f(x'') \quad \text{for all } \|v\| \leqq \hat{\varepsilon}. \tag{2.20}$$

Letting $v^* = g(x') - g(x'')$, we find that

$$\begin{aligned} G\hat{f}_{v^*}(x'') &= \left\{ \eta \;\middle|\; \eta = \frac{\partial g(x'')^T}{\partial x} y,\ y \in \partial\phi(g(x'') + [g(x') - g(x'')]) \right\} \\ &= \left\{ \eta \;\middle|\; \eta = \frac{\partial g(x'')^T}{\partial x} y,\ y \in \partial\phi(g(x')) \right\}. \end{aligned} \tag{2.21a}$$

Since $\|v^*\| \leqq \hat{\varepsilon}$, for $y' \in \partial\phi(g(x'))$ as above, we must have $(\partial g(x'')^T/\partial x)y' \in G\hat{f}_{v^*}(x'') \subset G_{\hat{\varepsilon}} f(x'')$. Now $\eta'' \triangleq (\partial g(x'')^T/\partial x)y' \in G_{\hat{\varepsilon}} f(x'')$ and

$$\|\eta' - \eta''\| \leqq \left\| \left[\frac{\partial g(x')^T}{\partial x} - \frac{\partial g(x'')^T}{\partial x} \right] \right\| \|y'\|. \tag{2.21b}$$

Since $\partial\phi(\cdot)$ is bounded on bounded sets and $\partial g(\cdot)/\partial x$ is uniformly continuous on $B(\hat{x}, \rho^*)$, it follows from (2.21b) that there is a $\hat{\rho} \in (0, \rho^*]$ such that for any $x', x'' \in B(\hat{x}, \hat{\rho})$, given an $\eta' \in \partial f(x')$ there exists a $\eta'' \in G_{\hat{\varepsilon}} f(x'')$ such that $\|\eta' - \eta''\| \leqq \hat{\delta}$. This completes our proof. □

Apart from the function $f(x)$ defined in (2.17a), which satisfies the assumptions of Proposition 2.2, we can cite the following two interesting examples which also fall within the framework of Proposition 2.2.[1]

Consider the function

$$f(x) = \sum_{j \in \mathbf{m}} |g^j(x)| \tag{2.22}$$

where the $g^j: \mathbb{R}^n \to \mathbb{R}^1$ are continuously differentiable. For any $v \in \mathbb{R}^m$, $\hat{f}_v(x) \triangleq \sum_{j \in \mathbf{m}} |g^j(x) + v^j|$ and hence, given any $\hat{x} \in \mathbb{R}^n$ and $\varepsilon > 0$, if we define $v_\varepsilon^j(\hat{x}) = -g^j(\hat{x})$ if $|g^j(\hat{x})| \leqq \varepsilon$ and set $v_\varepsilon^j(\hat{x}) = 0$ otherwise, we find that $G\hat{f}_v(\hat{x}) \subset G\hat{f}_{v_\varepsilon(\hat{x})}(\hat{x})$ for all $v \in \mathbb{R}^m$

[1] All the examples that we give involve semi-smooth functions [M2]. However, we do not seem to need this fact explicitly.

such that $\|v\|_\infty \leqq \varepsilon$. This is clear from the fact that, with $\hat{z} = g(\hat{x})$,

$$G\hat{f}_v(\hat{x}) = \sum_{j \notin J(\hat{z}+v)} \operatorname{sgn}[g^j(x)] \nabla g^j(\hat{x}) + \sum_{j \in J(\hat{z}+v)} \operatorname{co}\{\nabla g^j(\hat{x}), -\nabla g^j(\hat{x})\} \tag{2.23}$$

where $J(\hat{z}+v) \triangleq \{j \in \mathbf{m} \,|\, |\hat{z}^j + v^j| = 0\}$.

Finally consider the function

$$f(x) = \max_{\omega \in \Omega} \zeta(x, \omega) \tag{2.24}$$

where $\zeta: \mathbb{R}^n \times \mathbb{R}^1 \to \mathbb{R}^1$ is continuously differentiable and $\Omega \subset \mathbb{R}^1$ is a compact interval. In this case the function $g(x)(\cdot) \triangleq \zeta(x, \cdot)$ assumes values not in $\mathbb{R}^m$, but in $L_\infty(\Omega)$. For any $v \in L_\infty(\Omega)$, we define

$$\hat{f}_v(x) = \max_{\omega \in \Omega} [\zeta(x, \omega) + v(\omega)] \tag{2.25a}$$

and obtain that

$$G\hat{f}_v(x) = \operatorname*{co}_{\omega \in \hat{\Omega}_v(x)} \{\nabla_x \zeta(x, \omega)\}, \tag{2.25b}$$

where $\hat{\Omega}_v(x) \triangleq \{\omega \in \Omega \,|\, \hat{f}_v(x) = \zeta(x, \omega) + v(\omega)\}$. Clearly, if we set $v_\varepsilon(x)(\omega) = \hat{f}(x) - \zeta(x, \omega) - \varepsilon$ for all $\omega \in \tilde{\Omega}_\varepsilon(x) \triangleq \{\omega \in \Omega \,|\, f(x) - \zeta(x, \omega) \leqq 2\varepsilon\}$, and $v_\varepsilon(x)(\omega) = 0$ for all other $\omega \in \Omega$, we find that $G\hat{f}_v(x) \subset G\hat{f}_{v_\varepsilon(x)}(x)$ for all $v \in L_\infty(\Omega)$ such that $\|v\|_\infty \leqq \varepsilon$. This results in

$$G_\varepsilon f(x) = \operatorname*{co}_{\omega \in \tilde{\Omega}_{2\varepsilon}(x)} \{\nabla_x \zeta(x, \omega)\}. \tag{2.26a}$$

The above set may have an infinite number of elements and hence is not a convenient set to use for finding descent directions. Referring to [G2], we find that when $\tilde{\Omega}_0(x)$ is a finite set for all $x \in \mathbb{R}^n$, it is possible to use the much smaller set

$$G_\varepsilon f(x) \triangleq \operatorname*{co}_{\omega \in \Omega_\varepsilon(x)} \{\nabla_x \zeta(x, \omega)\} \tag{2.26b}$$

where $\Omega_\varepsilon(x) \triangleq \{\omega \in \tilde{\Omega}_\varepsilon(x) \,|\, \omega$ is a local maximizer of $\zeta(x, \cdot)\}$. It is easy to see that $G_\varepsilon f(x)$ corresponds to the perturbation function $\hat{f}_{v_\varepsilon(x)}(\cdot)$, with $v_\varepsilon(x)(\omega) = f(x) - \zeta(x, \omega)$ for all $\omega \in \Omega_\varepsilon(x)$ and is arbitrary otherwise up to the requirement that $\hat{\Omega}_{v_\varepsilon(x)}(x) = \Omega_\varepsilon(x)$. Quite clearly, $G_\varepsilon f(x)$, as defined in (2.26a) does not satisfy the assumptions of Proposition 2.2. However, showing that the maps $G_\varepsilon f(\cdot)$, defined by (2.26a), are c.d.f. maps is a great deal simpler than the original proof of convergence in [G2], as we now show.

PROPOSITION 2.3. *Consider the function $f(\cdot)$ defined by* (2.24) *with $\Omega = [\omega_0, \omega_f]$. Suppose that for every $x \in \mathbb{R}^n$ $\zeta_\omega(x, \bar{\omega}) = 0$ for at most a finite number of $\bar{\omega} \in \Omega$ and that $\zeta_\omega(x, \omega) \neq 0$ for $\omega \in \{\omega_0, \omega_f\}$ (where $\zeta_\omega \triangleq \partial\zeta/\partial\omega$). Then the maps $\{G_\varepsilon f(\cdot)\}_{\varepsilon \geqq 0}$ defined by* (2.26b) *form a family of c.d.f. maps.*

The proof of Proposition 2.3 requires the following three facts which we establish first.

FACT 2.1. *For any $\varepsilon > 0$, $x \in \mathbb{R}^n$, let*

$$\check{\Omega}_\varepsilon(x) = \{\omega \in \tilde{\Omega}_\varepsilon(x) \,|\, \zeta_\omega(x, \omega) = 0\} \cup [\{\omega_0, \omega_f\} \cap \tilde{\Omega}_\varepsilon(x)]. \tag{2.27}$$

Then $\check{\Omega}_\varepsilon(\cdot)$ is u.s.c.

Proof. Suppose $x_i \to \hat{x}$ as $i \to \infty$ and $\omega_i \in \check{\Omega}_\varepsilon(x_i)$ are such that $\omega_i \to \hat{\omega}$ as $i \to \infty$. Then (i) $f(x_i) - \zeta(x_i, \omega_i) \leqq \varepsilon$ for all i and hence, by continuity of $f(\cdot)$ and $\zeta(\cdot, \cdot)$, $\hat{\omega} \in \tilde{\Omega}_\varepsilon(x)$,

and (ii) either by continuity of $\zeta_\omega(\cdot,\cdot)$ $\zeta_\omega(\hat{x},\hat{\omega})=0$ or $\hat{\omega}\in\{\omega_0,\omega_f\}$. Thus $\hat{\omega}\in\check{\Omega}_\varepsilon(\hat{x})$, which completes the proof. □

FACT 2.2. *Given* $\hat{x}\in\mathbb{R}^n$, $\hat{\varepsilon}>0$, *there exists a* $\hat{\rho}>0$ *such that for any* $x', x''\in B(\hat{x},\hat{\rho})$, $\Omega_0(x')\subset\tilde{\Omega}_{\hat{\varepsilon}}(x'')$.

Proof. Since Ω is compact, there exists a $\hat{\rho}>0$ such that for any $x', x''\in B(\hat{x},\hat{\rho})$ if $\omega'\in\Omega$ is such that

$$f(x')-\zeta(x',\omega')=0 \tag{2.28a}$$

then

$$f(x'')-\zeta(x'',\omega')\leqq\varepsilon, \tag{2.28b}$$

i.e. $\Omega_0(x')\subset\tilde{\Omega}_\varepsilon(x'')$. □

The following result is obvious.

FACT 2.3. *Let* $\mu_\varepsilon(x)\triangleq\text{meas}(\tilde{\Omega}_\varepsilon(x))$. *Then* (i) $\varepsilon'<\varepsilon''\Rightarrow\mu_{\varepsilon'}(x)\leqq\mu_{\varepsilon''}(x)$, *and* (ii) $\mu_\varepsilon(\cdot)$ *is continuous.*

Proof of Proposition 2.3. Referring to Definition 2.1, we find that (i) $\partial f(x)=G_0 f(x)$ and (ii) that $\varepsilon<\varepsilon'\Rightarrow G_\varepsilon f(x)\subset G_{\varepsilon'}f(x)$, by construction in (2.26b). (iii) Clearly, $G_\varepsilon(x)$ is always convex and bounded and bounded sets. (iv) Next, let $x\in\mathbb{R}^n$ be arbitrary. Since by assumption the set $\check{\Omega}_\varepsilon(\hat{x})$ is finite for all $\varepsilon\geqq 0$, there exists an $\hat{\varepsilon}>0$ such that $\Omega_\varepsilon(\hat{x})=\check{\Omega}_\varepsilon(\hat{x})$ for all $\varepsilon\in[0,\hat{\varepsilon}]$. It now follows from the u.s.c. of $\check{\Omega}_\varepsilon(\cdot)$ that $\Omega_\varepsilon(\cdot)$ is u.s.c. at $\hat{x}$ for all $\varepsilon\in[0,\hat{\varepsilon}]$ and hence, from the continuity of $\nabla_x\zeta(\cdot,\cdot)$, it follows that $G_\varepsilon f(x)$ is u.s.c. at $(0,\hat{x})$. We now turn to property (v) in Definition 2.1. Let $x\in\mathbb{R}^n$, $\hat{\varepsilon}>0$ and $\hat{\delta}>0$ be given. Then (a) there exists a $\rho_1>0$ such that for all $x', x''\in B(\hat{x},\rho_1)$ and $\omega',\omega''\in[\omega_0,\omega_f]$ satisfying $|\omega'-\omega''|\leqq\rho_1$, we have

$$\|\nabla_x\zeta(x',\omega')-\nabla\zeta_x(x'',\omega'')\|\leqq\hat{\delta}. \tag{2.29}$$

(b) Since $\Omega_0(\hat{x})$ is a discrete set, there exists an $\varepsilon_1\in(0,\hat{\varepsilon}]$ such that $\mu_{\varepsilon_1}(\hat{x})\leqq\rho_1/2$. Hence, by continuity of $\mu_{\varepsilon_1}(\cdot)$, there exists a $\rho_2\in(0,\rho_1]$ such that $\mu_{\varepsilon_1}(x)\leqq\rho_1$ for all $x\in B(\hat{x},\rho_2)$. (c) By Fact 2.2, there exists a $\hat{\rho}\in(0,\rho_2]$ such that $\Omega_0(x')=\tilde{\Omega}_0(x')\subset\tilde{\Omega}_{\varepsilon_1}(x'')$ for all $x', x''\in B(\hat{x},\hat{\rho})$. (d) Now consider any $x', x''\in B(\hat{x},\hat{\rho})$. If $\eta'\in\partial f(x')$, then

$$\eta'=\sum_{\omega'\in\Omega_0(x')}\mu^{\omega'}\nabla_x\zeta(x',\omega') \tag{2.30}$$

where $\mu^{\omega'}\geqq 0$ and $\sum_{\omega'\in\Omega_0(x')}\mu^{\omega'}=1$. By (c) $\Omega_0(x')\subset\tilde{\Omega}_{\varepsilon_1}(x'')\subset\tilde{\Omega}_{\hat{\varepsilon}}(x'')$ and $\mu_{\varepsilon_1}(x'')\leqq\rho_1$. Since every disjoint interval of $\tilde{\Omega}_{\varepsilon_1}(x'')$ must contain at least one $\omega''\in\Omega_{\varepsilon_1}(x'')$, it follows that for every $\omega'\in\Omega_0(x')$ there exists an $\omega_{\omega'}\in\Omega_{\varepsilon_1}(x'')\subset\Omega_{\hat{\varepsilon}}(x'')$ such that $|\omega'-\omega_{\omega'}|\leqq\rho_1$. Hence the vector $\eta''\in G_{\hat{\varepsilon}}(x'')$ defined by

$$\eta''=\sum_{\omega'\in\Omega_0(x')}\mu^{\omega'}\nabla_x\zeta(x'',\omega_{\omega'}) \tag{2.31}$$

satisfies

$$\|\eta''-\eta'\|\leqq\sum_{\omega'\in\Omega_0(x')}\mu^{\omega'}\|\nabla_x\zeta(x'',\omega_{\omega'})-\nabla_x\zeta(x',\omega')\|\leqq\hat{\delta} \tag{2.32}$$

which completes our proof. □

Next we turn to problems involving eigenvalues of Hermitian matrices, see e.g. [C3], [P6]. We shall consider only one case. Let $Q:\mathbb{R}^n\to\mathbb{C}^{m\times m}$ be a continuously differentiable, complex matrix valued function such that $Q(x)$ is Hermitian for all x. For any $m\times m$ Hermitian matrix M, we denote its eigenvalues as $\sigma^1[M]\geqq\sigma^2[M]\cdots\geqq\sigma^m[M]$ and we consider the case where

$$f(x)\triangleq\sigma^1[Q(x)]. \tag{2.33}$$

Thus for any Hermitian matrix $V \in \mathbb{C}^{m\times m}$, we define the perturbation function $\hat{f}_v(\cdot)$ by

$$\hat{f}_V(x) = \sigma^1[Q(x)+V]. \tag{2.34}$$

We proceed by analogy with the example in (2.17) in defining a "maximal" perturbation matrix $V_\varepsilon(x)$. Clearly, there exists a matrix of complex orthonormal left eigenvectors $U(x)$ such that $U^*(x)U(x) = I$ and

$$Q(x) = U^*(x)\Sigma(x)U(x) \tag{2.35}$$

where $\Sigma(x) \triangleq \operatorname{diag}(\sigma^1[Q(x)], \cdots, \sigma^m[Q(x)])$. Given $\varepsilon > 0$, we define $V_\varepsilon(x)$ by

$$V_\varepsilon(x) \triangleq U^*(x)\Lambda_\varepsilon(x)U(x) \tag{2.36}$$

where $\Lambda_\varepsilon(x) \triangleq \operatorname{diag}(\lambda^i_\varepsilon(x))$, with $\lambda^i_\varepsilon(x) = \sigma^1[Q(x)] - \sigma^i[Q(x)] - \varepsilon$ for all $i \in I_\varepsilon[Q(x)]$ and $\lambda^i_\varepsilon(x) = 0$ otherwise, where $I_\varepsilon[Q(x)] \triangleq \{i \in \mathbf{m} \mid \sigma^1[Q(x)] - \sigma^i[Q(x)] \leqq 2\varepsilon\}$. This choice of $V_\varepsilon(x)$ clearly "maximizes" the set

$$G\hat{f}_V(x) \triangleq \operatorname{co}\left\{ y \,\middle|\, y^i = \left\langle U_V(x)z, \frac{dQ(x)}{dx^i} U_V(x)z \right\rangle, i = 1, 2, \cdots, n, \|z\| = 1 \right\} \tag{2.37}$$

where, given that $I_0[Q(x)+V] = \{1, 2, \cdots, k_V(x)\}$, $U_V(x)$ consists of the first $k_V(x)$ columns of $U(x)$ (see [P6] for a proof that $\partial f(x) = G\hat{f}_0(x)$). We claim that $G_\varepsilon f(x) \triangleq G\hat{f}_{v_\varepsilon(x)}(x)$, as defined by (2.36) and (2.37) is a c.d.f. map, but we omit a proof, which can be constructed by referring to [P6]. We note that in [P6] a somewhat *larger* set $G_\varepsilon f(x)$ was used so as to avoid computational difficulties caused by the need to distinguish between eigenvalues that are very close to being equal. We find that in [P6], for any $\varepsilon \geqq 0$, $x \in \mathbb{R}^n$

$$k_\varepsilon(x) \triangleq \max_{k\in\mathbf{m}} (k+1 \mid \sigma^i[Q(x)] - \sigma^{i+1}[Q(x)] \leqq 2\varepsilon \text{ for all } i \leqq k\} \tag{2.38a}$$

which leads to the definition

$$G_\varepsilon f(x) \triangleq \operatorname{co}\left\{ y \,\middle|\, y^i = \left\langle U_\varepsilon(x)z, \frac{\partial Q(x)}{\partial x^i} U_\varepsilon(x)z \right\rangle, i = 1, 2, \cdots, n, \|z\| = 1 \right\} \tag{2.38b}$$

where $U_\varepsilon(x)$ consists of the first $k_\varepsilon(x)$ columns of $U(x)$. Since $k_\varepsilon(x) \geqq k_{v_\varepsilon(x)}(x)$, it is clear that (2.38b) results in a larger set than (2.37) for $V = V_\varepsilon(x)$; however, the general properties relevant to convergence of the two sets are the same.

This concludes our demonstration that a large number of seemingly unrelated algorithms for various unconstrained nondifferentiable optimization problems can be seen as manifestations of a single, relatively simple principle.

Before proceeding to constrained optimization problems, it remains to point out that when $f(\cdot)$ is a locally Lipschitz continuous function from a Banach space $\mathscr{X}$ into $\mathbb{R}$, one must use an alternate definition of $\partial f(x)$, see [C1]. Also, the computation of a descent direction according to

$$h(x) \in \arg\min_{h\in\mathscr{X}} \{\tfrac{1}{2}\|h\|^2 + d_\varepsilon f(x, h)\} \tag{2.39a}$$

may not be a tractable problem. In that case, (2.39a) may be replaced by

$$h(x) \in \arg\min_{\|h\|\leqq 1} d_\varepsilon f(x, h) = \arg\min_{\|h\|\leqq 1} \max_{\xi(x_i)\in G_\varepsilon f(x)} (\xi, h) \tag{2.39b}$$

where the action of a $\xi \in \mathscr{X}'$, the dual of $\mathscr{X}$, on an $h \in \mathscr{X}$ is denoted by (ξ, h). All the proofs in this section have valid analogs when (2.39a) is replaced by (2.39b).

3. Constrained optimization. In this section we restrict ourselves to problems of the form

$$\min \{f^0(x) \mid f^j(x) \leqq 0, j \in \mathbf{m}\}, \tag{3.1}$$

where $\mathbf{m} \triangleq \{1, 2, \cdots, m\}$ and $f^j: \mathbb{R}^n \to \mathbb{R}^1$, $j \in \{0\} \cup \mathbf{m}$, are locally Lipschitz continuous functions.

We shall assume that we have for all the functions f^j, $j \in \{0\} \cup \mathbf{m}$, families of convergent direction finding maps $\{G_\varepsilon f^j(\cdot)\}_{\varepsilon \geqq 0}$ (see Definition 2.1). We define $\psi: \mathbb{R}^n \to \mathbb{R}^1$ and $\psi(\cdot)_+$, as follows

$$\psi(x) \triangleq \max_{j \in \mathbf{m}} f^j(x), \tag{3.2}$$

$$\psi(x)_+ = \max \{0, \psi(x)\}. \tag{3.3}$$

We are about to state a phase I-phase II algorithm of a form quite similar to the ones treated in [P2], [P7]. First, for any $x \in \mathbb{R}^n$ and $\varepsilon \geqq 0$, we define the ε-most violated constraint index set

$$I_\varepsilon(x) \triangleq \{j \in \mathbf{m} \mid f^j(x) \geqq \psi(x)_+ - \varepsilon\} \tag{3.4a}$$

and we set

$$J_\varepsilon(x) \triangleq \{0\} \cup I_\varepsilon(x). \tag{3.4b}$$

Next, for $\varepsilon \geqq 0$ given, we define the phase I ε-search direction at x by

$$\begin{aligned} h_{\psi\varepsilon}(x) &\triangleq \arg\min_h \{\tfrac{1}{2}\|h\|^2 + \max_{j \in I_\varepsilon(x)} d_\varepsilon f^j(x, h)\} \\ &= -\arg\min_h \{\tfrac{1}{2}\|h\|^2 \mid h \in \operatorname*{co}_{j \in I_\varepsilon(x)} \{G_\varepsilon f^j(x)\}\}, \end{aligned} \tag{3.5}$$

where co denotes the convex hull, and we define the phase II ε-search direction at x by

$$\begin{aligned} h_{f\varepsilon}(x) &\triangleq \arg\min_h \{\tfrac{1}{2}\|h\|^2 + \max_{j \in J_\varepsilon(x)} d_\varepsilon f^j(x, h)\} \\ &= -\arg\min_h \{\tfrac{1}{2}\|h\|^2 \mid h \in \operatorname*{co}_{j \in J_\varepsilon(x)} \{G_\varepsilon f^j(x)\}\}. \end{aligned} \tag{3.6}$$

Finally, we define the cross-over function

$$\Gamma(x) \triangleq e^{-\gamma\psi(x)_+} \tag{3.7}$$

where $\gamma > 0$ is a parameter. It will become clear shortly that when $\psi(x) > 0$, for appropriate values of ε, $h_{\psi_\varepsilon}(x)$ is a descent direction for $\psi(x)$, while when $\psi(x) \leqq 0$, $h_{f\varepsilon}(x)$ is a *feasible* descent direction for $f(\cdot)$. The cross-over from one to the other is incorporated in the search direction

$$h_\varepsilon(x) \triangleq \Gamma(x) h_{f\varepsilon}(x) + (1 - \Gamma(x)) h_{\psi_\varepsilon}(x). \tag{3.8}$$

As we have already seen in the preceding section, we need a mechanism for driving ε to zero. To this end, (cf. [P7, (3.13)]) we define

$$\theta_\varepsilon(x) \triangleq \max \{\|\Gamma(x) h_{f\varepsilon}(x)\|^2, \|(1 - \Gamma(x)) h_{\psi\varepsilon}(x)\|^2\} \tag{3.9}$$

and, with $\mathscr{E}$ as in (2.5), and $\delta > 0$,

$$\varepsilon(x) \triangleq \max \{\varepsilon \in \mathscr{E} \mid \theta_\varepsilon(x) \geqq \delta\varepsilon\}. \tag{3.10}$$

ALGORITHM MODEL 3.1.

Parameters: $\delta > 0$, ε_0 (for $\varepsilon(x)$); $\alpha, \beta \in (0, 1)$ (for Armijo step size rule). Families $\{G_\varepsilon f^j(\cdot)\}_{\varepsilon \geqq 0}$, $j \in \{0\} \cup \mathbf{m}$ of c.d.f. maps.

Data: $x_0 \in \mathbb{R}^n$.

Step 0: Set $i = 0$.

Step 1: Compute $\varepsilon(x_i)$ and $h_i \triangleq h_{\varepsilon(x_i)}(x_i)$.
If $\varepsilon(x_i) = 0$, stop.

Step 2: If $\psi(x_i)_+ > 0$, compute the largest $\lambda_i = \beta^{k_i}$, $k_i \in \mathbb{N}_+$ such that

$$\psi(x_i + \lambda_i h_i) - \psi(x_i) \leqq -\lambda_i \alpha \delta \varepsilon(x_i). \tag{3.11}$$

If $\psi(x_i)_+ \leqq 0$, compute the largest $\lambda_i = \beta^{k_i}$, $k_i \in \mathbb{N}_+$ such that

$$f^0(x_i + \lambda_i h_i) - f^0(x_i) \leqq -\lambda_i \alpha \delta \varepsilon(x_i) \tag{3.12a}$$

and

$$\psi(x_i + \lambda_i h_i) \leqq 0. \tag{3.12b}$$

Step 3: Set $x_{i+1} = x_i + \lambda_i h_i$, set $i = i + 1$ and go to Step 1.

To ensure that the above algorithm does not jam at an infeasible point, we must introduce the following commonly used hypothesis.

Assumption 3.1. For every $x \in \mathbb{R}^n$ such that $\psi(x) \geqq 0$, $h_{\psi 0}(x) \neq 0$. □

This assumption guarantees the existence of a point x^* such that $\psi(x^*) < 0$.

To establish the convergence properties of Algorithm Model 3.1, we shall need the following results.

LEMMA 3.1. *For every* $\varepsilon \geqq 0$ *and any* $x \in \mathbb{R}^n$

$$\|h_\varepsilon(x)\|^2 \geqq \theta_\varepsilon(x). \tag{3.13}$$

For a proof of this lemma, see [P7, Lemma 3.1].

LEMMA 3.2. *The function* $\theta_\varepsilon(x)$ *defined in* (3.9) *has the following properties*: a) *For any* $x \in \mathbb{R}^n$, *if* $\varepsilon'' \geqq \varepsilon' \geqq 0$, *then* $\theta_{\varepsilon''}(x) \leqq \theta_{\varepsilon'}(x)$. b) $\theta_\varepsilon(x)$ *is l.s.c. in* (ε, x) *at any* $(0, \hat{x})$.

Proof. a) Since $\varepsilon'' > \varepsilon'$ implies that $I_{\varepsilon''}(x) \supset I_{\varepsilon'}(x)$ and that $G_{\varepsilon''} f^j(x) \supset G_{\varepsilon'} f^j(x)$, this part is obvious. b) By definition of c.d.f. maps, $G_\varepsilon f^j(x)$ is u.s.c. in (ε, x) at any $(0, \hat{x})$, $j = 0, 1, \cdots, m$. Let $\hat{x} \in \mathbb{R}^n$ be arbitrary. Then, given $\hat{\varepsilon} > 0$, there exists a $\hat{\rho} > 0$ such that $I_\varepsilon(x) \subset I_{\hat{\varepsilon}}(x) \subset I_{\hat{\varepsilon}}(\hat{x})$ for all $(\varepsilon, x) \in [0, \hat{\varepsilon}] \times B(\hat{x}, \hat{\rho})$. Because of this and the u.s.c. of the $G_\varepsilon f^j(x)$ at $(0, \hat{x})$, there exist $\varepsilon^* \in (0, \hat{\varepsilon}]$ and $\rho^* \in (0, \hat{\rho}]$ such that

$$\operatorname*{co}_{j \in I_\varepsilon(x)} \{G_\varepsilon f^j(x)\} \subset \operatorname*{co}_{j \in I_{\hat{\varepsilon}}(\hat{x})} \{G_\varepsilon f^j(x)\} \subset N_\delta\left(\operatorname*{co}_{j \in I_{\hat{\varepsilon}}(\hat{x})} \{G_\varepsilon f^j(\hat{x})\}\right) \tag{3.14}$$

where $N_\delta(\cdot)$ denotes a δ neighborhood of the set in parentheses. Consequently, $\|h_{\psi\varepsilon}(x)\|$ is l.s.c. in (ε, x) at any $(0, \hat{x})$. Similarly, it can be shown that $\|h_{f\varepsilon}(x)\|$ is l.s.c. in (ε, x) at any $(0, \hat{x})$. Since $\Gamma(\cdot)$ is continuous, it follows that $\theta_\varepsilon(x)$ is l.s.c. in (ε, x) at any $(0, \hat{x})$, which completes our proof. □

The following result can be established in essentially the same way as Proposition 2.1 and hence a proof will be omitted.

COROLLARY 3.1. *For every* $\hat{x} \in \mathbb{R}^n$ *such that* $\theta_0(\hat{x}) > 0$, *there exists a* $\hat{\rho} > 0$ *and a* $\hat{\varepsilon} > 0$, $\hat{\varepsilon} \in \mathscr{E}$ *such that* $\varepsilon(x) \geqq \hat{\varepsilon} > 0$ *for all* $x \in B(\hat{x}, \hat{\rho})$.

THEOREM 3.1. *Suppose that Assumption* 3.1 *holds and that* $\{x_i\}$ *is a sequence constructed by Algorithm Model* 3.1. a) *If* $\{x_i\}$ *is finite, with last element* x_k, *then* $\psi(x_k) \leqq 0$ *and*

$$0 \in \operatorname*{co}_{j \in J_0(x_k)} \{\partial f^j(x_k)\} \tag{3.15a}$$

b) *If* $\{x_i\}$ *is infinite, then for any accumulation point* $\hat{x}$ *of* $\{x_i\}$, *we have* $\psi(\hat{x}) \leqq 0$ *and*

$$0 \in \underset{j \in J_0(\hat{x})}{\text{co}} \{\partial f^j(\hat{x})\} \tag{3.15b}$$

Proof. a) Suppose that $\{x_i\}_{i=1}^{k}$ is finite. Then, by construction, $\varepsilon(x_k) = 0$ and hence, by Corollary 3.1, $\theta_0(x_k) = 0$, so that

$$\Gamma(x_k) h_{f0}(x_k) = (1 - \Gamma(x_k)) h_{\psi 0}(x_k) = 0. \tag{3.16}$$

Suppose now that $\Gamma(x_k) < 1$, i.e. $\psi(x_k) > 0$, then (3.16) implies that $h_{\psi 0}(x_k) = 0$. But this contradicts Assumption 3.1 and hence we must have $\psi(x_k) = 0$. Since $\Gamma(x_k) = 1$, $h_{f0}(x_k) = 0$ and hence, since $\partial f^j(x_k) = G_0 f^j(x_k)$, $j = 0, 1, \cdots, m$, we find that (3.15a) must hold. b) Suppose that $\{x_i\}_{i=0}^{\infty}$ has an accumulation point $\hat{x}$, i.e., that $x_i \to^K \hat{x}$, with $K \subset \mathbb{N}_+$ infinite, that $\psi(\hat{x}) \geqq 0$ and (3.15b) fails to hold. We consider two cases.

Case 1. $\psi(x_i) > 0$ for all $i \in \mathbb{N}_+$. Then, by (3.11) $\{\psi(x_i)\}_{i=0}^{\infty}$ is monotone decreasing, and $\psi(x_i) \to^K \psi(\hat{x})$ by continuity of $\psi(\cdot)$. Hence $\psi(x_i) \searrow \psi(\hat{x})$ as $i \to \infty$. We shall now show that this leads to a contradiction and in the process also show that this part of the Algorithm Model 3.1 is well defined.

Since $\psi(x_i) \geqq 0$ for all i, we must have $\psi(\hat{x}) \geqq 0$ and hence $h_{\psi 0}(\hat{x}) \neq 0$ by Assumption 3.1. Consequently, $\theta_0(\hat{x}) > 0$ either because $\psi(\hat{x}) > 0$ or because (3.15b) fails, i.e., because $\psi(\hat{x}) = 0$ and $h_{f0}(\hat{x}) \neq 0$. Thus, by Corollary 3.1, there exists an i_0 such that $\varepsilon(x_i) \geqq \hat{\varepsilon} > 0$ for all $i \in K$, $i \geqq i_0$. Since the sets $G_{\varepsilon_0} f^j(x_i)$ are bounded on bounded sets and $G_{\varepsilon(x_i)} f^j(x_i) \subset G_{\varepsilon_0} f^j(x_i)$ by (ii) of Definition 2.1, it follows, via Lemma 3.1, that there exists a $b \in (0, \infty)$ such that

$$0 < \delta\hat{\varepsilon} < \delta\varepsilon(x_i) \leqq \|h_i\|^2 \leqq b, \tag{3.17}$$

for all $i \in K$, $i \geqq i_0$. Now, by the mean value theorem of Lebourg [L1], for $j \in \mathbf{m}$

$$f^j(x_i + \lambda_i h_i) - \psi(x_i) = [f^j(x_i) - \psi(x_i)] + \lambda_i \langle \xi^j_{\lambda_i}, h_i \rangle, \tag{3.18}$$

where $\xi^j_{\lambda_i} \in \partial f^j(x_i + s\lambda_i h_i)$ with $s \in (0, 1)$. Now, for any $\xi \in G_{\varepsilon(x_i)} f^j(x_i)$, $j \in I_{\varepsilon(x_i)}(x_i)$, we have by construction that

$$\begin{aligned} \langle -h_i, \xi \rangle &= \Gamma(x_i) \langle -h_{f\varepsilon(x_i)}(x_i), \xi \rangle + [1 - \Gamma(x_i)] \langle -h_{\psi\varepsilon(x_i)}(x_i), \xi \rangle \\ &\geqq \Gamma(x_i) \|h_{f\varepsilon(x_i)}(x_i)\|^2 + [1 - \Gamma(x_i)] \|h_{\psi\varepsilon(x_i)}(x_i)\|^2 \\ &\geqq \|h_i\|^2. \end{aligned} \tag{3.19}$$

Hence, proceeding as in the proof of Theorem 2.1, we conclude that there is a $\hat{\lambda}_1 = \beta^{\hat{k}_1}$, $\hat{k}_1 \in \mathbb{N}_+$, such that

$$f^j(x_i + \beta^{\hat{k}_1} h_i) - \psi(x_i) \leqq -\beta^{\hat{k}_1} \alpha \|h_i\|^2 \leqq -b^{\hat{k}_1} \alpha\delta\varepsilon(x_i) \leqq -\beta^{\hat{k}_1} \alpha\delta\hat{\varepsilon} < 0 \tag{3.20a}$$

for all $j \in I_{\varepsilon(x_i)}(x_i)$, (where we have made use of the fact that $f^j(x_i) - \psi(x_i) \leqq 0$). Next, since $f^j(x_i) - \psi(x_i) < -\varepsilon(x_i) \leqq -\hat{\varepsilon} < 0$ for all $j \notin I_{\varepsilon(x_i)}(x_i)$, and since the h_i are bounded for $i \in K$, it follows by uniform continuity of f^j, ψ on bounded sets, that there exists a $0 < \hat{\lambda} = \beta^{\hat{k}} \leqq \hat{\lambda}_1$ such that

$$f^j(x_i + \hat{\lambda} h_i) - \psi(x_i) \leqq -\hat{\lambda} \alpha\delta\varepsilon(x_i) \tag{3.20b}$$

for all $j \notin I_{\varepsilon(x_i)}(x_i)$, $i \in K$, $i \geqq i_0$. Combining (3.20a) and (3.20b) we conclude that $\lambda_i \geqq \hat{\lambda}$ for all $i \in K$, $i \geqq i_0$ and hence that

$$\psi(x_{i+1}) - \psi(x_i) \leqq -\hat{\lambda} \alpha\delta\hat{\varepsilon} \tag{3.21}$$

for all $i \geqq i_0$, $i \in K$. But this contradicts the fact that $\psi(x_i) \searrow \psi(\hat{x})$ and hence we are done.

Case 2. There exists an i_0 such that for all $i \geqq i_0$, $\psi(x_i) \leqq 0$. If $x_i \to^K \hat{x}$ and $\psi(\hat{x}) < -\varepsilon(\hat{x})$, then the theorem follows directly from Theorem 2.1. Hence we only need to consider the case where $\psi(\hat{x}) \geqq -\varepsilon(\hat{x})$. Now, for this case, we conclude from Case 1 above that there is an i_0 and a $\hat{\lambda}_1 = \beta^{\hat{k}_1}$ such that for all $i \geqq i_0$, $i \in K$ and $\lambda \in [0, \hat{\lambda}_1]$

$$\psi(x_i + \lambda h_i) \leqq 0, \tag{3.22}$$

and from the proof of Theorem 2.1 that there exists an $i_1 \geqq i_0$ and a $\hat{\lambda} = \beta^{\hat{k}} \leqq \hat{\lambda}_1$ such that (3.12a) is satisfied for all $i \geqq i_1$, $i \in K$, with $\lambda_i = \hat{\lambda}$. Hence we must have that $\lambda_i \geqq \hat{\lambda}$ and for all $i \geqq i_1$, $i \in K$

$$f^0(x_{i+1}) - f^0(x_i) \leqq -\hat{\lambda}\alpha\delta\hat{\varepsilon}. \tag{3.23}$$

But $f^0(x_i) \searrow f^0(\hat{x})$ since $x_i \to^K \hat{x}$ and $f^0(\cdot)$ is continuous, which is contradicted by (3.23) and hence the proof is complete. □

Since for continuously differentiable functions $f^j(\cdot)$ we may set $G_\varepsilon f^j(x) = \nabla f^j$, it should now be clear that any combination of differentiable functions and functions such as those defined in (2.17), (2.22) and (2.24), (2.33) may appear in the constraints.

Finally, it remains to point out that when the substitute formula (2.39) is used for problems in Banach spaces, (3.5) and (3.6) become replaced by

$$h_{\psi_\varepsilon}(x) \in \arg\min_{\|h\| \leqq 1} \max_{j \in I_\varepsilon(x)} d_\varepsilon f^j(x, h) = \arg\min_{\|h\| \leqq 1} \max_{\substack{\xi \in \operatorname{co}\{G_\varepsilon f^j(x)\} \\ j \in I_\varepsilon(x)}} (\xi, h) \tag{3.24}$$

and

$$h_{f_\varepsilon}(x) \in \arg\min_{\|h\| \leqq 1} \max_{j \in J_\varepsilon(x)} d_\varepsilon f^j(x, h) = \arg\min_{\|h\| \leqq 1} \max_{\substack{\xi \in \operatorname{co}\{G_\varepsilon f^j(x)\} \\ j \in J_\varepsilon(x)}} (\xi, h) \tag{3.25}$$

respectively. Again, the arg min max may be set valued.

4. Conclusion. We have shown that a rather large number of nondifferentiable optimization algorithms can be presented and analyzed in a unified way. We have also shown that for an important class of optimization problems, defined by composite functions, efficient nondifferentiable optimization algorithms can be constructed by using the generalized gradients of perturbation functions. Furthermore we have established rules for the construction of these perturbation functions.

Acknowledgment. The authors wish to thank Dr. C. Lemarechal for pointing out that the property stated in Lemma 2.1 is equivalent to (v) in Definition 2.1.

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 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] J. CULLUM, W. E. DONATH AND P. WOLFE, *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices*, in Nondifferentiable Optimization, Math. Programming Study No. 3, M. L. Balinski and P. Wolfe, eds., North-Holland, Amsterdam, 1975, pp. 35-55.

[G1] A. A. GOLDSTEIN, *Optimization of Lipschitz continuous functions*, Math. Programming, 13 (1977), pp. 14-22.

[G2] C. GONZAGA, E. POLAK AND R. TRAHAN, *An improved algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automatic Control, AC-25 (1979), pp. 49-54.

[L1] C. LEBOURG, *Valeur moyenne pour gradient generalisé*, C.R. Acad. Sci., Paris, 281 (1975), pp. 795-797.

[L2] C. LEMARECHAL, *Nondifferentiable optimization, subgradient and ε-subgradient methods*, Lecture Notes in Optimization and Operations Research, 17, Springer-Verlag, New York, 1976.

[L3] ———, *Minimization of nondifferentiable functions with constraints*, Proc. 12th Allerton Conference on Circuit Theory, Univ. Illinois, Urbana, pp. 16-24, 1974.

[L4] ———, *Extensions diverses des methodes de gradient et applications*, Thèse, Docteur d'Etat, Univ. Paris, 1980.

[M1] R. MIFFLIN, *An algorithm for constrained optimization with semi-smooth functions*, Math. Oper. Res., 2 (1977), pp. 191-207.

[M2] ———, *Semi-smooth and semi-convex functions in constrained optimization*, this Journal, 15 (1977), pp. 959-973.

[M3] D. Q. MAYNE AND E. POLAK, *A quadratically convergent algorithm for solving infinite dimensional inequalities*, Univ. California Electronics Research Laboratory, Memo. No. UCB/ERL M/80/11, 1980.

[P1] E. POLAK AND D. Q. MAYNE, *An algorithm for optimization problems with functional inequality constraints*, IEEE Trans. Automatic Control, AC-21 (1976), pp. 181-194.

[P2] E. POLAK, R. TRAHAN AND D. Q. MAYNE, *Combined phase-I-phase-II methods of feasible directions*, Math. Programming, 17 (1979), pp. 32-61.

[P3] E. POLAK AND A. SANGIOVANNI VINCENTELLI, *Theoretical and computational aspects of the optimal design centering, tolerancing and tuning problem*, IEEE Trans. Circuits and Systems, CAS-26 (1979), pp. 795-813.

[P4] E. POLAK AND D. Q. MAYNE, *On the solution of singular value inequalities over a continuum of frequencies*, IEEE Trans. Automatic Control, AC-26 (1981), pp. 690-695.

[P5] E. POLAK AND A. TITS, *A recursive quadratic programming algorithm for semi-infinite optimization problems*, Univ. California, Berkeley, Electronics Research Laboratory, Memo No. UCB/ERL M80/50, 1980, J. Applied Math. Optim., in press.

[P6] E. POLAK AND Y. Y. WARDI, *A nondifferentiable optimization algorithm for the design of control systems subject to singular value inequalities over a frequency range*, Automatica, 18 (1982), pp. 267-283.

[P7] E. POLAK, D. Q. MAYNE AND Y. Y. WARDI, *On the extension of constrained optimization algorithms from differentiable to nondifferentiable problems*, Univ. California, Berkeley, Electronics Research Laboratory Memo No. UCB/ERL M81/78, April 14, 1981; this Journal, 21 (1983), pp. 179-203.

[P8] E. POLAK AND Y. Y. WARDI, *A study of minimizing sequences*, Univ. California, Berkeley, Electronics Research Laboratory Memo No. UCB/ERL M82/22, March 22, 1982.

[P9] E. POLAK, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, 1981.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

SEMI-DEFINITE MATRIX CONSTRAINTS IN OPTIMIZATION*

R. FLETCHER†

Abstract. Positive semi-definite matrix constraints arise in a number of optimization problems in which some or all of the elements of a matrix are variables, such as the educational testing and matrix modification problems. The structure of such constraints is developed, including expressions for the normal cone, feasible directions and their application to optimality conditions. A computational framework is given within which these concepts can be exploited and which permits the quantification of second order effects. The matrix of Lagrange multipliers in this formulation is shown to have an important relationship to the characterization of the normal cone. Modifications of various iterative schemes in nonlinear programming are considered in order to develop an effective algorithm for the educational testing problem, and comparative numerical experiments are described. It is shown that a particular choice of the penalty parameter for an l_1 exact penalty function is appropriate for this type of problem. The behaviour of the extreme eigenvalues (or sums thereof) of a matrix is related to the ideas of the paper, and the convexity of such functions is proved, together with expressions for subgradients.

Key words. positive semi-definite matrix, nonsmooth optimization, educational testing, extreme eigenvalues

1. Introduction. My interest in semi-definite matrix constraint problems was kindled some time ago (Fletcher 1981a) by studying the *educational testing problem*; that is given a symmetric positive definite matrix S how much can be subtracted from the diagonal of S and still retain a positive semi-definite matrix. Use of the l_1 norm as a measure gives rise to the problem

$$\begin{aligned} &\text{maximize} && e^T\theta, \qquad \theta \in \mathbb{R}^n \\ &\text{subject to} && S - \operatorname{diag} \theta_i \geqq 0, \\ & && \theta \geqq 0, \end{aligned} \tag{1.1}$$

where $e = (1, 1, \cdots, 1)^T$. The first constraint is the semi-definite matrix constraint, adopting the notation that a positive definite matrix A, defined by

$$z^T A z \geqq 0 \quad \forall z, \tag{1.2}$$

is written as $A \geqq 0$. Likewise $A > 0$ denotes a strictly positive definite matrix. Early attempts to solve this problem were not very successful (some references are given by Fletcher (1981a)) and the same is true for early efforts which I counselled (Jayarajan (1979), Pang (1981)) in which the semi-definite constraint is reduced to an eigenvalue constraint and standard nonlinear programming techniques are used. In my case this was due to a presumption that the eigenvalue constraint would be smooth at the solution, except in rare cases. This has turned out to be incorrect and in fact the large majority (although not all) of such problems are nonsmooth at the solution.

A related problem to (1.1) is the *matrix modification problem* which arises in Newton-like methods for unconstrained optimization. In this case G is a symmetric but indefinite matrix and the question is how little must be added to the diagonal of G to make G positive semi-definite. Again use of the l_1 norm as a measure gives rise to the problem

$$\begin{aligned} &\text{minimize} && e^T\theta \\ &\text{subject to} && G + \operatorname{diag} \theta_i \geqq 0, \\ & && \theta \geqq 0. \end{aligned} \tag{1.3}$$

* Received by the editors March 17, 1983, and in revised form February 16, 1984.
† Department of Mathematical Sciences, University of Dundee, Dundee DD1 4HN, Scotland.

The difference from (1.1) is essentially that the bounds on θ are reversed. Another problem arising in this application would be to look for the least change to all the elements of G (this would involve the ideas of § 2 directly).

In this paper the structure of the semi-definite constraint $A \geqq 0$ is analysed, firstly in regard to variations in all the elements of A. In § 2 various expressions are given for the normal cone of the set, and for feasible directions and optimality conditions. Section 3 describes a more practical computational framework within which these concepts can be exploited, and which allows for the quantification of second order effects. The matrix of Lagrange multipliers in this formulation is shown to be closely related to a matrix which arises in the characterization of the normal cone. In § 4 an important subclass is considered in which variations in only the diagonal elements of A are permitted (corresponding more directly to (1.1) and (1.3)), and the analogous structure is set out. In § 5 various recent iterative schemes for nonlinear programming are considered (applied to the framework of § 3) in an attempt to derive an effective algorithm to solve the educational testing problem (1.1). One particular feature of interest is the use of an l_1 exact penalty function, for which a particular choice of the penalty parameter is appropriate. The numerical results and a comparison with Woodhouse's (1976) method are given in § 6.

The development of §§ 2 and 4 is also related to the behaviour of the extreme eigenvalues (or in general sums of extreme eigenvalues) of A. Applications of optimization involving such functions occur in the graph partitioning problem (Cullum, Donath and Wolfe (1975)) and in control theory problems (for example Mayne and Polak (1982)). The convexity of such functions, and expressions for subgradients similar to those in §§ 2 and 4, are developed in the Appendix to this paper.

This introduction finishes with a few remarks about notation. Conventionally $A=[a_{ij}]$, $B=[b_{ij}]$ etc. is used to refer to individual elements of matrices. The solution to a problem is denoted by a superscript * as in θ^*, A^* etc. Iterates in an iteration scheme are denoted by superscript (k), for example $\theta^{(k)}$, $A^{(k)}$ etc. Quantities computed from these quantities are superscripted correspondingly, for example $D(A^*)$ is denoted by D^* and so on.

2. The positive semi-definite matrix cone. The set of all $n \times n$ symmetric positive semi-definite matrices

$$K=\{A \mid A \in \mathbb{R}^{n\times n},\ A^T=A,\ A \geqq 0\} \tag{2.1}$$

is a closed convex cone of dimension $\frac{1}{2}n(n+1)$. The convexity is an immediate consequence of the definition (1.2) and the dimension is the number of free parameters in a symmetric matrix. For example if $n=2$ and $A=\begin{bmatrix} x & z \\ z & y \end{bmatrix}$, then the cone K is defined by the inequalities $x \geqq 0$, $y \geqq 0$, $xy \geqq z^2$ and is illustrated in Fig. 1. It can be seen that matrices on the boundary of the cone are singular, whereas those in the interior are positive definite.

To derive optimality conditions for problems which involve any closed convex set $K \subset \mathbb{R}^n$, it is convenient to introduce the concept of the *normal cone* ∂K. If x' is on the boundary of K, then $\partial K(x')$ is the set of outward pointing gradients (normals) of all supporting hyperplanes at x' (see Fig. 2). Consequently any vector $g \in \partial K(x')$ satisfies $g^T(x-x') \leqq 0$ for all $x \in K$. Thus the normal cone can be defined by

$$\partial K(x')=\left\{ g \;\middle|\; g^Tx'=\sup_{x\in K} g^Tx \right\}. \tag{2.2}$$

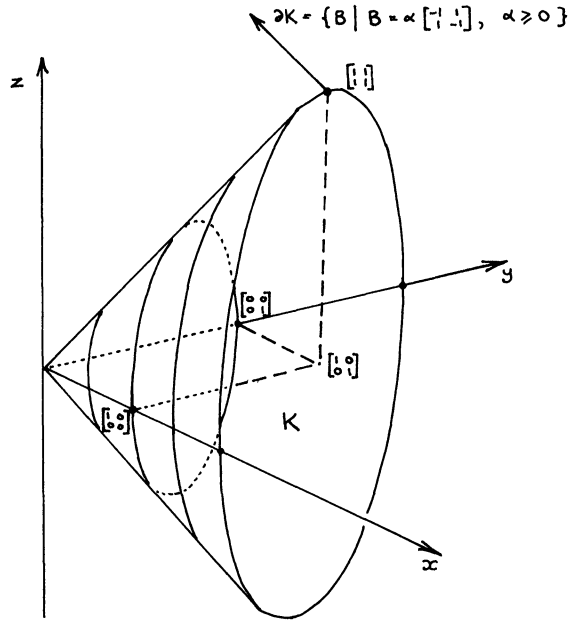


FIG. 1. *The positive semi-definite matrix cone K.*

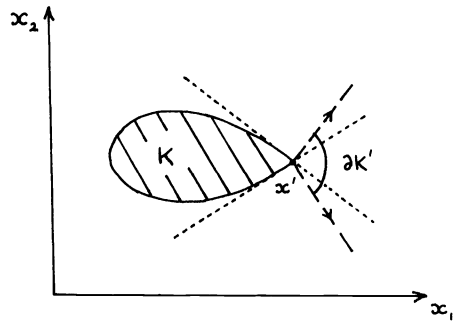


FIG. 2. *The normal cone ∂K to a convex set K.*

If x' is not on the boundary of K, then it is convenient to define either $\partial K(x') = \{0\}$ if x' is interior to K, or $\partial K(x') = \varnothing$ if x' is exterior to K. Both these definitions are consistent with (2.2), the latter by virtue of the separating hyperplane theorem.

In order to analyse problems in which the variables are matrices, an innerproduct defined by

$$A : B = \sum a_{ij} b_{ij} = \operatorname{tr}(A^T B) \tag{2.3}$$

is suitable. It follows from (2.2) that

$$\partial K(A) = \left\{ B \;\middle|\; A : B = \sup_{A \in K} A : B \right\}. \tag{2.4}$$

Observe that if $B \in \partial K$, then $B + S \in \partial K$ where S is any skew-symmetric matrix, since $A : S = 0$. Unsymmetric matrices in ∂K are of no interest (indeed they could be avoided by eliminating the symmetry constraint in (2.1) and working directly in a space of

dimension $\frac{1}{2}n(n+1)$: this requires an inner product equivalent to (2.3) in which off-diagonal terms are doubled). Therefore ∂K subsequently refers to the symmetric normal cone

$$\partial K(A) = \left\{ B \;\middle|\; B^T = B, A:B = \sup_{A \in K} A:B \right\}. \tag{2.5}$$

The case in which K is the positive semi-definite matrix cone (2.1) is now considered, and some equivalent forms of (2.5) are derived which are more useful.

THEOREM 2.1. *If K is defined by* (2.1), *then* (2.5) *is equivalent to*

$$\partial K(A) = \begin{cases} \varnothing & \text{if } A \notin K, \\ \{B \mid B^T = B, A:B = 0, B \leqq 0\} & \text{if } A \in K. \end{cases} \tag{2.6}$$

Proof. If $A \notin K$, then $\partial K(A) = \varnothing$ (see discussion after (2.2)).

Let $A \in K$. First consider $\sup_{A \in K} A:B$ for fixed B. Let $B = X\Omega X^T$ be the spectral decomposition of B with X being the orthogonal matrix of eigenvectors and $\Omega = \operatorname{diag} \omega_i$ the diagonal matrix of eigenvalues. Then using (2.3)

$$\sup_{A \in K} A:B = \sup_{A \in K} C:\Omega = \sup_{C \in K} C:\Omega = \sup_{c_{ii} \geqq 0} \sum c_{ii}\omega_i$$

$$= \begin{cases} 0 & \omega_i \leqq 0 \quad \forall i, \\ \infty & \text{otherwise}, \end{cases}$$

where $C = X^T A X$ is in K iff A is in K. Since the ω_i are the eigenvalues of B, the second part of (2.6) follows directly from (2.5). □

When A is interior to K ($A > 0$), $B = 0$ is the only element of $\partial K(A)$. The most interesting case is when A is on the boundary of K ($A \geqq 0$ and singular). Let M be the diagonal matrix whose elements are the nonzero eigenvalues of A and let the columns of Y be a corresponding orthonormal set of eigenvectors. Then $A = YMY^T$ and the condition $A:B = 0$ becomes $\operatorname{tr}(MY^TBY) = 0$. Because M is diagonal and $M > 0$, the diagonal elements of Y^TBY are zero. Now let the columns of Z be an orthonormal basis for the null space of A, so that $[Y\ Z]$ is an orthogonal matrix. Expressing B as

$$B = [Y\ \ Z] \begin{bmatrix} R & S \\ S^T & T \end{bmatrix} [Y\ \ Z]^T,$$

then $Y^TBY = R$ so it follows that R has zero diagonal elements. Hence from (2.6), $B \leqq 0$ implies that $R = 0$ and $S = 0$, and hence that $T \leqq 0$. Thus in this case an equivalent form of (2.6) is

$$\partial K(A) = \{B \mid B = -Z\Lambda Z^T, \Lambda = \Lambda^T, \Lambda \geqq 0\}. \tag{2.7}$$

An interpretation of the matrix Λ which appears in this result is given in § 3. An illustration of (2.7) is provided by the matrix $\left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$ (see Fig. 1). Then $Z = \left[\begin{smallmatrix} -1 \\ 1 \end{smallmatrix}\right]$ and $\Lambda = [\alpha] \geqq 0$, so that $\partial K = \{\alpha \left[\begin{smallmatrix} -1 & 1 \\ 1 & -1 \end{smallmatrix}\right], \alpha \geqq 0\}$ which is the normal to the cone at this point. Yet another equivalent expression which follows from (2.9) using Carathéodory's theorem is

$$\partial K = \operatorname*{conv}_{x:Ax=0} -xx^T. \tag{2.8}$$

An alternative way of handling the positive semi-definite matrix cone is to write

$$K = \{A \mid A^T = A, \lambda_n(A) \geqq 0\} \tag{2.9}$$

where λ_n refers to the smallest eigenvalue of A. It is shown in the Appendix (Corollary A.2 with $m=1$) that $\lambda_n(A)$ is a concave function, so the constraint $A \in K$ can be handled as a regular inequality. Of course the function $\lambda_n(A)$ may be nonsmooth and the subdifferential of $-\lambda_n(A)$ is required in the optimality conditions. By virtue of Theorem A.4, this can be expressed in various equivalent forms, namely

$$\partial(-\lambda_n(A)) = \{B \mid B^T = B,\ B \geqq 0,\ \operatorname{tr} B = 1,\ B : A = \lambda_n(A)\} \tag{2.10}$$

$$= \{B \mid B = X \Lambda X^T,\ \Lambda \geqq 0,\ \operatorname{tr} \Lambda = 1,\ X \text{ is an orthonormal basis for } Ax = \lambda_n x\} \tag{2.11}$$

$$= \operatorname{conv} xx^T \quad \forall x : Ax = \lambda_n x,\quad x^T x = 1, \tag{2.12}$$

corresponding to (2.6), (2.7) and (2.8) respectively, together with a normalization condition. There is a slight conflict of notation in (2.11): λ_n is the least eigenvalue of A, and is not related to the matrix Λ. The results in the Appendix are expressed in a more general form that can be useful in certain applications. For example Theorem A.4 together with the ideas of § 4 gives a direct expression for the subdifferential in the graph partitioning problem studied by Cullum et al. (1975).

In addition to the normal cone another important set associated with any point x' in a general convex set $K \subset \mathbb{R}^n$ is the set of feasible directions. This can be expressed as

$$\mathcal{F}(x') = \{s \mid \exists \{x^{(k)}\},\ x^{(k)} \to x',\ s^{(k)} \to s,\ \alpha^{(k)} \downarrow 0\} \tag{2.13}$$

where $\alpha^{(k)} \geqq 0$ and $s^{(k)} \in \mathbb{R}^n$ satisfy $\alpha^{(k)} s^{(k)} = x^{(k)} - x'$. Thus a feasible direction is the limiting direction of any feasible directional sequence at x'. This is related to the dual (or polar) cone of $\partial K(x')$, that is the set

$$F(x') = \{s \mid s^T g \leqq 0\ \forall g \in \partial K(x')\}, \tag{2.14}$$

which is the set of feasible directions for the supporting cone of all supporting hyperplanes at x'.

In the case of the positive semi-definite matrix cone (2.1) a feasible direction becomes a symmetric matrix S, but similar definitions to (2.13) and (2.14) hold, involving the inner product (2.3). It follows from (2.14) and (2.3) that if Z is a basis matrix for the null space of A, then

$$F(A) = \{S \mid \Lambda : (Z^T S Z) \geqq 0\ \forall \Lambda \geqq 0\}$$

and hence that

$$F(A) = \{S \mid Z^T S Z \geqq 0\}. \tag{2.15}$$

It is easily shown by taking limits in (2.13) that $\mathcal{F} \subset F$. Conversely by taking a direction $S \in F$, a direction $S \in \mathcal{F}$ can be constructed, which demonstrates that these sets are in fact equivalent. To show this construction, let $X = [Y\ Z]$ as before be the eigenvector matrix for A (that is, $X^T X = I$ and $X^T A X = \left[\begin{smallmatrix} M & 0 \\ 0 & 0 \end{smallmatrix}\right]$ where $M > 0$ is the diagonal matrix of nonzero eigenvalues). In the case that $Z^T S Z > 0$ then the trajectory

$$A_\varepsilon = A + \varepsilon S \tag{2.16}$$

gives rise to

$$X^T A_\varepsilon X = \begin{bmatrix} M + \varepsilon Y^T S Y & \varepsilon Y^T S Z \\ \varepsilon Z^T S Y & \varepsilon Z^T S Z \end{bmatrix}. \tag{2.17}$$

To show that $X^TA_\varepsilon X$ (and equivalently A_ε) is feasible, it is sufficient to show that

$$M+\varepsilon Y^TSY>0 \tag{2.18}$$

and

$$\varepsilon Z^TSZ-\varepsilon^2 Y^TSZ(M+\varepsilon Y^TSY)^{-1}Z^TSY\geqq 0 \tag{2.19}$$

(by virtue of the existence of block Choleski factors). Clearly for ε sufficiently small, (2.18) holds by virtue of $M>0$ and (2.19) by $Z^TSZ>0$. In the case that $Z^TSZ\geqq 0$ and singular, then the trajectory

$$A_\varepsilon = A+\varepsilon S+\alpha\varepsilon^2 I \tag{2.20}$$

is similarly shown to be feasible if $\alpha>\|M^{-1}\|\,\|S\|^2$ is chosen. Thus in both cases a feasible arc of matrices in the direction S is constructed; the existence of a feasible directional sequence in $\mathscr{F}$ follows by taking $\varepsilon=\varepsilon_k$ for any sequence $\varepsilon_k\downarrow 0$.

The value of the expressions developed in this section lies in their application to optimization problems involving the constraint $A\in K$. Expression (2.15) provides a characterization of a feasible direction of search which is readily verified. The expressions for the normal cone ((2.7) in particular is useful) play the part of the subdifferential in the statement of optimality conditions. This can be done in various ways, for instance if the functions $f(A)$ and $c(A)$ ($\mathbb{R}^{n\times n}\to\mathbb{R}$) are convex and nonsmooth, and if a regularity assumption holds, then first order necessary conditions can be stated as follows.

THEOREM 2.2. *If A^* solves the problem*

$$\begin{aligned}&\text{minimize} \quad f(A)\\ &\text{subject to} \quad A\in K, \qquad c(A)\leqq 0,\end{aligned} \tag{2.21}$$

then A^* is feasible and there exist matrices $G^*\in\partial f^*$, $B^*\in\partial K^*$, $C^*\in\partial c^*$ and a multiplier $\pi^*\geqq 0$, $\pi^*c^*=0$ such that

$$G^*+B^*+\pi^*C^*=0. \tag{2.22}$$

Proof. See for example Rockafellar (1981, Chap. 5). □

Problem (2.21) subsumes problems with smooth convex constraints $c_i(x)\leqq 0$ $i=1,2,\cdots,m$ through the transformation $c(x)=\max_i c_i(x)$. The theorem also generalizes to nonconvex problems to some extent, in which case the subdifferential is replaced by the generalized gradient.

Theorem 2.2 is obtained as a consequence of the fact that no feasible directions of strict descent exist at a minimizing point. In some cases it can be shown that all feasible directions are strict ascent directions, in which case the conditions in Theorem 2.2 are sufficient. Often however Theorem 2.2 may be satisfied but there may exist feasible directions along which $f(A)$ has a zero directional derivative. In this case second order information is required in order to confirm or deny the existence of a minimizer and to provide effective algorithms, and this point is considered in the next section.

3. A computational framework. The theory of the previous section is valuable in that it gives a characterization of first order conditions for problems involving the constraint $A\in K$, in particular the result that the matrix Λ in (2.7) is positive semi-definite. However a disadvantage is that it does not take into account second order effects, whereas it may be important to do this in order to obtain a second order rate of convergence in an algorithm. Also nothing is said about how to compute a basis

matrix Z for the null space of A. These omissions are rectified in this section in the context of a partial LDL^T factorization of A. Assume that the rank of A^* is known to be m $(1 \leqq m < n)$. Permuting rows and columns if necessary, and partitioning

$$A = \begin{bmatrix} A_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix}$$

where A_{11} is $m \times m$, then for A sufficiently close to A^*, partial factors

$$A = LDL^T, \qquad L = \begin{bmatrix} L_{11} & \\ L_{21} & I \end{bmatrix}, \qquad D = \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} \tag{3.1}$$

can be calculated where L_{11} is $m \times m$ unit lower triangular, D_1 is $m \times m$ diagonal and $D_1 > 0$, but D_2 has no particular structure other than symmetry. At the solution $D_2^* = 0$. In general

$$D_2 = A_{22} - A_{21} A_{11}^{-1} A_{21}^T \tag{3.2}$$

and this expression enables the constraint $A \in K$ to be written in the form

$$D_2(A) = 0. \tag{3.3}$$

The advantage of this formulation is that expressions for both first and second derivatives of the constraints with respect to the elements of A can be obtained; an example is given in the next section. The condition that A is close to A^* is needed to ensure that $D_1(A) > 0$.

The partial factorization (3.1) also gives rise to a readily available basis matrix Z. Define

$$V = L^{-T} = \begin{bmatrix} V_{11} & V_{12} \\ & I \end{bmatrix}. \tag{3.4}$$

Then

$$Z = \begin{bmatrix} V_{12} \\ I \end{bmatrix} = \begin{bmatrix} -A_{11}^{-1} A_{21}^T \\ I \end{bmatrix}$$

is a basis matrix for the null space of A when $D_2 = 0$ (because $AZ = 0$ and the rank of Z is $n - m$). In this case (3.2) and (3.3) can also be expressed as

$$D_2(A) = Z^T A Z = 0. \tag{3.5}$$

This formulation also sheds light on the matrix Λ that appears in (2.7). Consider for example the problem

$$\begin{aligned} &\text{minimize} \quad f(A) \\ &\text{subject to} \quad Z^T A Z = 0, \qquad c(A) \leqq 0. \end{aligned} \tag{3.6}$$

Introducing a matrix Λ of Lagrange multipliers for the constraint (3.5), the Lagrangian for the problem can be expressed as

$$\mathscr{L}(A, \Lambda, \pi) = f(A) - \Lambda : (Z^T A Z) + \pi c(A). \tag{3.7}$$

Assume that Z is a fixed matrix computed from A^*. Observing that $\Lambda : (Z^T A Z) = A : (Z \Lambda Z^T)$, then the condition $\nabla_A \mathscr{L} = 0$ gives rise to

$$\nabla_A f^* - Z \Lambda^* Z^T + \pi^* \nabla_A c^* = 0. \tag{3.8}$$

These terms can be identified with the corresponding terms in (2.22), and so the matrix Λ that appears in (2.7) can be interpreted as the *Lagrange multiplier matrix for the constraints* $D_2(A)=0$ relative to the basis Z. This can be important in computation because it enables a solution of (3.6) to be verified as being a solution of (2.21); that is to say it confirms that the correct choice of m has been made.

4. The diagonal restriction of K. The applications described in § 1 have the particular feature that the off-diagonal elements of A do not vary, and so it is important to consider the special case in which the positive semi-definite matrix cone K is regarded solely as a function of the diagonal elements of A. The vector $a=(a_{ii})$ $(a\in\mathbb{R}^n)$ is used to denote these elements and $\hat{K}$ is used to denote the restricted set, either as a function of A

$$\hat{K}(A)=\{A\,|\,A\in K,\ a_{ij}=a'_{ij},\ i\neq j\} \tag{4.1}$$

where the $a'_{ij}=a'_{ji}$ are some fixed values, or as a function of a

$$\hat{K}(a)=\{a\,|\,A'+\operatorname{diag} a_i\in K\} \tag{4.2}$$

where A' refers to the matrix $[a'_{ij}]$ with zero diagonal elements. For example problem (1.1) can be expressed as

$$\begin{aligned} &\text{minimize} && e^Ta, && a\in\mathbb{R}^n,\\ &\text{subject to} && a\in\hat{K}(a), && a\leqq a' \end{aligned} \tag{4.3}$$

in this notation, where $S=A'+\operatorname{diag} a'_i$ and $\theta=a'-a$.

Each of $\hat{K}(A)$ and $\hat{K}(a)$ is a closed convex set (but no longer usually a cone) so it is important to derive an expression for the normal cone, in particular $\partial\hat{K}(a)$. In fact it can be shown that this set is obtained merely by taking the diagonal elements of the matrices $B\in\partial K$, that is

$$\partial\hat{K}(a)=\{b\,|\,b=(b_{ii}),\ B\in\partial K(A)\}. \tag{4.4}$$

This result is by no means immediate, however, since it may not be true for other types of convex sets. (For example in $\mathbb{R}^2$ if $K=\{x\,|\,x_1\geqq 0, x_2\geqq x_1^2\}$, then $\partial K(0)=\{g\,|\,g_1\leqq 0, g_2\leqq 0\}$, whereas if K is restricted to variations in x_1 with $x_2=0$, then $\hat{K}(x_1)=\{0\}$ and $\partial\hat{K}=\mathbb{R}$. Thus any element $g_1>0$ is an element of $\partial\hat{K}$ but there is no $g\in\partial K$ with $g_1>0$. However the converse that $g\in\partial K\Rightarrow g_1\in\partial\hat{K}$ is always true as shown below.)

THEOREM 4.1. *$\partial\hat{K}(a)$ is given by* (4.4).

Proof. If $A'+\operatorname{diag} a_i\notin K$, then both ∂K and $\partial\hat{K}$ are empty, so (4.4) is trivial. Otherwise $A'+\operatorname{diag} a_i\in K$, and let $B\in\partial K$. Then

$$\begin{aligned} \mathrm{B}:(\mathrm{A}'+\operatorname{diag} a_i) &\geqq B:A && \forall A\in K,\\ &\geqq B:(A'+\operatorname{diag} a_i) && \forall a\in\hat{K}(a), \end{aligned}$$

since $\hat{K}(A)\subset K(A)$ and hence denoting $b=(b_{ii})$,

$$b^Ta=\sup_{a\in\hat{K}(a)} b^Ta$$

so that $b\in\partial\hat{K}(a)$.

Conversely let $b\in\partial\hat{K}(a)$. The hypothesis that there is no $B\in\partial K$ such that $(b_{ii})=b$ can be contradicted as follows. The vectors (b_{ii}), $B\in\partial K$ form a closed convex cone so by the separating hyperplane theorem there exists a vector s such that $s^Tb>0$ and $s^T(b_{ii})\leqq 0\forall(b_{ii})$, $B\in K$. The latter condition implies (using (2.14) applied to ∂K) that

the matrix $S = \operatorname{diag} s_i$ satisfies $S \in F(A' + \operatorname{diag} a_i)$. Using construction (2.20), a feasible arc A_ε can be constructed and in fact $A_\varepsilon \in \hat{K}(A)$. Hence $s \in \hat{F}(a)$ and so $s^T b \leqq 0$ which contradicts $s^T b > 0$. □

Various equivalent forms of $\partial \hat{K}(a)$ can be deduced using (2.6), (2.7) and (2.8). The most useful of these is probably that deriving from (2.7),

$$\partial \hat{K}(a) = \{b \mid b = (b_{ii}),\ B = -Z\Lambda Z^T,\ \Lambda = \Lambda^T,\ \Lambda \geqq 0\} \tag{4.5}$$

that is the set of vectors that are the diagonal elements of all matrices of the form $-Z\Lambda Z^T$ where Λ is any symmetric positive semi-definite matrix and Z is the null space matrix. Likewise feasible directions for the set $\hat{K}(a)$ are given by

$$\hat{\mathcal{F}}(a) = \hat{F}(a) = \{s \mid Z^T[\operatorname{diag} s_i]Z \geqq 0\} \tag{4.6}$$

by virtue of (2.15) and the construction of feasible arcs which follows. Optimality conditions also follow a similar pattern to (2.21). For example *first order necessary conditions* for a^* to solve (4.3) are that a^* is feasible and there exist vectors $b^* \in \partial \hat{K}^*$ and $\pi^* \geqq 0$ ($\pi^* \in \mathbb{R}^n$) such that

$$e + b^* + \pi^* = 0, \tag{4.7}$$

$$\pi^{*T}(a'^* - a^*) = 0. \tag{4.8}$$

As an example of these conditions consider problem (4.3) in which

$$A' = \begin{bmatrix} 0 & 2 & 3 \\ 2 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix}, \qquad a' = \begin{pmatrix} 3\frac{1}{2} \\ 6 \\ 8 \end{pmatrix}. \tag{4.9}$$

The solution is $a^* = (2, 2, 4\frac{1}{2})^T$, no bounds are active ($\pi^* = 0$), and the set

$$\hat{K}(a) = \left\{ a \;\middle|\; \begin{bmatrix} a_1 & 2 & 3 \\ 2 & a_2 & 3 \\ 3 & 3 & a_3 \end{bmatrix} \geqq 0 \right\} \tag{4.10}$$

is illustrated in the vicinity of a^* in Fig. 3. It can be observed that $\hat{K}$ is convex but not a cone, and is nonsmooth at a^*. The rank of $A^* = A' + \operatorname{diag} a_i^*$ is $m = 1$, and its partial factors are

$$D = \begin{bmatrix} 2 & & \\ & 0 & \\ & & 0 \end{bmatrix}, \quad L = \begin{bmatrix} 1 & & \\ 1 & 1 & \\ 1\frac{1}{2} & & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & -1 & -1\frac{1}{2} \\ \vdots & 1 & \\ \vdots & & 1 \\ \vdots & & \end{bmatrix}. \tag{4.11}$$

$$\underbrace{\qquad\qquad}_{Z}$$

The vector $b^* = -e$ satisfies (4.7) and the corresponding $B^* \in \partial K$ is generated by the matrix

$$\Lambda^* = \begin{bmatrix} 1 & -\frac{3}{4} \\ -\frac{3}{4} & 1 \end{bmatrix}$$

($\Lambda^* > 0$ as required) and is

$$B^* = -Z\Lambda^* Z^T = -\begin{bmatrix} 1 & \frac{1}{8} & -\frac{3}{4} \\ \frac{1}{8} & 1 & -\frac{3}{4} \\ -\frac{3}{4} & -\frac{3}{4} & 1 \end{bmatrix}. \tag{4.12}$$

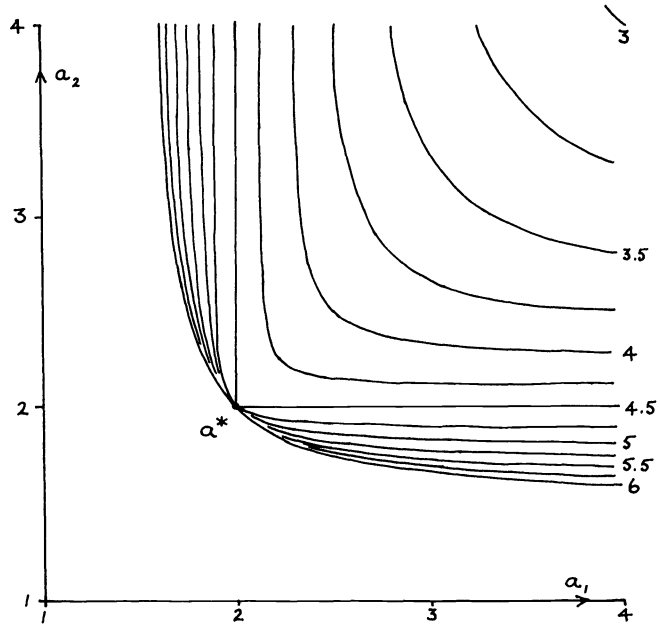


FIG. 3. *The boundary of the restricted cone* $\hat{K}(a)$ *in* (4.10) (*contours of* a_3).

Another example for which $n=4$ is

$$A' = \begin{bmatrix} 0 & 1 & 2 & -2 \\ 1 & 0 & 3 & 2 \\ 2 & 3 & 0 & 1 \\ -2 & 2 & 1 & 0 \end{bmatrix}, \qquad a' = \begin{pmatrix} 2 \\ 4 \\ 8 \\ 10 \end{pmatrix}. \tag{4.13}$$

It is readily verified that the solution is $a^* = (2, 2\frac{1}{2}, 4, 6\frac{1}{2})^T$ and in fact the problem reduces to (4.9) after the first stage of the factorization. The bound $a_1 \leqq a_1'$ is active and has a Lagrange multiplier $\pi_1^* = \frac{29}{8}$. However if a_1' is increased (to $a_1' = 6$ say) in (4.13), then the bound $a_1 \leqq a_1'$ is no longer active, and the vector $a = (2, 2\frac{1}{2}, 4, 6\frac{1}{2})^T$ is feasible but not optimal (note $e^T a = 15$). The same factorization is obtained as before, so that $D_2 = 0$ (feasibility), but the conditions (4.7) and (4.8) do not hold. In fact the solution to this modified problem is

$$a^* = (3.455534, 3.183300, 3.183300, 3.455534)^T \tag{4.14}$$

to six decimal places (note $e^T a^* = 13.277668 < 15$) and $m = 2$. In this modified problem second order effects become important in locating the solution and there exist feasible directions of zero slope at a^*.

There are also some further observations about the computational framework in § 3 which are relevant when considering problems involving the restricted cone $\hat{K}(a)$. In this case there are at most n free variables, and this can be reduced to $n-p$ if there are p active bounds at the solution. The equation $D_2 = 0$ imposes $\frac{1}{2}(n-m+1)(n-m)$ conditions (because $D_2 \in \mathbb{R}^{(n-m)\times(n-m)}$ and using symmetry) so except in degenerate cases it follows that

$$n - p \geqq \tfrac{1}{2}(n-m+1)(n-m) \tag{4.15}$$

which imposes a substantial restriction on the dimensions of D_2. For example if $n-p$ is 20, then $n-m$ can be no larger than 5. Whenever there is no slack in (4.15), then

solving the system $D_2 = 0$ completely determines the solution of the problem. This occurs in problem (4.9) with $(n, m, p) = (3, 2, 0)$ and in problem (4.13) with $(n, m, p) = (4, 2, 1)$. Otherwise there are more free variables than equations and second order effects are important in locating the solution. This occurs in the modified form of problem (4.13) with $(n, m, p) = (4, 2, 0)$. Another important feature in (4.3) is that if there are active bounds on the variables a_i $i > m$, then these would interfere with the formulation of § 3. It is most convenient computationally to permute the variables so that the active bounds are those on variables a_i $i = 1, 2, \cdots, p$.

5. An algorithm. In this section an algorithm is described for solving problem (4.3) which arises from the educational testing problem (1.1). The main idea is to replace the constraint $a \in \hat{K}(a)$ by the set of nonlinear equations (3.3) in order to utilize current nonlinear programming methods which are globally convergent at a second order rate. The ready availability of second derivatives of (3.3) enables the sequential quadratic programming (SQP) method (e.g. Fletcher (1981b)) to be used. This provides locally second order convergence and is an improvement over previous algorithms (Woodhouse (1976), Jayarajan (1979), Pang (1981)) which appear to converge linearly. The convergence characteristics of the SQP method are improved by the incorporation of an l_1 exact penalty function and the use of second order corrections (Fletcher (1982a)). However there are also some difficulties that arise when replacing the constraint $a \in \hat{K}(a)$ by (3.3) which require special attention. One of these is that the index m (= rank (A^*)) used in partitioning A is not known in advance. However it is shown that by solving a sequence of problems for different m, each of which is well-behaved, the correct value of m can be located. Some difficulties arising from the bounds $a \leqq a'$ are also discussed; these are handled by a permutation technique which is successful in practice, although it does not allow the standard proof of global convergence to be used.

In developing an algorithm for solving (4.3), the ideas of § 3 are followed and it is temporarily assumed that $m = \text{rank}\,(A^*)$ is known $(A^* = A' + \text{diag}\, a_i^*)$, and also that the variables have been permuted so that the bounds $a_i \leqq a_i'$ are inactive for $i > m$. Then (4.3) can be expressed as

$$\begin{aligned} &\text{minimize} \quad e^T a, \qquad a \in \mathbb{R}^n \\ &\text{subject to} \quad D_2(a) = 0, \qquad a \leqq a' \end{aligned} \tag{5.1}$$

where $D_2(a)$ refers to (3.2) in which A is the matrix $A' + \text{diag}\, a_i$. The Lagrangian for this problem is

$$\mathscr{L}(a, \Lambda, \pi) = e^T a - \Lambda : D_2(a) + \pi^T (a - a') \tag{5.2}$$

and the first order conditions are given by (4.7) and (4.8). By virtue of (4.4), b^* is a vector whose elements are the diagonal elements of the matrix $-Z^T \Lambda^* Z$, where Λ^* $(= [\lambda_{ij}^*])$ is the matrix of Lagrange multipliers for the constraints $D_2(a) = 0$, and Z is the null space matrix for A^*. Setting derivatives of (5.2) to zero and using (3.2) gives

$$\frac{\partial \mathscr{L}}{\partial a_i} = 1 - \lambda_{ii} + \pi_i = 0, \qquad i = m+1, \cdots, n. \tag{5.3}$$

Since the bounds are inactive for $i > m$, π_i is zero and

$$\lambda_{ii} = 1, \qquad i = m+1, \cdots, n, \tag{5.4}$$

follows. (For convenience the elements λ_{ij} of Λ are indexed from $m+1, \cdots, n$ to correspond to the elements d_{ij} of D_2.)

The simple form of (3.2) can be exploited to use the diagonal elements of $D_2(a)$

$$d_{ii}(a) = a_i - \sum_{kl} a_{ik}[A_{11}^{-1}]_{kl}a_{il} = 0 \tag{5.5}$$

to eliminate the variables a_i. (Conventionally i and j will refer to indices whose scope is from $m+1$ to n whilst k, l, r and s refer to indices whose scope is from 1 to m.) The variables a_k are the unknowns, and will subsequently be referred to by a ($a \in \mathbb{R}^m$). Thus (5.1) reduces to

$$\begin{aligned} &\text{minimize} \quad f(a) \triangleq \sum a_k + \sum a_i(a), \\ &\text{subject to} \quad d_{ij}(a) = 0, \qquad i \neq j, \quad a \leqq a', \end{aligned} \tag{5.6}$$

where $a_i(a)$ indicates that a_i is the function of a determined by (5.5). It is easily seen that the Lagrangian function for (5.6) is again given by (5.2) if the result that $\lambda_{ii} = 1$ is used. In fact (5.6) is unnecessarily redundant in that the equivalent constraints $d_{ij}(a) = 0$ and $d_{ji}(a) = 0$ are both present. In practice the constraints would only be stated for $i > j$ and the Lagrange multiplier for each constraint in this system would be $2\lambda_{ij}$. However it is notationally more convenient to refer to (5.6).

The application of standard nonlinear programming techniques to solve (5.6) is now considered. In order to write down the SQP method it is necessary to derive expressions for ∇d_{ij} and $\nabla^2 d_{ij}$ where $\nabla = (\partial/\partial a_1, \cdots, \partial/\partial a_m)^T$. Now $\partial A_{11}/\partial a_r = e_r e_r^T$ and so it follows by differentiating in $A_{11}A_{11}^{-1} = I$ that

$$\frac{\partial A_{11}^{-1}}{\partial a_r} = -A_{11}^{-1} e_r e_r^T A_{11}^{-1}. \tag{5.7}$$

Hence from (3.2)

$$\frac{\partial D_2}{\partial a_r} = A_{21}A_{11}^{-1} e_r e_r^T A_{11}^{-1} A_{21}^T = V_{12}^T e_r e_r^T V_{12} \tag{5.8}$$

or

$$\frac{\partial d_{ij}}{\partial a_r} = v_{ri}v_{rj}. \tag{5.9}$$

Differentiating again in (5.7) gives

$$\frac{\partial^2 A_{11}^{-1}}{\partial a_r\, \partial a_s} = A_{11}^{-1}(e_s e_s^T A_{11}^{-1} e_r e_r^T + e_r e_r^T A_{11}^{-1} e_s e_s^T) A_{11}^{-1} \tag{5.10}$$

and hence

$$\frac{\partial^2 d_{ij}}{\partial a_r\, \partial a_s} = -(v_{ri}v_{sj} + v_{si}v_{rj})[A_{11}^{-1}]_{rs}. \tag{5.11}$$

Each iteration of the SQP method applied to (5.6) requires the solution of the QP subproblem

$$\begin{aligned} &\text{minimize} \quad f^{(k)} + \nabla f^{(k)T}\delta + \tfrac{1}{2}\delta^T \nabla^2 \mathscr{L}^{(k)}\delta, \qquad \delta \in \mathbb{R}^m \\ &\text{subject to} \quad d_{ij}^{(k)} + \nabla d_{ij}^{(k)T}\delta = 0, \qquad i \neq j, \\ &\qquad\qquad a^{(k)} + \delta \leqq a' \end{aligned} \tag{5.12}$$

giving a correction vector $\delta^{(k)}$, so that $a^{(k+1)} = a^{(k)} + \delta^{(k)}$. Also the Lagrange multipliers of the equations in (5.12) become the elements $\lambda_{ij}^{(k+1)}$ for the next iteration. Formulae

for the derivatives required in (5.12) are obtained from (5.9) and (5.11). ∇d_{ij} is given directly from (5.9), and from (5.6) and (5.5)

$$\nabla f = e - \sum_i \nabla d_{ii}. \tag{5.13}$$

From (5.2)

$$\nabla^2 \mathscr{L}(a^{(k)}, \Lambda^{(k)}, \pi) = \sum_{ij} - \lambda_{ij}^{(k)} \nabla^2 d_{ij}(a^{(k)}) \tag{5.14}$$

(note that the diagonal terms are included with $\lambda_{ii}^{(k)} = 1$ from (5.4)). This can be rearranged using (5.11) as the matrix $\nabla^2 \mathscr{L}^{(k)}$ with elements

$$\begin{aligned} [\nabla^2 \mathscr{L}^{(k)}]_{rs} &= 2[V_{12} \Lambda^{(k)} V_{12}^T]_{rs} [A_{11}^{-1}]_{rs} \\ &= 2[V_{12} \Lambda^{(k)} V_{12}^T]_{rs} [V_{11} D_1^{-1} V_{11}^T]_{rs}. \end{aligned} \tag{5.15}$$

The submatrices of V and D on the right-hand side of (5.15) are calculated using the partial factors of the matrix $A^{(k)}$ $(= A' + \text{diag } a_i^{(k)})$ in accordance with § 3. Since $A^* \geqq 0$ from (4.5), it can be observed that the matrix $\nabla^2 \mathscr{L}^*$ $(= \nabla^2 \mathscr{L}(a^*, \Lambda^*, \pi^*))$ is positive semi-definite. This follows from (5.15) because

$$\begin{aligned} z^T \nabla^2 \mathscr{L}^* z &= 2 \text{ tr } (V_{12} \Lambda^* V_{12}^T [\text{diag } z_r] A_{11}^{-1} [\text{diag } z_s]) \\ &= 2 \text{ tr } (D_1^{-1/2} V_{11}^T [\text{diag } z_s] V_{12} \Lambda^* V_{12}^T [\text{diag } z_r] V_{11} D_1^{-1/2}) \\ &\geqq 0 \end{aligned} \tag{5.16}$$

since the matrix inside the trace is symmetric positive semi-definite. Usually $\nabla^2 \mathscr{L}^*$ is positive definite in which case, if $a^{(k)}$ is sufficiently close to a^*, the basic SQP method converges and the rate is second order (e.g. Fletcher (1981b)).

Globally however the SQP method may not converge and it is usual to modify the basic method in some way by introducing an l_1 exact penalty function. In this case the l_1 exact penalty function for (5.6) is

$$\phi(a) = \sum a_k + \sum a_i + \sigma \left\{ \sum_{i \neq j} |d_{ij}| + \sum_i \max (a_i - a_i', 0) \right\} \tag{5.17}$$

where the quantities d_{ij} and a_i are functions of a $(a \in \mathbb{R}^m)$ as above. The assumption that the bounds are inactive at the solution implies that the max terms are zero if $a^{(k)}$ is sufficiently close to a^*. To ensure that a minimizer of (5.17) satisfies first order conditions for (5.6), the penalty parameter σ in (5.17) must satisfy $\sigma \geqq \max_{ij} |\lambda_{ij}^*|$ (e.g. Fletcher (1981b)). But $\Lambda^* \geqq 0$ and $\lambda_{ii}^* = 1$ imply that $\max_{ij} |\lambda_{ij}^*| \leqq 1$, and equality can be obtained. Thus $\sigma \geqq 1$ must hold. Since it is advantageous to choose σ as close to this threshold as possible, the choice $\sigma = 1$ is recommended. (In practice if the non-redundant form of (5.6) is used with summation over indices $i > j$, then a similar summation is used in (5.17) and the choice $\sigma = 2$ is appropriate.)

Various methods of integrating (5.12) and (5.17) exist and have been tried here. The simplest, that of Han (1977), uses the solution of (5.12) as a search direction, and the next point is accepted only if it significantly reduces the value of $\phi(a)$. Unfortunately large values of σ are required to impose a descent property on the method and no successful algorithm of this type has been obtained. Another algorithm closely related to the SQP method, but having stronger convergence properties than the Han method, is suggested by Fletcher (1981c) (Algorithm 1). It uses the same approximating functions

as (5.12) but solves a subproblem of the form

$$\begin{aligned} &\text{minimize} \quad f^{(k)}+\nabla f^{(k)T}\delta+\tfrac{1}{2}\delta^T\nabla^2\mathscr{L}^{(k)}\delta+\sigma\sum_{i\neq j}|d_{ij}^{(k)}+\nabla d_{ij}^{(k)T}\delta| \\ (5.18)\qquad &\text{subject to} \quad a^{(k)}+\delta\leqq a', \\ &\qquad\qquad \|\delta\|_\infty\leqq\rho^{(k)} \end{aligned}$$

which is more directly related to (5.17). The parameter $\rho^{(k)}$ is the radius of a trust region constraint and is varied in accordance with standard practice to achieve a significant decrease in $\phi(a)$. The subproblem can be solved by techniques analogous to those used in QP, and asymptotically the SQP method (5.12) and the method based on (5.18) are equivalent. Method (5.18) solves many of the test problems described in § 6; in some cases, however, slow convergence is obtained, apparently caused by following curved grooves in the graph of $\phi(a)$ caused by the nonsmooth terms $|d_{ij}(a)|$. This type of behaviour has become better understood recently, and a remedy is to modify the basic algorithm to allow a "second order correction step" in certain circumstances (e.g. Fletcher (1982a)). With this modification, the resulting algorithm succeeded in solving all the test problems. In principle however, any effective algorithm for nonlinear programming can be used to solve (5.6); what is proposed here is in line with current (albeit recent) methodology, and so is not described in too much detail. More is given in the report (Fletcher (1982b)) which precedes this paper.

The above description of the algorithm does not, however, take account of certain features of the semi-definite matrix constraint which require attention. The most important consideration is how the integer $m^*=\text{rank}\,(A^*)$ can be identified correctly. Let the current estimate of m^* be denoted by $m^{(k)}$. Any change to $m^{(k)}$ causes an unpredictable change to $\phi(a^{(k)})$, conflicting with the global convergence strategy of reducing $\phi(a^{(k)})$ monotonically, so it is unwise to change $m^{(k)}$ frequently. Consider therefore the effect of making a *fixed* but incorrect estimate m of m^*. If $m<m^*$, then the second order correction method converges satisfactorily and ultimately at a second order rate to the minimizer of $\phi(a)$. However, because m is too small, there are too many conditions $d_{ij}(a)=0$ in (5.6) and so it happens that $d_{ij}\neq 0$ for some indices $i\neq j$ at the minimizer of $\phi(a)$. Also $\Lambda\geqq 0$ does not usually hold. Therefore the minimizer of $\phi(a)$ is not a solution of (5.6). If $m>m^*$, then the second order correction method may converge to the minimizer of $\phi(a)$, which is the solution of (5.6) in this case, but the rate is so slow as to be unacceptable in practice. In the case $m=n-1>m^*$ it is similar to the algorithm described by Jayarajan (1979) and Fletcher (1981a). This very slow rate of convergence indicates that the nonsmooth nature of the problem has not been accounted for. These observations suggest the following approach in which m approaches m^* from below. Initially m is chosen as the smallest integer, $\bar{m}$ say, compatible with (4.15). Then $\phi(a)$ is minimized using the second order correction method, starting from $a^{(1)}=a'$, $\Lambda^{(1)}=0$ and $\rho^{(1)}=\bar{\rho}$ which is user supplied. If $D_2(a)=0$ at the minimizer, then the solution of (5.6) has been determined and the process terminates. Otherwise if $D_2(a)\neq 0$, then m is increased by one and the process is repeated. Thus a sequence of nonlinear programming problems, each of which is well-behaved, is solved until the correct value m^* is identified. For the values of n described here ($n\leqq 20$) only a few values of m need be tried (see Table 6.3) and the process has proved to be very reliable and reasonably efficient.

Another feature that needs attention is the following. A requirement of the approach based on (3.1) ff. is that the variables $a\in\mathbb{R}^m$ must permit the matrix $A'+\text{diag}\,a_i$ to be factorized with $D_1>0$, and this restricts the choice of variables $a^{(k)}$ in the

nonlinear programming method. The simplest way to deal with this when using (5.18) is to decrease the trust region radius (choosing $\rho^{(k+1)} = \rho^{(k)}/4$, $a^{(k+1)} = a^{(k)}$, $\Lambda^{(k+1)} = \Lambda^{(k)}$) whenever the solution of the subproblem gives rise to a matrix whose partial factors (3.1) do not exist. To discourage the algorithm from choosing such matrices, some additional linear constraints have also been added to the subproblem—more details are given by Fletcher (1982b). The main feature of these heuristics is that they only come into play when $a^{(k)}$ is remote from the minimizer of $\phi(a)$ and so do not affect the second order rate of convergence. I think it is also likely that the global convergence property of the second order correction method will remain valid.

Finally, another restriction on the variables $a \in \mathbb{R}^m$ of (5.6) is that the bounds $a_i \leqq a_i'$ $i > m$ must remain inactive. In practice this can only be achieved by permuting the variables, and a suitable permutation is not known in advance, so the following heuristic has been adopted. At the start of an iteration each variable is examined in turn: if it is active and is preceded by inactive variables then these variables are permuted cyclically so that the active variable is first. Consequently the active bounds are those on variables $a_r^{(k)}$ $r = 1, 2, \cdots, p$. However any such permutation causes a complete change to the factorization (3.2). The matrix D_2 and the basis matrix Z are redefined, and the Lagrange multiplier estimates are reset to zero since they are no longer appropriate to the new basis. Likewise the function $\phi(a)$ in (5.17) is redefined, which causes an arbitrary change in the value of $\phi(a^{(k)})$. This conflicts with the global convergence strategy of reducing $\phi(a^{(k)})$ monotonically. Therefore any convergence proof is valid only if the number of permutations made during the course of the algorithm is finite. Whilst practical experience suggests that permutations occur infrequently, and there is no evidence of zig-zagging, it is an open question as to whether a counter-example could be constructed.

6. Numerical experiments. The algorithm of § 5 is applied to solve the set of educational testing problems (1.1) given by Woodhouse (1976). The results show that the method is effective and reliable, and give an indication of how much computational effort is required. The computations have been carried out on a DEC 10 computer with a single length precision of 7–8 decimal digits.

In the educational testing problem, the $n \times n$ matrix S is constructed from an $N \times n$ data matrix X $(N > n)$ in the following way. The column means $\bar{x}_j = \sum_i x_{ij}/N$ are calculated and then

$$s_{jk} = \frac{1}{N-1} \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{6.1}$$

determines S. Since this computation essentially "squares" the matrix X, which can cause substantial loss of precision, both (6.1) and the subsequent factorization of $S - \text{diag}\ \theta_i$ are carried out in double precision. An alternative to this which has not been followed up would be to use some form of Square Root Free Givens operations to compute the factors of $S - \text{diag}\ \theta_i$ directly from X and θ in single precision. Double precision is also used for accumulation of scalar products with single precision for all other operations. The particular set of problems given by Woodhouse (1976) is derived from a 64×20 data set which is reproduced by Fletcher (1981a). Unfortunately due to a poor photocopy, there are some errors in the table given by Fletcher. For this table to correspond exactly to the Woodhouse table the following corrections are necessary, the erroneous figures being given in brackets: $x_{30,1} = 59$ (39), $x_{14,13} = 31$ (21), $x_{22,13} = 36$ (56), $x_{17,18} = 37$ (27), $x_{4,20} = 28$ (38), $x_{55,20} = 76$ (75). The corrected figures check with the S matrix given by Woodhouse: it is sufficient to check the diagonal elements which are given in Table 6.1.

TABLE 6.1
Diagonal elements of the matrix S for the Woodhouse 1-20 problem.

i	s_{ii}	i	s_{ii}	i	s_{ii}	i	s_{ii}
1	407.5394	6	51.71032	11	199.7688	16	325.8311
2	358.5970	7	329.8482	12	317.3013	17	512.7696
3	329.5784	8	275.5652	13	277.6419	18	414.6704
4	317.4363	9	300.9117	14	160.7054	19	479.7054
5	145.1069	10	232.2299	15	104.3175	20	459.2835

Woodhouse uses the 64×20 data set to generate various test problems by selecting various subsets of columns to form the matrix X. These subsets are those given in the first column of Table 6.2, the number of elements in each subset being the value of n. The results given by Woodhouse are reproduced in Table 6.2 for comparison purposes. Woodhouse only gives his results to one decimal place, and his solutions do not always correspond to feasible points, although nearly so. For an accurate comparison, one (two for the 1-18 case) of the variables is adjusted so that the matrix $S - \text{diag}\ \theta_i$ is exactly singular and positive semi-definite. The adjustment required is small: 1.6 units

TABLE 6.2
The results given by Woodhouse (1976).

Columns which determine X	θ_i^* $i = 1, 2, \cdots, n$				$\sum \theta_i^*$
1, 2, 5, 6	174.8	235.8	103.4	28.76130	542.76130
1, 3, 4, 5	155.9	240.9	128.8	107.3818	632.98179
1, 2, 3, 6, 8, 10	0	101.7	20.2	31.5	
	82.3	69.78108			305.48108
1, 2, 4, 5, 6, 8	86.0	119.6	58.4	108.6	
	45.6	30.34421			448.54421
1-6	116.0	21.2	104.1	131.5	
	113.2	41.86110			527.86111
1-8	32.0	54.7	79.1	147.9	
	98.9	29.8	105.2248	80.7	628.32481
1-10	8.6	73.1	57.9	123.7	
	122.1	28.7	87.1	60.5	
	37.41804	60.8			659.91805
1-12	18.3	72.2	59.6	125.2	
	98.9	23.3	95.9	40.4	741.51527
	36.5	48.8	64.5	57.91526	
1-14	11.7	40.9	95.1	138.2	
	99.1	22.5	34.5	27.3	596.78513
	46.4	29.4	22.9	24.5	
	2.3	1.985111			
1-16	6.2	47.4	84.5	127.2	
	107.7	28.6	36.9	46.4	
	31.3	4.1	5.9	16.5	563.09605
	3.2	9.6	0.1960485	7.4	
1-18	4.4	59.4	72.4	124.7	607.75524
	93.2	27.4	26.4	41.9	
	31.5	44.1	10.4	6.7	
	0	1.2	0.05523563	4.0	
	25.9	34.1			
1-20	no solution obtained				

TABLE 6.3
Results for method of § 5.

Columns which determine X	$\bar{m}$	m^*	NQP	θ_i^* $i=1,2,\cdots,n$				$\sum \theta_i^*$
1, 2, 5, 6	2	3	14	173.1174	236.8681	103.8765	28.91159	542.77356
1, 3, 4, 5	2	2	12	156.2324	240.9354	128.7423	107.2478	633.15784
1, 2, 3, 6, 8, 10	3	5	9	0	102.0205	19.87713	31.46058	305.48170
				82.28315	69.84034			
1, 2, 4, 5, 6, 8	3	4	13	59.63703	214.0408	69.79544	115.7275	564.46331
				47.03951	58.22299			
1-6	3	4	14	152.7057	54.47563	82.93145	99.64148	535.36227
				104.6550	40.95294			
1-8	5	6	29	14.03226	38.54178	95.09897	158.9009	641.83848
				120.3823	28.37133	106.7753	79.73562	
1-10	6	8	34	0	43.89229	80.71647	132.8874	690.78040
				126.8620	28.03018	56.62000	92.61001	
				61.33628	67.82578			
1-12	8	9	29	18.63315	61.86325	63.42740	127.5681	747.48921
				99.97348	30.77045	96.53485	45.28755	
				41.60155	45.32906	64.04083	52.45959	
1-14	10	12	36	0	59.49895	62.91230	109.9237	671.27506
				99.94911	32.71942	79.07282	31.73815	
				47.42095	33.78883	41.95283	63.59558	
				4.251664	4.450777			
1-16	11	14	42	0	63.48564	52.38936	108.1923	663.46204
				92.39511	34.56210	85.75734	21.95673	
				37.54895	32.96661	28.51149	54.57159	
				12.92956	4.102086	6.706302	27.38691	
1-18	13	15	27	0	58.38015	62.16198	107.2306	747.50574
				80.28730	25.38330	70.70336	24.31729	
				52.43795	41.69481	24.29243	39.17598	
				15.76098	6.861478	3.259009	14.59315	
				68.80438	52.16162			
1-20	15	18	39	0	47.37281	76.58167	101.0016	820.34265
				63.44995	13.38219	41.48301	4.300287	
				56.36490	33.98319	33.76988	29.95976	
				17.59711	0	4.328106	13.69029	
				45.58721	51.58627	57.20664	128.6977	

for the {1-10} case, 0.8 units in the {1, 2, 4, 5, 6, 8} case, 0.6 units in the {1-14} case, otherwise at most ~0.2 units. The results obtained by the new method of § 5 are tabulated (in unpermuted form) in Table 6.3. The initial value of $\bar{\rho}$ is 20. By solving some of the problems with a different initial $\bar{\rho}$ it is estimated that these results are accurate to 5-6 decimal places in the variables θ_i and 7-8 decimal places in the function value $\sum \theta_i$. The maximum residual $\|d_{ij}\|_1$ on any problem for the matrix $D_2(a)$ is $0.2_{10^{-5}}$, which is as good as can be expected in single length with elements $s_{ij} \sim 100$. By comparing the values in Tables 6.2 and 6.3 it can be seen that, except for the smallest problems, the Woodhouse method only gives an order of magnitude estimate of the correct variables, and there are substantial differences in the best function values that are achieved. In Table 6.3 the column headed NQP gives the number of times that the major l_1QP problem (5.18) is solved. For each of these, a parametric solution of the l_1QP problem is also carried out to calculate the second order correction and a partial factorization of a matrix $S - \text{diag}\ \theta_i$ is attempted. This amount of computation is well within the capacity of modern computers, and is quite acceptable for obtaining accurate

solutions to the educational testing problem. However the amount involved is not small and would not be acceptable in practice for solving the matrix modification problem (1.3) which must be solved on every iteration of a modified Newton method. Therefore it is of some interest to enquire whether yet more effective methods can be developed.

In addition Table 6.3 gives the initial value $\bar{m}$ of m and the correct value m^* for each particular problem. It can be seen that only in two cases (1-16 and 1-20) is the minimization of $\phi(a)$ repeated for four different values of m, otherwise fewer repeats are required. In most cases it is observed that fewer iterations are required to minimize $\phi(a)$ as m increases. For each value of m, second order convergence of the method is observed as predicted by the theory, and there is no evidence of the Maratos effect, presumably because second order corrections are being used.

The results of Tables 6.2 and 6.3 are given in detail, not only because of their intrinsic interest, but also because they could be used as test problems for nonsmooth optimization. One possibility could be to use the problem

$$\begin{aligned} &\text{maximize} && e^T\theta \\ &\text{subject to} && \lambda_n(S-\operatorname{diag}\theta_i)\geqq 0, \qquad \theta\geqq 0 \end{aligned} \tag{6.2}$$

which has a nonsmooth constraint. Another possibility is to use the eigenvalue constraint in (6.2) to eliminate one of the variables, giving rise to a nonsmooth optimization problem with simple bounds. Yet another possibility would be to create an l_1 exact penalty function

$$\phi(\theta)=-e^T\theta+\sigma(\min(\lambda_n(S-\operatorname{diag}\theta_i),0)+\textstyle\sum\min(\theta_i,0)) \tag{6.3}$$

for sufficiently large σ.

Finally it is worth remarking that the new method of § 5 was initially tested out on some random matrix problems in which the elements of the matrix S in (1.1) is generated by using random integers in $[-10, 10]$. To make S positive definite, a matrix cI (c an integer) is then added in. The method had no difficulty in handling such problems, and various problems of dimension $n\leqq 12$ were used. One point which is worth making is that by taking c close to the threshold value for which S is singular, the resulting problem is likely to have a number of active bounds. The permutation technique described in § 5 works well in handling this situation. In contrast the Woodhouse test problems have very few active bounds and the permutation technique is rarely required.

Appendix. Subgradients for extreme eigenvalues. These results all concern the space of $n\times n$ symmetric matrices A and the function

$$f(A)=\sum_{i=1}^{m}\lambda_i(A) \tag{A.1}$$

which is the sum of the m largest eigenvalues of A where $1\leqq m\leqq n$ and where $\lambda_i(A)$ $i=1,2,\cdots,n$ are the eigenvalues of A ordered according to $\lambda_1\geqq\lambda_2\geqq\cdots\geqq\lambda_n$. The results specialize to the largest eigenvalue alone ($m=1$), and there are corresponding results for the least eigenvalues by considering $f(-A)$.

THEOREM A.1. *$f(A)$ is convex.*

Proof. An equivalent Rayleigh quotient form of $f(A)$ is

$$f(A)=\max_{\{Q\mid Q^TQ=I\}}\operatorname{tr}(Q^TAQ), \qquad Q\in\mathbb{R}^{n\times m}. \tag{A.2}$$

(This follows by expanding $Q = XY$ where X is the matrix whose columns are the eigenvectors of A, and $Y \in \mathbb{R}^{n \times m}$. Then $\operatorname{tr} Q^T A Q = \operatorname{tr} Y^T \Lambda Y = \operatorname{tr} \Lambda Y Y^T$. The condition $Q^T Q = I$ implies $Y^T Y = I$, so $\operatorname{tr} Y^T Y = \operatorname{tr} Y Y^T = m$ and hence $\operatorname{tr} \Lambda Y Y^T$ is maximized when $Y = \left[\begin{smallmatrix} I \\ 0 \end{smallmatrix}\right]$.) Now consider $f(A)$ for two symmetric matrices A_0 and A_1. Then for any $\theta \in [0, 1]$, and using linearity of the trace,

$$\begin{aligned} f((1-\theta)A_0 + \theta A_1) &= \max_{\{X \mid X^T X = I\}} \operatorname{tr} X^T((1-\theta)A_0 + \theta A_1)X \\ &\leqq \max_{\{X \mid X^T X = I\}} (1-\theta) \operatorname{tr} X^T A_0 X + \max_{\{X \mid X^T X = I\}} \theta \operatorname{tr} X^T A_1 X \\ &= (1-\theta) f(A_0) + \theta f(A_1). \end{aligned}$$

Thus $f(A)$ is convex. □

COROLLARY A.2. *The sum of the m least eigenvalues*

$$g(A) = \sum_{i=n-m+1}^{n} \lambda_i(A) \tag{A.3}$$

is concave (since $g(A) = -f(-A)$*).* □

Before proving the second theorem, a preliminary result is required.

LEMMA A.3. *If* a_{ii} $i = 1, 2, \cdots, n$ *are fixed, then* $f(A)$ *is minimized by the diagonal matrix* $D = \operatorname{diag}(a_{ii})$.

Proof. Consider a perturbation of D to $D + \varepsilon M$ where M is symmetric and $m_{ii} = 0$ $i = 1, 2, \cdots, n$. Define the orthogonal matrix $Q = e^{\varepsilon S}$ where S is skew-symmetric. Then $D + \varepsilon M$ is similar to $Q^T(D + \varepsilon M)Q$ and the first order terms of this matrix are $\varepsilon(DS - SD + M)$. Choosing

$$s_{ij} = \begin{cases} m_{ij}/(a_{jj} - a_{ii}) & \text{if } a_{ii} \neq a_{jj}, \\ 0 & \text{otherwise} \end{cases}$$

eliminates off-diagonal terms when $a_{ii} \neq a_{jj}$. Thus, apart from the $O(\varepsilon^2)$ terms, $Q^T(D + \varepsilon M)Q$ is block diagonal, and within each block the diagonal elements are equal. Now eliminate each off-diagonal element in each block (to within $O(\varepsilon^2)$) by a Jacobi rotation of angle $\pi/4$. Each rotation leaves the trace of each block unchanged, but changes the ii and jj diagonal elements by $-\varepsilon m_{ij}$ and εm_{ij} respectively. Thus the eigenvalues of $D + \varepsilon M$ can be expressed as

$$a_{ii} + \varepsilon b_i + O(\varepsilon^2)$$

where $\sum b_j = 0$, summed over any subset of all the indices for which the a_{jj} are equal. Hence

$$f(D + \varepsilon M) \geqq f(D) + O(\varepsilon^2).$$

Taking the limit $\varepsilon \downarrow 0$ it follows that the directional derivative at D is nonnegative in any direction M, and so by convexity $f(A)$ is minimized by $D = \operatorname{diag}(a_{ii})$. □

It is now possible to prove the second main result which characterizes the symmetric subgradients of $f(A)$. As in (2.4) $B + S$ is also a subgradient where S is skew-symmetric.

THEOREM A.4.

$$\partial f(A) = \{B \mid B^T = B,\ B \geqq 0,\ \operatorname{tr} B = m,\ B : A = f(A)\}. \tag{A.4}$$

Proof. B is a subgradient of $f(A)$ iff

$$B : A = f(A) + f^c(B) \tag{A.5}$$

where f^c is the conjugate function defined by

$$f^c(B) = \sup_A (B : A - f(A)) \tag{A.6}$$

(Rockafellar (1970)). Let $B = X\Omega X^T$ be the spectral decomposition of the fixed matrix B in (A.6), where $\Omega = \text{diag}(\omega_i)$ and $\omega_1 \geqq \omega_2 \geqq \cdots \geqq \omega_n$. Using the definition of $B:A$ $(= \text{tr}\, X\Omega X^T A = \Omega : C$ say, where $C = X^T A X)$, then

$$f^c(B) = \sup_C (\Omega : C - f(C)) \tag{A.7}$$

where the sup is taken over symmetric matrices C. Consider changing only the off-diagonal elements of C; the term $\Omega : C$ is unaffected and by Lemma A.3, $f(C)$ is minimized when the off-diagonal elements are zero. Hence only variations in $C = \text{diag}\, c_i$ need be considered. In this case $f(C)$ is just the sum of the m largest elements c_i, and

$$\sup_C (\textstyle\sum \omega_i c_i - f(C)) = \begin{cases} \infty & \text{if any } \omega_i < 0 \quad (\text{let } c_i \to -\infty), \\ \infty & \text{if } \sum \omega_i > m \quad (\text{let } c_i = \alpha \to \infty), \\ \infty & \text{if } \sum \omega_i < m \quad (\text{let } c_i = \alpha \to -\infty), \\ 0 & \text{if } \omega_i \geqq 0 \quad \text{and} \quad \sum \omega_i = m, \end{cases}$$

since in the last case $\sum \omega_i c_i \leqq f(C)$ with equality when

$$C = \begin{bmatrix} I_m & 0 \\ 0 & 0 \end{bmatrix}.$$

Applying these results together with (A.5) gives the required conditions (A.4). □

COROLLARY A.5. *B is a symmetric subgradient iff $B = XX^T$ where columns of $X \in \mathbb{R}^{n \times m}$ are any orthogonal set of eigenvectors for the eigenvalues $\lambda_1, \cdots, \lambda_m$ of A (follows directly from* (A.4)).

REFERENCES

J. CULLUM, W. E. DONATH AND P. WOLFE (1975), *The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices* in Nondifferentiable Optimization, M. Balinski and P. Wolfe, eds., Mathematical Programming Study 3, North-Holland, Amsterdam.

R. FLETCHER (1981a), *A nonlinear programming problem in statistics (educational testing)*, SIAM J. Sci. Stat. Comput., 2, pp. 257-267.

——— (1981b), *Practical Methods of Optimization, Volume 2, Constrained Optimization*, John Wiley, Chichester.

——— (1981c), *Numerical experiments with an exact l_1 penalty function method*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer and S. M. Robinson, eds., Academic Press, New York.

——— (1982a), *Second order corrections for nondifferentiable optimization*, in Numerical Analysis, Dundee 1981, G. A. Watson, ed., Lecture Notes in Mathematics 912, Springer-Verlag, Berlin.

——— (1982b), *Semi-definite matrix constraints in optimization*, Dept. Mathematical Sciences, Numerical Analysis Report NA/61, Univ. Dundee, Scotland.

S. P. HAN (1977), *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22, pp. 297-309.

S. JAYARAJAN (1979), *A nonlinear optimization problem in educational testing*, MSc thesis, Dept. Mathematics, Univ. Dundee, Scotland.

D. Q. MAYNE AND E. POLAK (1982), *Algorithms for the design of control systems subject to singular value inequalities*, in Algorithms and Theory in Filtering and Control, D. C. Sorensen and R. J-B. Wets, eds., Mathematical Programming Study 18, North-Holland, Amsterdam.

S. P. PANG (1981), *A numerical method for an optimization problem in statistics*, MSc thesis, Dept. Mathematics, Univ. Dundee, Scotland.

R. T. ROCKAFELLAR (1970), *Convex Analysis*, Princeton Univ. Press, Princeton, NJ.

——— (1981), *The Theory of Subgradients and Its Applications to Problems of Optimization. Convex and Nonconvex Functions*, Research and Education in Mathematics 1, Heldermann Verlag, Berlin.

B. WOODHOUSE (1976), *Lower bounds for the reliability of a test*, MSc thesis, Dept. Statistics, Univ. Wales, Aberystwyth.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

002

CONVERGENCE OF VIABLE SOLUTIONS OF DIFFERENTIAL INCLUSIONS WITH CONVEX COMPACT GRAPHS*

ARIE LEIZAROWITZ†

Abstract. A convergence property for the viable solutions of a differential inclusion with convex and compact graph is established. We show that if there is a unique stationary point and no elliptic solutions, then all the viable solutions converge uniformly to the stationary point. We show that the convergence property is generic for the set of differential inclusions with a unique stationary point.

Key words. differential inclusions, set valued functions, viable solutions, stationary points

1. Introduction. In this work we establish a convergence property for the viable solutions of a differential inclusion with a convex and compact graph. Viable solutions are those which are defined for all positive times. The main result asserts that if there is a unique stationary point and no elliptic solutions, then all the viable solutions converge, uniformly, to the stationary point. This result is presented in § 3; in § 2 we display our notations and some definitions. In § 4 we show that the convergence property is generic, namely, for most differential inclusions with convex and compact graphs, a sufficient condition for all the viable solutions to converge is the existence of a unique stationary point.

The motivation for studying this property arises from the study of the following infinite horizon problem. Consider, for every trajectory $z(\cdot):[0,\infty)\to R^n$, the cost flow

$$c(t)=\int_0^t L(z(s),\dot{z}(s))\,ds$$

where $L(\cdot,\cdot)$ is a convex scalar function defined in $R^n\times R^n$. An overtaking optimal trajectory is one whose cost flow is less than the cost flow of any other trajectory with the same initial value, from a certain time on. The problem is the existence of an overtaking optimal trajectory, given an initial value $z(0)=z_0$. This problem is studied in Brock and Haurie [2] and Leizarowitz [5]. In [5] the existence of overtaking optimal trajectories is established while assuming that a certain differential inclusion $\dot{z}\in G(z)$ (which will be described below) is such that all its viable solutions converge uniformly to a certain point. The set valued function $z\to G(z)$ is related to $L(\cdot,\cdot)$ as follows: It is assumed that L has the representation

$$L(z,w)=L_0(z,w)+\eta' w+\alpha$$

where $\alpha\in R^1$, $\eta\in R^n$, and $L_0(z,w)\geqq 0$ is a convex function with the property

$$L_0(z,0)=0 \quad \text{if and only if} \quad z=\bar{z}$$

for a certain $\bar{z}\in R^n$. We define the function $z\to G(z)$ by defining its graph in $R^n\times R^n$ as

$$\text{graph } G=\{(z,w):\ L_0(z,w)=0\}$$

(namely $w\in G(z)$ if and only if $(z,w)\in G$). The main result of [5] is that if the viable solutions of $\dot{z}\in G(z)$ have the desired convergence property, and if the initial value

* Received by the editors October 13, 1983, and in revised form June 29, 1984.

† Department of Theoretical Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel. Present address, Institute for Mathematics and Its Applications, University of Minnesota, Minneapolis, Minnesota 55455.

can be steered in a finite time to their common limit, then there exists an overtaking optimal trajectory satisfying the initial condition.

2. Notation and terminology. We consider a differential inclusion

$$\frac{dz(t)}{dt} \in G(z(t)) \tag{2.1}$$

where $z \to G(z)$ is a set valued function defined on a bounded domain in R^n, such that $G(z) \subset R^n$. We assume throughout that the graph of the set valued map, namely the set

$$G = \{(z, w) \in R^n \times R^n : w \in G(z)\}$$

is compact and convex.

An absolutely continuous function $z:(t_1, t_2) \to R^n$ with $t_1 < t_2$ is called *a solution of* (2.1) if it satisfies (2.1) almost everywhere in the interval (t_1, t_2). A *viable solution of* (2.1) is a solution on $(0, \infty)$, namely one which persists for infinite time. (This terminology is due to J.-P. Aubin. For a detailed discussion concerning differential inclusions see Aubin and Cellina [1].)

A function $z:(0, \infty) \to R^n$ is called an *elliptic solution* of (2.1) if it is a viable solution of (2.1) having the form

$$z(t) = a \cos \alpha t + b \sin \alpha t$$

for some scalar $\alpha \neq 0$ and $a, b \in R^n$ with $|a| + |b| \neq 0$ (here $|\cdot|$ is the Euclidean norm on R^n).

We denote the scalar product of the vectors $x, y \in R^n$ by $\langle x, y \rangle$. We shall say that the point $z \in R^n$ *is a stationary point* of G if $0 \in G(z)$. One can easily verify that the existence of a viable solution of (2.1) implies the existence of a stationary point of G. (Indeed, if $z(\cdot)$ is a viable solution, then the values

$$\left(\frac{1}{T_i}\int_0^{T_i} z(t)\,dt, \frac{1}{T_i}\int_0^{T_i} \dot{z}(t)\,dt\right)$$

converge to some $(\bar{z}, 0) \in G$ for a suitable sequence $T_i \to \infty$.) Thus the existence of a stationary point will be assumed and for convenience we assume that 0 is such a point, namely:

$$0 \in G(0). \tag{2.2}$$

3. The main result. In the proof of the main result, Theorem 3.2, we shall use the following well-known fact.

LEMMA 3.1. *The set of solutions of* (2.1) *defined on the interval* $[0, T]$ *is closed as a subset of* $C[0, T]$ (*namely, the set of all continuous functions on* $[0, T]$ *endowed with the norm* $\|z(\cdot)\| = \max_{0 \leq t \leq T} |z(t)|$).

THEOREM 3.2. *Let* $z \to G(z)$ *be a set valued function with a convex and compact graph* $G \subset R^n \times R^n$ *and such that* $0 \in G(0)$.

Then all the viable solution of (2.1) *converge, in a uniform rate, to zero if and only if*

(i) *zero is the only stationary point of* G,

(ii) *there are no elliptic solutions of* (2.1).

A remark. By saying that all the viable solutions of (2.1) converge uniformly to zero we mean that, given an $\varepsilon > 0$, there is a $T > 0$ such that $|z(t)| < \varepsilon$ for every $t > T$ and every viable solution $z(\cdot)$.

A remark. Note that the assertion of the theorem concerns only the viable solutions; in particular it holds when there are no viable solutions at all. For example, let $C \subset R^n$ be a compact and convex set and A be an $n \times n$ matrix, with all the eigenvalues having a nonzero real part. Then the theorem applies to the differential equation $dz(t)/dt = Az(t)$ subject to the constraint $z(t) \in C$ for all $t \geqq 0$. Clearly when all the eigenvalues have a positive real part, there are no nontrivial viable solutions. Thus the asymptotic stability property which we study here is different from the usual asymptotic stability of differential inclusions. To the latter applies, for example, the method of Lyapunov functions (see Roxin [6]), which seems to be inadequate in our context.

A remark. We discuss here the meaning and some consequences of the assumptions concerning the set valued function $z \to G(z)$, in particular Assumption (i) of Theorem 3.2. A discussion concerning Assumption (ii) of that theorem is presented in § 4. In the special case $z \in R^1$ our assumptions imply that the graph of G in R^2 is a nonhorizontal line segment which contains the origin or else, it is contained in the upper or lower half of the plane. In higher dimensional spaces the structure is more complicated. Consider, for example, a control system $\dot{z} = Az + Bu$, $z \in K$, $u \in U$, where $K \subset R^n$ and $U \subset R^m$ are convex and compact sets. Then the set $G = \{(z, \dot{z})\colon z \in K, u \in U\}$ satisfy our assumptions provided that $A(K) \cap B(-U) = \{0\}$ and $Az = 0$, $z \in K$ imply $z = 0$. On the other hand, not every set $G \subset R^n \times R^n$ satisfying our assumptions can be represented as above as arising from a linear control system.

A counterexample is the following: Consider $z = (z_1, z_2) \in R^2$, let $e = (1, 0)$ and denote $C = \{z\colon (z_1 - 1)^2 + z_2^2 = 1\}$. Define $g\colon C \to R^2$ by $g(z) = |z|e$ and let $G = \operatorname{conv}\{(z, g(z))\colon z \in C\}$. Then G satisfies our assumptions. On C the function G is single valued; hence it coincides with g there. We claim that there are no A, B and U such that $G = \{(z, Az + Bu)\colon z \in K, u \in U\}$, since such an equality would imply that $B(U)$ consists of only one point, say $a \in R^2$; hence $g(z) = Az + a$ for all $z \in C$. Substituting $z = 0$, we obtain $a = 0$; therefore $g(z) = Az$, which contradicts the choice $g(z) = |z|e$ (e.g. $g((1, 1)) + g((1, -1)) = 2\sqrt{2}\, e \neq g((2, 0))$).

A discussion running along very similar lines, yields the following conclusion: Not every set $G \subset R^n \times R^n$ satisfying our assumptions can be represented as $G = \{Az \in B(z)u\colon z \in K, u \in U\}$ where $K \subset R^n$ and $U \subset R^m$ are compact and convex, $B(z)$ is continuous in z, and U has a nonempty interior in R^m.

Proof of Theorem 3.2. The "only if" part of the assertion is obvious and we shall prove here the "if" part.

For any compact and convex set $F \subset R^n \times R^n$ we denote

$$F_W = \{w \in R^n\colon (z, w) \in F \text{ for some } z \in R^n\},$$

$$F_Z = \{z \in R^n\colon (z, w) \in F \text{ for some } w \in R^n\}.$$

For G, the graph of $z \to G(z)$, we have $0 \in G_W$ and $0 \in G_Z$, this since $(0, 0) \in G$. Let K be the face of 0 in G_W, namely the largest convex subset of G_W which contains zero in its relative interior. We consider the set $H \subset G$ given by

$$H = \{(z, w) \in G\colon w \in K\}. \tag{3.1}$$

The idea of the proof is the following: We shall first prove that all the solutions of $dz(t)/dt \in H(z(t))$ converge uniformly to zero. Here $z \to H(z)$ is the set valued mapping whose graph is H. Then we shall prove that every solution of $dz(t)/dt \in G(z(t))$ converges uniformly on compact intervals, to the set of solutions of $dz(t)/dt \in H(z(t))$. Combining these two pieces of information will verify the assertion of the theorem.

With the above notation for H_W and H_Z we claim that for each $v \in H_W$ there is a unique element $s \in H_Z$ satisfying $(s, v) \in H$. To see this we consider the set valued function $w \to T(w)$ defined by $T(w) = \{s \in H_Z : (s, w) \in H\}$. We note that this function has a convex graph, that by (i) of Theorem 3.2 we have $T(0) = \{0\}$, and that zero belongs to the relative interior of H_W. These facts imply that $T(v)$ is a singleton for every $v \in H_W$.

Being single valued and of a convex graph, T is the restriction to H_W of some linear map B: Span $H_W \to$ Span H_Z.

CLAIM 1. *All the viable solutions* $z(\cdot)$ *of*

$$\frac{dz(t)}{dt} \in H(z(t)) \tag{3.2}$$

converge uniformly to zero.

We prove the claim by constructing a subset $H' \subset H$ with the following properties:

(a) Span $H'_W =$ Span H'_Z.

(b) The differential inclusion $dz(t)/dt \in H'(z(t))$ has the same set of viable solutions as the relation (3.2). Here $z \to H'(z)$ is the set valued function whose graph is H'.

Assume for a moment that such a set H' does exist. Then the restriction of B to Span H'_W is one-to-one and onto itself, and let us denote its inverse by A. Then every solution of $(z(t), \dot{z}(t)) \in H'$ satisfies the equation $dz(t)/dt = Az(t)$ in H'_Z. Since zero belongs to the relative interior of H'_Z, and since by our assumption there are no elliptic trajectories, it follows that A has no purely imaginary eigenvalues. Thus all the bounded viable solutions of $dz(t)/dt = Az(t)$ converge uniformly to zero as $t \to \infty$, and the assertion of the claim follows.

We now construct a set H' with the desired properties. We denote $H_0 = H$ and $X_0 =$ Span $(H_0)_W$, $Y_0 =$ Span $(H_0)_Z$. We define $X_{i+1} = X_i \cap Y_i$, $Y_{i+1} = B(X_{i+1})$ and $H_{i+1} = \{(z, w) \in H_i : z \in Y_{i+1}, w \in X_{i+1}\}$. We show that H_i and H_{i+1} have the same set of viable solutions and since $H_{i+1} \subset H_i$ it is enough to show that every solution of H_i is a solution of H_{i+1}. Let $z(\cdot)$ satisfy $dz(t)/dt \in H_i(z(t))$. Then by the definition of H_i we have that $z(t) \in Y_i$ for all $t \geqq 0$; hence $\dot{z}(t) \in Y_i \cap X_i = X_{i+1}$ for all $t \geqq 0$, which proves that $z(\cdot)$ solves $dz/dt \in H_{i+1}(z(t))$.

It follows from $X_i \neq Y_i$ that dim $X_{i+1} <$ dim X_i; therefore we conclude that $X_i = Y_i$ for some $i \geqq 0$, and then $H' = H_i$ has the desired properties. Thus Claim 1 is established.

Recall that the set K is the face of zero in G_W. If $K = G_W$, then (3.1) implies that $H = G$ and the assertion of the theorem follows from Claim 1. Assume therefore that K is a proper subset of G_W. Then there is an $\eta \in R^n$ such that

$$\langle \eta, w_0 \rangle \geqq 0 \quad \text{for every } w_0 \in G_W, \qquad \langle \eta, v \rangle > 0 \quad \text{for some } v \in G_W, \tag{3.3}$$

$$\langle \eta, w_0 \rangle = 0 \quad \text{for every } w_0 \in K. \tag{3.4}$$

CLAIM 2. *Denoting* $\Sigma = \{s \in G_Z : \langle \eta, s \rangle = 0\}$, *we have for every viable solution* $z(\cdot)$ *of* (2.1) *the following*:

$$\operatorname{dist}(z(t), \Sigma) \to 0 \quad \text{as } t \to \infty, \text{ uniformly in } z(\cdot). \tag{3.5}$$

Let $z(\cdot)$ be any viable solution of (2.1). Then the function $t \to \langle \eta, z(t) \rangle$ is nondecreasing (by (3.3)); hence it converges, say to α. This clearly implies that $\langle \eta, (1/T) \int_0^T z(t)\, dt \rangle$ converges to α as $T \to \infty$.

We claim that for every viable solution of (2.1) we have $(1/T) \int_0^T z(t)\, dt \to 0$ as $T \to \infty$. Otherwise, there is an $\varepsilon > 0$ and a sequence $T_i \to \infty$ such that $\left|(1/T_i) \int_0^{T_i} z(t)\, dt\right| \geqq \varepsilon$ for each i. Since the convexity of G implies $((1/T_i) \int_0^{T_i} z(t)\, dt, (1/T_i)[z(T_i) - z(0)]) \in G$

and clearly $(1/T_i)(z(T_i)-z(0))\to 0$ as $i\to\infty$, the compactness of G implies a contradiction to condition (i) of Theorem 3.2.

Thus we conslude that $\langle \eta, z(t)\rangle \to 0$ as $t\to\infty$. We claim that this convergence is uniform in $z(\cdot)$. This is a consequence of the monotonicity of $t\to\langle \eta, z(t)\rangle$. If the convergence is nonuniform in $z(\cdot)$, then, for some $\varepsilon>0$, we can find a sequence of viable solutions $\{z_k(\cdot)\}$ and an increasing sequence $t_k\to\infty$ as $k\to\infty$, such that $\langle \eta, z_k(t)\rangle \leqq -\varepsilon$ for all $0\leqq t\leqq t_k$. It is easy to see from Lemma 3.1 that a subsequence of $\{z_k(\cdot)\}$ converges on every compact interval $[0, T]$, to some viable solution $z_0(\cdot)$ of (2.1). But then $\langle \eta, z_0(t)\rangle \leqq -\varepsilon$ for all $t\geqq 0$, contradicting $\langle \eta, z_0(t)\rangle\to 0$ as $t\to\infty$. The uniform convergence to zero of $\langle \eta, z(t)\rangle$ implies, by (3.4), the validity of (3.5) and concludes the proof of the claim.

We now construct a sequence of convex and compact sets

$$G_W = K_0 \supset K_1 \supset \cdots \supset K_r = K$$

as follows:

Let η_1 be the vector η which appears in (3.3) and (3.4), and define

$$K_1 = \{w\in K_0 : \langle \eta_1, w\rangle = 0\}.$$

Clearly, by (3.4), $K\subset K_1$.

If zero belongs to the relative interior of K_1, then we stop. Otherwise we repeat the construction and choose a vector $\eta_2\in R^n$ such that $\langle \eta_2, w\rangle \geqq 0$ for all $w\in K_1$ and $\langle \eta_2, v\rangle > 0$ for some $v\in K_1$, $\langle \eta_2, w\rangle = 0$ for all $w\in K$. We define

$$K_2 = \{w\in K_1 : \langle \eta_2, w\rangle = 0\}.$$

We continue this construction until we arrive at a set K_r which contains zero in its relative interior (which must occur since the dimensions of the K_i are strictly decreasing as long as the construction continues). It is clear from the construction that $K\subset K_r$ and from the definition of K that $K_r\subset K$, thus $K_r = K$.

CLAIM 3. *Let $z(\cdot)$ be a solution of* (2.1) *defined on* $(-\infty,\infty)$ *and such that for some* $0\leqq i<r$ *the following holds*: $dz(t)/dt\in K_i$ *for all* $-\infty<t<\infty$. *Then*

$$\frac{dz(t)}{dt}\in K_{i+1} \quad \textit{for all } -\infty<t<\infty.$$

We consider in Claim 2 the set K_i instead of G_W. Analogous to Σ we define

$$\Sigma_{i+1} = \{s\in G_Z : \langle \eta_{i+1}, s\rangle = 0\}.$$

Then it follows from Claim 2 that $\operatorname{dist}(z(t), \Sigma_{i+1}) = 0$ for all $-\infty<t<\infty$; therefore $\langle \eta_{i+1}, z(t)\rangle = 0$ for all $-\infty<t<\infty$. Hence $\langle \eta_{i+1}, dz(t)/dt\rangle = 0$ for all $-\infty<t<\infty$, proving the claim.

CLAIM 4. *All the viable solutions of* (2.1) *converge, uniformly, to zero.*

Assume that the claim is false. Then there are an $\varepsilon>0$ and a sequence of solutions $\{z_k(\cdot)\}_{k=1}^{\infty}$ with a sequence $t_k\to\infty$ as $k\to\infty$ such that $|z_k(t_k)|\geqq\varepsilon$ for all $k\geqq 1$. We consider the sequence of solutions $\{s_k(\cdot)\}$ defined by:

$$s_k(t) = z_k(t+t_k) \quad \text{for } -t_k\leqq t<\infty.$$

Then we have $|s_k(0)|\geqq\varepsilon$ for all $k\geqq 1$.

It is easy to see from Lemma 3.1 that a subsequence of $\{s_k(\cdot)\}$ converges, uniformly on compact intervals $[-T, T]$, to some solution $s_0(t)$ of (2.1) ($s_0(t)$ is a solution in the interval $(-\infty,\infty)$). It follows from Claim 3 that $ds_0(t)/dt\in K$ for all $-\infty<t<\infty$ and

by (3.1) we have that $ds_0(t)/dt \in H(s_0(t))$ for all $-\infty < t < \infty$. Now Claim 1 implies that $s_0(t) \equiv 0$, contradicting $|s_0(0)| \geqq \varepsilon$ and concluding the proof of the theorem. □

Example 3.3. Let $G \subset R^n \times R^n$ be the convex and compact set consisting of points $(z, w) \in R^n \times R^n$ which satisfy

$$|w - w_0| + |w - Az| \leqq |w_0|$$

for some nonzero $w_0 \in R^n$ and a nonsingular $n \times n$ matrix A. Let $z \to G(z)$ be the set valued function whose graph is G. For each z, the set $G(z)$ is an ellipsoid of revolution with foci in w_0 and Az. We check that the conditions in the sufficiency part of Theorem 3.2 are fulfilled. If $0 \in G(z)$, then $|Az| \leqq 0$; hence $z = 0$ by the regularity of A. We claim that there are no elliptic solutions for $dz/dt \in G(z(t))$. For every solution we have that $|\dot{z}(t) - w_0| \leqq |w_0|$; therefore $\langle w_0, \dot{z}(t) \rangle \geqq 0$ with equality only when $\dot{z}(t) = 0$. The existence of an elliptic solution $z(\cdot)$ of period $T > 0$ would imply $\int_0^T \langle w_0, \dot{z}(t) \rangle \, dt > 0$, contradicting the periodicity of $z(\cdot)$. Thus we conclude that all the viable solutions converge, at a uniform rate, to zero.

We consider now a slightly different problem and ask the following question: When does every viable solution $z(\cdot)$ of (2.1) converge to some constant? Now it is not necessary for G to have a unique stationary point, and we define P to be the set of all stationary points of G, namely

$$P = \{z \in R^n : (z, 0) \in G\}. \tag{3.6}$$

By (2.2) we have $0 \in P$.

The following lemma, which will imply our result, shall be needed for future reference.

LEMMA 3.4. *Let $z \to G(z)$ be a set valued function with a convex and compact graph $G \subset R^n \times R^n$. Let P be defined by* (3.6) *and assume that*

$$(\text{Span } P) \cap G_W = \{0\}. \tag{3.7}$$

Then all the viable solutions of (2.1) *which satisfy $z(0) = z_0$ for a fixed z_0, are viable solutions of a differential inclusion*

$$\frac{dz(t)}{dt} \in F(z(t)) \tag{3.8}$$

where the mapping $z \to F(z)$ has a unique stationary point.

Proof. Let $z(\cdot)$ be a viable solution of (2.1). Then $z(t) \in z(0) + G_W$ for all $t \geqq 0$. Hence $z(\cdot)$ is a viable solution of (3.8) where the mapping $z \to F(z)$ is the set valued function whose graph is

$$F = \{(z, v) \in G : z \in z(0) + G_W\}.$$

Then (3.7) implies that there is at most one stationary point for F and by the existence of a viable solution there is exactly one stationary point. Thus the assertion follows. □

The following result is a consequence of Theorem 3.2 and Lemma 3.4.

THEOREM 3.5. *Let $z \to G(z)$ be a set valued function with a convex and compact graph $G \subset R^n \times R^n$. Let P be as in* (3.6) *and assume that* (3.7) *holds. Then every viable solution of* (2.1) *converges, as $t \to \infty$, to some point in P if and only if there is no elliptic solution of* (2.1).

4. The nonexistence of elliptic solutions is generic. The condition which appears in (ii) of Theorem 3.2 is that there is no elliptic solution which solves (2.1). In the proof of Theorem 3.2 this is shown to be equivalent to the nonexistence of some purely

imaginary eigenvalue for a certain matrix. This suggests that in some natural sense, condition (ii) in Theorem 3.2 is generic. In this section we establish this property.

Let Y be the metric space of all the compact sets $K \subset R^n \times R^n$ with the Hausdorff metric ρ which is defined as

$$\rho(K_1, K_2) = \max \{\max_{a \in K_1} \min_{b \in K_2} d(a, b), \max_{b \in K_2} \min_{a \in K_1} d(a, b)\}$$

where $d(\cdot, \cdot)$ is the following metric on $R^n \times R^n$:

$$d((x, y), (x', y')) = |x - x'| + |y - y'|.$$

(See Kelly and Weiss [4, p. 237], for a verification that Y is a complete metric space.)

We consider the set $X \subset Y$ of all the convex and compact sets $G \subset R^n \times R^n$ for which condition (i) of Theorem 3.2 holds, namely, $(z, 0) \in G$ if and only if $z = 0$. Then X itself, with the induced topology, is a metric space. Notice that $X \neq \text{cl } X \neq Y$ (where cl X is the closure of X in Y), and X is not a complete metric space.

We are interested in the following subset of X:

$$A = \{G \in X: \text{there is no elliptic solution of (2.1)}\}.$$

We shall show that A is dense in X and that it contains a denumerable intersection of open and dense sets in X. Since X is not a complete metric space, this does not imply yet that A is big in the topological sense. However, we shall prove that A is residual in cl X too. In this sense the condition of Theorem 3.2 holds generically for sets G in X. (For more details concerning residual sets in a complete metric space consult Kelley [3, pp. 200-202].)

PROPOSITION 4.1. *The set A is dense in X and contains a denumerable intersection of open and dense sets in X.*

Proof. Let us introduce the following terminology. We consider a linear function M from a subspace S of R^n into R^n. We enlarge M to a linear function $\bar{M}$ on the entire R^n by defining $\bar{M}$ to be zero on the complement of S. We shall say that M *has no purely imaginary eigenvalues* if $\bar{M}$ does not have such eigenvalues.

It is remarked in the beginning of the proof of Theorem 3.2 that one can associate with each set $K \in X$ a linear function M_K, defined on the face of the point 0 in K_W. Then it is easy to see that $K \in A$ if and only if M_K has no purely imaginary eigenvalues.

Let $K \in X$ and $\varepsilon > 0$. Then there is an $n \times n$ matrix M such that

$$|M - \bar{M}_K| < \varepsilon \tag{4.1}$$

and the restriction of M to the face of zero in K_W has no purely imaginary eigenvalues. The set K is the graph of the following set valued function $w \to T(w)$ defined by

$$T(w) = \{z \in R^n: (z, w) \in K\}.$$

We define the set valued function

$$\bar{T}(w) = T(w) + (M - \bar{M}_K)w$$

both defined on the same domain K_W. The graph of $\bar{T}$, denoted $\bar{K}$, is a set in X, and $\rho(K, \bar{K})$ is small if ε in (4.1) is small. Clearly, $\bar{K} \in A$; thus A is dense in X.

We prove now that A contains a denumerable intersection of open and dense sets in X. For each integer $m \geqq 2$ we define the family F_m of all sets $K \in X$ with the following property: There exists an elliptic solution $z(\cdot)$, $z(t) = a \cos \alpha t + b \sin \alpha t$, satisfying $(z(t), \dot{z}(t)) \in K$ such that $1/m \leqq |\alpha| \leqq m$ and $|a| + |b| \geqq (1/m)$. We shall prove that F_m is closed in X. Since F_m is disjoint from A and since A is dense in X, we have that

each $X \setminus F_m$ is an open and dense set in X. Thus the relation $A = \bigcap_{m=2}^{\infty} (X \setminus F_m)$ will complete the proof of the proposition.

Let $K_i \in F_m$ for every $i \geqq 1$, and let $\rho(K_i, K) \to 0$ as $i \to \infty$, for some $K \in X$. We should prove that $K \in F_m$. For each i let $z_i(t) = a_i \cos \alpha_i t + b_i \sin \alpha_i t$, where $1/m \leqq |\alpha_i| \leqq m$ and $1 \geqq |a_i| + |b_i| \geqq 1/m$, be an elliptic solution for K_i. It may be assumed that $a_i \to a$, $b_i \to b$ and $\alpha_i \to \alpha$ as $i \to \infty$, where $1/m \leqq |\alpha| \leqq m$ and $1 \geqq |a| + |b| \geqq 1/m$. Then the elliptic solution $z(t) = a \cos \alpha t + b \sin \alpha t$ satisfies $(z(t), \dot{z}(t)) \in K$ for all $t \geqq 0$, proving that $K \in F_m$ and concluding the proof of the proposition. □

THEOREM 4.2. *The set A is residual in* cl X, *i.e., A contains a denumerable intersection of open and dense sets in* cl X.

Proof. We shall prove that cl $X \setminus X$ is of first category in cl X, namely

$$\text{cl } X \setminus X = \bigcup_{k=1}^{\infty} H_k \tag{4.2}$$

where H_k are closed and nowhere dense in cl X. Once we have proved this, let $\{O_i\}_{i=1}^{\infty}$ be the open and dense sets in X which are guaranteed by Proposition 4.1, that is:

$$O_i = U_i \cap X \tag{4.3}$$

for some open set $U_i \subset Y$, and

$$A \supset \bigcap_{i=1}^{\infty} O_i.$$

Then, defining

$$V_i = (U_i \cap \text{cl } X) \setminus \bigcup_{k=1}^{i} H_k$$

with U_i and H_k as in (4.3) and (4.2) respectively, we have that each V_i is open and dense in cl X, and $\bigcap_{i=1}^{\infty} V_i \subset \bigcap_{i=1}^{\infty} (U_i \cap \text{cl } X) \setminus \bigcup_{k=1}^{\infty} H_k \subset \bigcap_{i=1}^{\infty} O_i \subset A$ proving that A is residual in cl X.

To prove that cl $X \setminus X$ is of first category in cl X, let us denote

$$W_j = \left\{ K \in \text{cl } X : (z, 0) \in K \text{ implies } |z| < \frac{1}{j} \right\};$$

then

$$X = \bigcap_{j=1}^{\infty} W_j. \tag{4.4}$$

We prove now that W_j is open in cl X. Let $\{K_l\}_{l=1}^{\infty} \subset \text{cl } X \setminus W_j$ be such that $K_l \to K_0$ as $l \to \infty$. For every $l \geqq 1$ there is a z_l such that $|z_l| \geqq 1/j$ and $(z_l, 0) \in K_l$. There is a subsequence of $\{z_l\}_{l=1}^{\infty}$ which converges to some point z_0 which satisfies $|z_0| \geqq 1/j$ and $(z_0, 0) \in K_0$. Thus each W_j is an open set, and it follows from (4.4) that it is also dense in cl X; hence X is residual in cl X, concluding the proof of the theorem. □

Acknowledgment. This work is part of a Ph.D. dissertation which was conducted under the supervision of Professor Zvi Artstein whom I wish to thank for fruitful discussions and very helpful advice.

REFERENCES

[1] J.-P. AUBIN AND A. CELLINA (1983), *Differential Inclusions*, Springer-Verlag, New York.

[2] W. A. BROCK AND A. HAURIE (1976), *On existence of overtaking optimal trajectories over an infinite time horizon*, Math. Oper. Res., 1, pp. 337-346.

[3] J. L. KELLEY (1955), *General Topology*, Van Nostrand, New York.
[4] P. J. KELLY AND M. L. WEISS (1979), *Geometry and Convexity*, John Wiley, New York.
[5] A. LEIZAROWITZ, *Existence of overtaking optimal trajectories for problems with convex integrands*, Math. Oper. Res., to appear.
[6] E. ROXIN (1965), *Stability in general control systems*, J. Differential Equations, 1, pp. 115–150.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

$(\mathrm{Ad}_{f,g})$, $(\mathrm{ad}_{f,g})$ AND LOCALLY $(\mathrm{ad}_{f,g})$ INVARIANT AND CONTROLLABILITY DISTRIBUTIONS*

ARTHUR J. KRENER†

Abstract. In the study of nonlinear control systems, the concepts of an invariant foliation and an invariant distribution play important roles. In this paper we explore various forms of these concepts and show how they occur in the study of controllability, observability and decoupling of nonlinear systems.

Key words. invariant foliation, invariant distribution, nonlinear controllability, nonlinear observability, nonlinear decoupling

1. Introduction. Through the work of many researchers over the past decade it has become clear that concepts from differential topology such as foliation and invariant distribution play a crucial roll in the study of nonlinear systems. These tools were first used in the study of nonlinear controllability and later observability. More recently they have arisen in the study of decoupling and linearization via feedback.

As their use has widened, a greater precision in their application has become necessary. This paper is an attempt at that precision at least as far as my own joint work with R. Hermann [12], A. Isidori, C. Gori-Giorgi and S. Monaco is concerned [7]. These papers use differential topological tools to extend to nonlinear systems the geometric approach to linear systems. Although they are quite successful, they do not have the same logical simplicity and elegance of the corresponding linear theory. This reflects a basic fact of mathematical life, nonlinearities are much messier to deal with, one usually must make strong regularity assumptions and distinguish between a much larger range of phenomena when in their presence.

In this paper we introduce the basic concepts needed for an understanding of controllability, observability and decoupling of nonlinear systems. Some of the theorems contained herein build on and are refinements of those appearing in [12] and [7]. By slightly modifying some definitions we achieve a synthesis of the previous work. From this firm platform we are able to treat controllability distributions in a precise manner and prove several interesting results.

2. Mathematical preliminaries. Throughout this paper we consider nonlinear systems of the form

$$\dot{x} = f(x, u) = g^0(x) + g(x)u, \tag{2.1a}$$

$$y = h(x), \tag{2.1b}$$

$$x(0) = x^0, \tag{2.1c}$$

where x denotes local coordinates on a smooth n-dimensional Hausdorff, paracompact connected manifold M, $u \in \mathbb{R}^m$, $y \in \mathbb{R}^p$, g^0 and $g^1, \cdots, g^m$, the m columns of g, are local coordinate descriptions of smooth vector fields globally defined on M. Smooth means either $\mathscr{C}^\infty$ or $\mathscr{C}^\omega$ (analytic). The definitions of differentiable manifold, etc. can be found in Boothby [18] or Spivak [22].

* Received by the editors October 4, 1983, and in revised form June 28, 1984. This research was supported in part by the National Science Foundation under grant MCS-8003263, by the National Aeronautics and Space Administration under grant NCA2-0Y-18 0-251 and by the Department of Energy under grant DE-AC01-80RA50421.

† Department of Mathematics, University of California, Davis, California 95616 and Flight Dynamics and Control Branch, NASA Ames Research Center 210-3, Moffet Field, California 94035.

Of course (2.1) is a local description; different descriptions of this type are valid in different coordinate neighborhoods of M. As far as possible we use local coordinate notation; hopefully this will make the paper accessible to a wider audience.

We denote by T_xM and T_x^*M the tangent and cotangent spaces at x, and by TM and T^*M the tangent and cotangent bundles. The ring of smooth real valued functions on M is denoted by $\mathcal{F}(M)$, the space of smooth vector fields (sections of TM) by $\mathcal{X}(M)$ and the space of smooth one forms (sections of T^*M) by $\mathcal{X}^*(M)$. $\mathcal{X}(M)$ and $\mathcal{X}^*(M)$ are real vector spaces and $\mathcal{F}(M)$ modules, and $\mathcal{X}(M)$ is a Lie algebra under the Lie (or Jacobi) bracket. Locally vector fields are represented by column n vectors and one forms by row n vectors. The bilinear pairing between a one form $\omega(x)$ and a vector field $X(x)$ is then the multiplication of $1\times n$ and $n\times 1$ matrices. It defines a function denoted by $\langle \omega, X\rangle \in \mathcal{F}(M)$.

A vector field X defines a flow $\Phi(t, x)$, the solution of the differential equation

$$\frac{\partial}{\partial t}\Phi(t, x) = X(\Phi(t, x)),$$

$$\Phi(0, x) = x.$$

For each x, $t \to \Phi(t, x)$ is a curve defined for t in some open interval depending on x. For some x the curve may escape from the manifold in finite time and hence not be definable for all t. We use the phrase "for all t" to mean "for all t where defined". For each t the map $x \to \Phi(t, x)$ is a smooth diffeomorphism where defined.

A vector field X or its flow $\Phi(t, x)$ acts on functions $\varphi \in \mathcal{F}(M)$, vector fields $Y \in \mathcal{X}(M)$ and one forms $\omega \in \mathcal{X}^*(M)$. The right side of the following are local coordinate descriptions which can be taken as the definitions of the symbols to the left.

(2.2a) $$\mathrm{Ad}_X^t(\varphi)(x) := \Phi(t)^*\varphi(x) := \varphi(\Phi(t, x)),$$

(2.2b) $$L_X(\varphi)(x) := \langle d\varphi, X\rangle(x),$$

(2.3a) $$\mathrm{Ad}_X^t(Y)(x) := (\Phi(-t)_* Y)(x) := \left.\frac{\partial \Phi(-t, z)}{\partial z}\right|_{z=\Phi(t, x)} Y(\Phi(t, x)),$$

(2.3b) $$\mathrm{ad}_X(Y)(x) := L_X Y(x) := [X, Y](x) := \frac{\partial Y}{\partial x}(x)X(x) - \frac{\partial X}{\partial x}(x)Y(x),$$

(2.4a) $$\mathrm{Ad}_X^t(\omega)(x) := \Phi(t)^*\omega(\Phi(t, x)) := \omega(\Phi(t, x))\left.\frac{\partial \Phi(t, z)}{\partial z}\right|_{z=x},$$

(2.4b) $$\mathrm{ad}_X(\omega)(x) := L_X(\omega)(x) := \left(\frac{\partial \omega'}{\partial x}(x)X(x)\right)' + \omega(x)\frac{\partial X}{\partial x}(x).$$

We use $'$ to denote transpose and $\partial/\partial x$ to denote partial differentiation. It is always applied to a column vector yielding a matrix with i the row and j the column index as in

$$d\varphi(x) := \frac{\partial \varphi}{\partial x}(x) = \left(\frac{\partial \varphi}{\partial x_j}\right),$$

$$\frac{\partial Y}{\partial x}(x) := \left(\frac{\partial Y_i}{\partial x_j}(x)\right),$$

$$\frac{\partial \omega'}{\partial x}(x) := \left(\frac{\partial \omega_i}{\partial x_j}(x)\right).$$

Equation (2.3b) defines the Lie bracket of vector fields. It is standard mathematical notation to denote (2.2b) by $X\varphi$ or $X(\varphi)$. We shall not employ these but instead use $X\varphi$ for X multiplied by φ.

The operator L_X of the above formulas is called Lie differentiation for

$$L_X(\varphi)(x) = \frac{d}{dt}\Big|_{t=0} \mathrm{Ad}_X^t(\varphi)(x), \tag{2.5a}$$

$$L_X(Y)(x) = \frac{d}{dt}\Big|_{t=0} \mathrm{Ad}_X^t(Y)(x), \tag{2.5b}$$

$$L_X(\omega)(x) = \frac{d}{dt}\Big|_{t=0} \mathrm{Ad}_X^t(Y)(x). \tag{2.5c}$$

The following Taylor series expansions are called Lie series:

$$\mathrm{Ad}_X^t(\varphi)(x) \approx \sum_{k+0}^{\infty} \frac{t^k}{k!} L_X^k(\varphi)(x), \tag{2.6a}$$

$$\mathrm{Ad}_X^t(Y)(x) \approx \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathrm{ad}_X^k(Y)(x), \tag{2.6b}$$

$$\mathrm{Ad}_X^t(\omega)(x) \approx \sum_{k=0}^{\infty} \frac{t^k}{k!} L_X^k(\omega)(x), \tag{2.6c}$$

where $\mathrm{ad}_X^k(Y) = [X, \mathrm{ad}_X^{k-1}(Y)]$.

Further identities are

$$\mathrm{Ad}_X^t(\langle \omega, Y\rangle)(x) = \langle \mathrm{Ad}_X^t(\omega), \mathrm{Ad}_X^t(Y)\rangle(x), \tag{2.7a}$$

$$L_X\langle \omega, Y\rangle(x) = \langle L_X(\omega), Y\rangle(x) + \langle \omega, L_X(Y)\rangle(x), \tag{2.7b}$$

$$\mathrm{Ad}_X^t([Y, Z])(x) = [\mathrm{Ad}_X^t(Y), \mathrm{Ad}_X^t(Z)](x), \tag{2.8a}$$

$$[X[Y, Z]](x) = [[X, Y]Z](x) + [Y[X, Z]](x) \text{ (Jacobi identity)}, \tag{2.8b}$$

and

$$\mathrm{Ad}_X^t(d\varphi)(x) = d(\mathrm{Ad}_X^t(\varphi))(x), \tag{2.9a}$$

$$L_X(d\varphi)(x) = d(L_X(\varphi))(x). \tag{2.9b}$$

A fundamental geometric concept in the study of nonlinear systems is the following.

DEFINITION. A distribution $\mathscr{D}$ is a submodule of $\mathscr{X}(M)$. We denote by $D(x)$ the subspace of T_xM obtained by evaluating the elements of $\mathscr{D}$ at x. The union $D = \bigcup_{x\in M} D(x)$ of these subspaces is called the *singular subbundle* of TM associated to $\mathscr{D}$. (By definition all singular subbundles of TM are associated to distributions.) If D (or $\mathscr{D}$) is nonsingular, i.e., the dimension of $D(x)$ is constant over all x, then D is a *subbundle* of TM (in the usual sense of the term).

A *local frame* for $\mathscr{D}$ (or D) on an open set $\mathscr{U} \subset M$ is a family of vector fields $\{X^1, \cdots, X^d\}$ such that for each $x \in \mathscr{U}$ the vectors $\{X^1(x), \cdots, X^d(x)\}$ are a basis for $D(x)$ (clearly D is nonsingular iff around each $x \in M$ it admits a local frame). Given a singular subbundle D associated to a distribution $\mathscr{D}$, we can define a second distribution $\Gamma(D)$ as the set of all smooth vector fields $X \in \mathscr{X}(M)$ such that $X(x) \in D(x)$, $\forall x \in M$. A distribution $\mathscr{D}$ is *complete* if $\mathscr{D} = \Gamma(D)$. (After this section all distributions will be assumed to be complete, and we shall use the term distribution to mean complete distribution.)

For nonsingular D, the distinction between D and $\mathscr{D}$ (or $\Gamma(D)$) is not particularly important and a certain amount of sloppiness is tolerable. However, one must be much more careful when considering the singular case. For example, the collection of all distributions on M forms a lattice partially ordered by inclusion under the operations of submodule addition and submodule intersection. If $\mathscr{D}^1$, $\mathscr{D}^2$ are complete distributions, D^1, D^2 their associated singular subbundles and D the singular subbundle associated to $\mathscr{D}^1 \cap \mathscr{D}^2$ then

$$D(x) \subset (D^1(x) \cap D^2(x))$$

but the inclusion can be proper for some x. For example, for $M = \mathbb{R}^2$ let $\mathscr{D}^1$ be the span of $\partial/\partial x_1$ and $\mathscr{D}^2$ be the span of $\partial/\partial x_1 + x_2 \partial/\partial x_2$ (span always means over $\mathscr{F}(M)$). Then $\mathscr{D}^1 \cap \mathscr{D}^2$ contains only the zero vector field so $D(x) = \{0\}$.

No such difficulty occurs with sums; if D is the singular subbundle associated to $\mathscr{D}^1 + \mathscr{D}^2$ then $D(x) = D^1(x) + D^2(x)$.

DEFINITION. An integral submanifold L of $\mathscr{D}$ is a connected, immersed submanifold $L \subset M$ such that for each $x \in L$, $T_x L = D(x)$. A distribution $\mathscr{D}$ is *integrable* if its maximal integral manifolds define a partition of M. This partition is called a *foliation* and the maximal integral submanifolds are its *leaves.*

DEFINITION. A distribution $\mathscr{D}$ is Ad_X *invariant* if $Y \in \mathscr{D}$ implies $\mathrm{Ad}_X^t(Y) \in \mathscr{D}$ for all t. A distribution $\mathscr{D}$ is ad_X *invariant* if $Y \in \mathscr{D}$ implies $\mathrm{ad}_X(Y) \in \mathscr{D}$.

Clearly from (2.5b), Ad_X invariance implies ad_X invariance but the converse need not hold. If everything is $\mathscr{C}^\omega$ then Lie series arguments (2.6b) imply the converse. If $\mathscr{D}$ is nonsingular then an argument of Hermann [16] also implies the converse.

DEFINITION. A distribution $\mathscr{D}$ is *involutive* if $\mathscr{D}$ is ad_X invariant for every $X \in \mathscr{D}$.

The basic integrability result is next.

THEOREM 2.1 (Sussmann [10]). *A distribution $\mathscr{D}$ is integrable iff $\mathscr{D}$ is Ad_X invariant for every $X \in \mathscr{D}$.*

This leads to the following corollaries.

COROLLARY 2.2 (Frobenius [18]). *For nonsingular distributions integrability and involutiveness are equivalent.*

COROLLARY 2.3. (Hermann [16], Nagano [17]). *For C^ω distributions integrability and involutiveness are equivalent.*

DEFINITION. A point x^0 is a *regular point* of the distribution $\mathscr{D}$ if the dimension of $D(x)$ is constant in a neighborhood of x^0 otherwise it is a *singular point.*

It is easy to see that the regular points of $\mathscr{D}$ form an open and dense submanifold of M.

COROLLARY 2.4. *An integrable distribution $\mathscr{D}$ is involutive. An involutive distribution $\mathscr{D}$ restricted to the submanifold of its regular points is integrable.*

The Ad_X and ad_X invariant distributions form lattices, while the integrable and involutive distributions form semilattices (closed under intersections but not sums). There exist minimal integrable and involutive distributions containing a given distribution $\mathscr{D}$, called the integrable and involutive closures of $\mathscr{D}$. From (2.8a,b) it follows that if $\mathscr{D}$ is Ad_X or ad_X invariant then so is its involutive closure.

DEFINITION. A *codistribution* $\mathscr{E}$ is a submodule of $\mathscr{X}^*(M)$. (Classically codistributions are called Pfaffian systems.) Associated to each codistribution $\mathscr{E}$ is a family of subspaces $E(x) \subset T_x^* M$ obtained by evaluating the one forms of $\mathscr{E}$ at x. The union $E = \cup E(x)$ is a *singular subbundle* of $T_x^* M$. Nonsingularity, local frame, completeness, etc. are all defined analogously.

There is a duality between distribution and codistributions. To each distribution $\mathscr{D}$ (codistribution $\mathscr{E}$) there is a codistribution $\mathscr{D}^\perp$ (distribution $\mathscr{E}^\perp$) called its annihilator

defined by

$$\mathscr{D}^{\perp} = \{\omega \in \mathscr{X}^*(M) : \langle \omega, X \rangle = 0, \forall X \in \mathscr{D}\},$$

$$(\mathscr{E}^{\perp} = \{X \in \mathscr{X}(M) : \langle \omega, X \rangle = 0, \forall \omega \in \mathscr{E}\}).$$

One has the inclusion

$$\mathscr{D} \subset \mathscr{D}^{\perp\perp} \qquad (\mathscr{E} \subset \mathscr{E}^{\perp\perp}),$$

which may be proper unless $\mathscr{D}(\mathscr{E})$ is nonsingular and complete. Moreover

$$(\mathscr{D}^1 + \mathscr{D}^2)^{\perp} = \mathscr{D}^{1\perp} \cap \mathscr{D}^{2\perp} \qquad ((\mathscr{E}^1 + \mathscr{E}^2)^{\perp} = \mathscr{E}^{1\perp} \cup \mathscr{E}^{2\perp}),$$

$$\mathscr{D}^{1\perp} + \mathscr{D}^{2\perp} \subset (\mathscr{D}^1 \cap \mathscr{D}^2)^{\perp} \qquad (\mathscr{E}^{1\perp} + \mathscr{E}^{2\perp} \subset (\mathscr{E}^1 \cap \mathscr{E}^2)^{\perp}).$$

If $\mathscr{D}^1$, $\mathscr{D}^2$ and $\mathscr{D}^1 \cap \mathscr{D}^2$ ($\mathscr{E}^1$, $\mathscr{E}^2$ and $\mathscr{E}^1 \cap \mathscr{E}^2$) are complete and nonsingular then the last inclusion is an identity.

DEFINITION. A codistribution $\mathscr{E}$ is Ad_X (ad_X) invariant if $\omega \in \mathscr{E}$ implies $\mathrm{Ad}_X^t(\omega) \in \mathscr{E}$ $\forall t$ ($\mathrm{ad}_X(\omega) \in \mathscr{E}$).

LEMMA 2.5. *If the distribution $\mathscr{D}$ is $\mathrm{Ad}_X(\mathrm{ad}_X)$ invariant then the codistribution $\mathscr{D}^{\perp}$ is also. If the codistribution $\mathscr{E}$ is Ad_X (ad_X) invariant then the distribution $\mathscr{E}^{\perp}$ is also.*

Proof. Suppose $\mathscr{D}$ is Ad_X invariant, $\omega \in \mathscr{D}^{\perp}$ and $Y \in \mathscr{D}$; then $\mathrm{Ad}_X^{-t}(Y) \in \mathscr{D}$. Using (2.7a) gives

$$\langle \mathrm{Ad}_X^t(\omega), Y \rangle = \langle \mathrm{Ad}_X^t(\omega), \mathrm{Ad}_X^t(\mathrm{Ad}_X^{-t}(Y)) \rangle = \mathrm{Ad}_X^t(\langle \omega, \mathrm{Ad}_X^{-t}(Y) \rangle) = 0,$$

so $\mathrm{Ad}_X^t(\omega) \in \mathscr{D}^{\perp}$ and $\mathscr{D}^{\perp}$ is Ad_X invariant. Suppose $\mathscr{D}$ is ad_X invariant, $\omega \in \mathscr{D}^{\perp}$ and $Y \in \mathscr{D}$; then $L_X(Y) \in \mathscr{D}$. Using (2.7b)

$$\langle L_X(\omega), Y \rangle = L_X \langle \omega, Y \rangle - \langle \omega, L_X(Y) \rangle = L_X(0) - 0 = 0$$

so $L_X(\omega) \in \mathscr{D}^{\perp}$ and $\mathscr{D}^{\perp}$ is ad_X invariant. The other assertion is proved similarly. QED

DEFINITION. A codistribution $\mathscr{E}$ is *integrable* if the distribution $\mathscr{E}^{\perp}$ is integrable.

Let $h: M \to \mathbb{R}^p$ be smooth. We denote by $\mathscr{R}(dh)$ the codistribution spanned by the one forms dh_i, $i = 1, \cdots, p$. We denote by $\mathscr{H}(dh)$ the distribution which annihilates $\mathscr{R}(dh)$, $\mathscr{H}(dh) = \mathscr{R}(dh)^{\perp}$.

LEMMA 2.6. *$\mathscr{R}(dh)$ is integrable.*

Proof. By definition we must show that the distribution $\mathscr{H}(dh)$ is integrable. By Sussmann's theorem this amounts to showing that $\mathscr{H}(dh)$ is Ad_X invariant for every $X \in \mathscr{H}(dh)$. By Lemma 2.6 this is equivalent to showing that $\mathscr{R}(dh)$ is Ad_X invariant for every $X \in \mathscr{H}(dh)$.

From the definition

$$\mathrm{Ad}_X^{s+t}(dh_i) = \mathrm{Ad}_X^s(\mathrm{Ad}_X^t(dh_i))$$

so

$$\frac{d}{ds}\mathrm{Ad}_X^s(dh_i) = \frac{d}{dt}\bigg|_{t=0} \mathrm{Ad}_X^{s+t}(dh_i) = \mathrm{Ad}_X^s \frac{d}{dt}\bigg|_{t=0} \mathrm{Ad}_X^t(dh_i).$$

By (2.5c) and (2.9b) this becomes

$$\frac{d}{ds}\mathrm{Ad}_X^s(dh_i) = \mathrm{Ad}_X^s L_X(dh_i) = \mathrm{Ad}_X^s d(L_X(h_i)).$$

But $X \in \mathscr{H}(dh)$ implies $L_X(h_i) = 0$ hence

$$\mathrm{Ad}_X^s(dh_i) = dh_i.$$

QED

3. Ad_f and ad_f invariance. In the study of linear systems of the form

$$\dot{x} = Ax + Bu, \tag{3.1a}$$

$$y = Cx, \tag{3.1b}$$

$$x(0) = x^0, \tag{3.1c}$$

the invariant subspaces of the matrix A play an important role. Suppose $V \subseteq \mathbb{R}^n$ is such a subspace, i.e., $AV \subseteq V$. Then V is spanned by the real and imaginary parts of a subset of eigenvectors and generalized eigenvectors of A. The invariant subspaces are the modal subspaces of A.

The nonlinear generalizations of this are several.

DEFINITION. A distribution or codistribution is Ad_f invariant (ad_f invariant) if it is $\mathrm{Ad}_{f(\cdot,u)}$ invariant ($\mathrm{ad}_{f(\cdot,u)}$ invariant) for each constant control $u \in \mathbb{R}^n$.

Clearly Ad_f invariance implies ad_f invariance but not the converse unless the distribution or codistribution is nonsingular or $\mathscr{C}^\omega$. It is easy to see that ad_f invariance is equivalent to ad_{g^j} invariance for $j = 0, \cdots, m$. What is not so obvious, but follows from Lemmas 3.2 and 3.3, is that Ad_f invariance is equivalent to Ad_{g^j} invariance for $j = 0, \cdots, m$. As one expects from the results of § 2, the sum and intersection of Ad_f or ad_f invariant (co) distributions is also, the involutive closure of an Ad_f or ad_f invariant distribution is also and the annihilator of an Ad_f or ad_f invariant (co) distribution is also.

Before we go any further let us relate these concepts to that of an invariant subspace of a linear system (3.1). Let V be an invariant subspace, and define $\mathscr{D}$ as the set of vector fields on $\mathbb{R}^n$ which take values in V. (We are using the canonical identification of $\mathbb{R}^n$ with each of its tangent spaces $T_x\mathbb{R}^n$.) The associated subbundle D is nonsingular with $D(x) = V$ (thought of as a subspace of $T_x\mathbb{R}^n$). For each constant control $u \in \mathbb{R}^m$ we obtain the vector field $f(x, u) = Ax + Bu$ and corresponding flow $\Phi(t, x) = e^{At}(x + \int_0^t e^{-As}\, Bu\, ds)$.

We claim that $\mathscr{D}$ is Ad_f and ad_f invariant. By the above remarks it suffices to verify that $\mathscr{D}$ is Ad_{g^j} and ad_{g^j} invariant for $j = 0, \cdots, m$. But $g^j(x) = B^j$ (the jth column of B), a constant vector field, and any basis for V considered as constant vector fields defines a global frame for $\mathscr{D}$. Let $v \in V$ considered as a constant vector field in $\mathscr{D}$ then

$$\mathrm{ad}_{g^0}(v) = -Av, \qquad \mathrm{Ad}^t_{g^0}(v) = e^{-At}v, \tag{3.2a}$$

$$\mathrm{ad}_{g^j}(v) = 0, \qquad \mathrm{Ad}^t_{g^0}(v) = v, \quad j = 1, \cdots, m. \tag{3.2b}$$

Since a frame for $\mathscr{D}$ is invariant, it follows that all of $\mathscr{D}$ is.

We refer to such a $\mathscr{D}$ as a constant distribution on $\mathbb{R}^n$ because it has a global frame of constant vector fields but of course $\mathscr{D}$ contains nonconstant vector fields. If g^j is a constant vector field (such as B^j) and $\mathscr{D}$ is a constant distribution then $\mathscr{D}$ is always Ad_{g^j} and ad_{g^j} invariant. Therefore one need only check the Ad_{g^0} and ad_{g^0} invariance of constant distributions. This fact frequently leads to differences between the formulation of a linear result and its nonlinear generalization as we shall see throughout this paper.

We have just noted that for a linear system the constant distributions which are Ad_f or ad_f invariant are precisely the invariant subspaces of A. One might ask whether there are any nonconstant distributions which are invariant. If one restricts to nonsingular distributions the answer is essentially no.

PROPOSITION 3.1. *Suppose the linear system* (3.1) *is controllable and $\mathscr{D}$ is a nonsingular* Ad_f (*equivalently* ad_f) *invariant distribution for* (3.1). *Then $\mathscr{D}$ is a constant distribution, hence corresponds to an invariant subspace of A.*

Proof. Let the dimension of $\mathscr{D}$ be d and let $X^1, \cdots, X^d$ be a local frame. By assumption for $j = 0, \cdots, m$ and $k = 1, \cdots, d$

$$[g^j, X^k] \in \mathscr{D}.$$

Using the Jacobi identity (2.7b)

$$[[g^i, g^j], X^k] = [g^i[g^j, X^k]] - [g^j[g^i, X^k]] \in \mathscr{D},$$

so $\mathscr{D}$ is invariant under any bracket $[g^i, g^j]$. By repeating this argument it follows that $\mathscr{D}$ is invariant under any multiple bracket $[g^{j_1} \cdots [g^{j_{r-1}}, g^{j_r}] \cdots]$.

Now $g^0 = Ax$, $g^j = B^j$ (jth column of B) and

$$\mathrm{ad}^r_{g^0}(g^j) = (-1)^r A^r B^j,$$

$$[g^i, \mathrm{ad}^r_{g^0}(g^j)] = 0,$$

where $i, j = 1, \cdots, m$ and $r \geqq 0$. The controllability assumption implies that there are n linearly independent vectors of the form $A^r B^j$. View these as constant vector fields and denote them by $Y^1, \cdots, Y^n$.

Since each Y^k is a bracket of g^j's, it leaves $\mathscr{D}$ invariant, hence there exist functions Γ_i^{kj} such that

$$[Y^k, X^j] = \sum_{i=1}^{d} X^i \Gamma_i^{kj}.$$

Let Γ^k denote the $d \times d$ matrix (Γ_i^{kj}) and X the $n \times d$ matrix $(X^1 \cdots X^d)$; we abbreviate the above as

$$[Y^k, X] = X\Gamma^k.$$

We make a change of local frame for $\mathscr{D}$ by choosing a $d \times d$ invertible matrix valued function γ, the new basis is the set of columns $\tilde{X}^1, \cdots, \tilde{X}^d$ of $\tilde{X} = X_\gamma$. We seek a basis which commutes with Y^k, i.e.

$$0 = [Y^k, \tilde{X}] = [Y^k, X\gamma] = [Y^k, X]\gamma + XL_{Y^k}(\gamma) = X(\Gamma^k\gamma + L_{Y^k}(\gamma)).$$

Hence γ should satisfy the linear partial differential equation

$$L_{Y^k}(\gamma) = -\Gamma^k\gamma.$$

There is a local solution to this equation if the integrability (mixed partial) conditions are satisfied. Since $[Y^k, Y^l] = 0$ these are

$$L_{Y^k}L_{Y^l}(\gamma) = L_{Y^l}L_{Y^k}(\gamma)$$

which reduce to

$$L_{Y^k}(\Gamma^l\gamma) = L_{Y^l}(\Gamma^k\gamma)$$

or

$$(L_{Y^k}(\Gamma^l) - \Gamma^l\Gamma^k)\gamma = (L_{Y^l}(\Gamma^k) - \Gamma^k\Gamma^l)\gamma.$$

But these follow from the Jacobi identity (2.8b) and the linear independence of the columns of X for

$$[Y^k[Y^l, X]] - [Y^l[Y^k, X]] = [[Y^k, Y^l]X] = 0,$$

$$[Y^k, X\Gamma^l] - [Y^l, X\Gamma^k] = 0,$$

$$X(\Gamma^k\Gamma^l + L_{Y^k}(\Gamma^l) - \Gamma^l\Gamma^k - L_{Y^l}(\Gamma^k)) = 0.$$

Hence we can find γ such that the vector fields $\tilde{X}^1, \cdots, \tilde{X}^d$ of the new local frame for $\mathscr{D}$ commute with the constant vector fields $Y^1, \cdots, Y^n$ which span $\mathbb{R}^n$. From this one can conclude that $\tilde{X}^1, \cdots, \tilde{X}^d$ are constant vector fields so locally $\mathscr{D}$ has a constant frame. On the common domain of definition of two such constant frames, the change of frame matrix must be constant so any such constant local frame extends to a constant global frame for $\mathscr{D}$. QED

The statement that $AV \subset V$ can be interpreted as the dynamics (3.1a) infinitesimally leaves the directions of V invariant. The statement that $e^{At}V \subset V$ can be interpreted as the flow of (3.1a) leaves the directions of V invariant. Both these statements have direct nonlinear generalizations. If $\mathscr{D}$ is ad_f invariant then the dynamics (2.1a) infinitesimally leaves the directions of $\mathscr{D}$ invariant. If $\mathscr{D}$ is Ad_f invariant then the flow of (2.1a) leaves the directions of $\mathscr{D}$ invariant.

The constant distribution $\mathscr{D}$ on $\mathbb{R}^n$ associated to any subspace V of $\mathbb{R}^n$ is integrable, the leaves of the foliation that it induces are the cosets $x+V$ for $x \in \mathbb{R}^n$. If V is an invariant subspace of A then the flow of (3.1a) for any fixed control $u(t)$ carries cosets into cosets. A concrete way of seeing this is to choose local coordinates $x = \binom{x_1}{x_2}$ such that

$$V = \{x: x_1 = 0\}.$$

In these coordinates the dynamics (3.1a) takes a block triangular form.

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} u. \tag{3.3}$$

The coset space $\mathbb{R}^n / V$ is coordinatized by x_1 and since x_1 evolves independently of x_2, the dynamics passes to this space.

In the nonlinear context a similar thing happens. Suppose $\mathscr{D}$ is a nonsingular, involutive Ad_f (equivalently ad_f) invariant distribution. Then locally one can choose coordinates $x = \binom{x_1}{x_2}$ so that the leaves of the foliation induced by $\mathscr{D}$ are given by $x_1 = \text{constant}$. In these coordinates the dynamics again assumes a triangular form

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, u) = g_1^0(x_1) + g_1(x_1)u, \\ \dot{x}_2 &= f_2(x_1, x_2, u) = g_2^0(x_1, x_2) + g_2(x_1, x_2)u, \end{aligned} \tag{3.4}$$

and the flow for any fixed control $u(t)$ carries leaves into leaves. If the foliation induced by $\mathscr{D}$ is *regular*, i.e., the space of leaves can be given a manifold structure, then this space is locally coordinatized by x_1 and the dynamics passes to it. See [1] for details. We close with some technical results regarding Ad_X invariance which we referred to in the beginning of this section and which will be used later on.

LEMMA 3.2. *Suppose $\mathscr{D}$ is an Ad_X invariant distribution and $c \in \mathbb{R}$; then $\mathscr{D}$ is $\mathrm{Ad}_{(cX)}$ invariant. Suppose $\mathscr{D}$ is an Ad_X invariant distribution, $X \in \mathscr{D}$ and $\varphi \in \mathscr{F}(M)$; then $\mathscr{D}$ is $\mathrm{Ad}_{(\varphi X)}$ invariant.*

Proof. The first statement follows immediately from the identity $\mathrm{Ad}_{(cX)}^t = \mathrm{Ad}_X^{ct}$. As for the second let $\tau(t, x)$ be the solution of

$$\frac{\partial}{\partial t}\tau(t, x) = \varphi(\Phi(\tau(t, x), x),$$

$$\tau(0, x) = 0,$$

where $\Phi(t, x)$ is the flow of X. Define $\Psi(t, x) = \Phi(\tau(t, x), x)$; then $\Psi(t, x)$ is the flow of φX for

$$\begin{aligned}\frac{\partial}{\partial t}\Psi(t, x) &= \frac{\partial \Phi}{\partial \tau}(\tau(t, x), x)\frac{\partial \tau}{\partial t}(\tau, x)\\ &= X(\Phi(\tau(t, x), x))\varphi(\Phi(\tau(t, x), x))\\ &= X(\Psi(t, x))\varphi(\Psi(t, x))\end{aligned}$$

and

$$\Psi(0, x) = \Phi(\tau(0, x), x) = \Phi(0, x) = x.$$

Now if $Y \in \mathscr{D}$ then by (2.3a)

$$\begin{aligned}\mathrm{Ad}^t_{(\varphi X)}(Y)(x) &= \frac{\partial \Psi}{\partial z}(-t, z)\, Y(\Psi(t, x))\\ &= \frac{\partial \Phi}{\partial z}(-\tau(t, z), z)\, Y(z)\\ &= \mathrm{Ad}^\tau_X(Y)(x) - \frac{\partial \Phi}{\partial \tau}(-\tau, z)\frac{\partial \tau}{\partial t}(t, z)\, Y(z)\end{aligned}$$

where $z = \Psi(t, x) = \Phi(\tau(t, x), x)$ and $\tau = \tau(t, x)$.

Since $\partial\Phi/\partial\tau(-\tau, z) = X(x) \in D(x)$ it follows that $\mathrm{Ad}^t_{(\varphi X)} Y \in \mathscr{D}$. QED

LEMMA 3.3. *Suppose $\mathscr{D}$ is Ad_{X^i} invariant for $i = 1, 2$. Then $\mathscr{D}$ is $\mathrm{Ad}_{(X^1+X^2)}$ invariant.*

Proof. Let $u(t) = (u_1(t), u_2(t))$ be a bounded measurable function. Let $\Phi_u(t, t_0, x^0)$ be the time dependent flow of the time dependent vector field $X_u(t, x) = X^1(x)u_1(t) + X^2(x)u_2(t)$, i.e.

$$\frac{d}{dt}\Phi_u(t, t_0, x^0) = X_u(t, \Phi_u(t, t_0, x^0)), \tag{3.5a}$$

$$\Phi_u(t_0, t_0, x^0) = x^0. \tag{3.5b}$$

By standard results from differential equations, for each t_1, the map $x_0 \to \Phi_u(t_1, t_0, x^0)$ is a local diffeomorphism. If $u^k(\cdot)$ converges to $u(\cdot)$ in the weak L^∞ topology on $[t_0, t_1]$ then $\Phi_{u^k}(t_1, t_0, \circ)$ converges to $\Phi_u(t_1, t_0, \circ)$ uniformly as small compact subsets. Moreover each of the derivatives does also.

$\mathscr{D}$ is Ad_{X^i} invariant iff the flow Φ^i of X^i carries the vector field of $\mathscr{D}$ back into $\mathscr{D}$, i.e. if $Y \in \mathscr{D}$ then $\Phi^i(t)_* Y \in \mathscr{D}$. Let $u^k(\cdot)$ be equal to $(2, 0)$ and $(0, 2)$ on intervals of length $1/k$; then u^k converges weakly to $u(t) = (1, 1)$. By assumption $\Phi_{u^k}(t_1, t_0)_* Y \in \mathscr{D}$ for all t_0, t_1 and $Y \in \mathscr{D}$. By continuity $\Phi_u(t_1, t_0)_* Y \in \mathscr{D}$, hence $\mathscr{D}$ is $\mathrm{Ad}_{(X^1+X^2)}$ invariant.

Remark. One could define Ad and ad invariance with respect to time dependent vector fields such as $X_u(t, x)$. By modifying the proofs of the above lemmas one can show that $\mathscr{D}$ is Ad_{X_u} (or ad_{X_u}) invariant for any bounded measurable $u(t)$ iff $\mathscr{D}$ is Ad_{X^i} (or ad_{X^i}) invariant for all i.

LEMMA 3.4. *Suppose $\mathscr{D}$ is Ad_{X^i} invariant for $i = 1, 2$. Then $\mathscr{D}$ is Ad_Z invariant where $Z = \mathrm{Ad}^\tau_{X^1}(X^2)$ for any τ.*

Proof. Let $\Phi^i(t, x)$ denote the flow of X^i and define

$$\Psi(t, x) = \Phi^1(-\tau, \Phi^2(t. \Phi^1(\tau, x))). \tag{3.6}$$

Since

$$\frac{\partial \Psi}{\partial t}(t, x) = \Phi^1(-\tau)_* X^2(\Phi^2(t, \Phi^1(t, x)))$$
$$= \Phi^1(-\tau)_* X^2(\Phi^1(\tau, \Psi(t, x)))$$
$$= Z(\Psi(t, x)),$$
$$\Psi(0, x) = x,$$

it follows that Ψ is the flow of Z. But then

$$\Psi(t)_* Y = \Phi^1(-\tau)_* \Phi^2(\tau)_* \Phi^1(\tau)_* Y,$$

hence $Y \in \mathscr{D}$ implies $\Psi(\tau)_* Y \in \mathscr{D}$. QED

4. Nonlinear controllability and observability. In this section we review the basic concepts of nonlinear controllability and observability because they are needed in the study of disturbance decoupling and noninteracting control and they are nowhere treated in an appropriate form. The closest reference is our joint work with Hermann [12] but we must apologize for the somewhat confusing terminology that we introduced there. We hope this section rectifies the situation.

The main difficulty in passing from linear to nonlinear is that typically there are several reasonable nonlinear generalizations of a single linear concept. The appropriate choice depends on the context.

Let $\mathscr{U}$ be an open connected subset of M and T a nonnegative real number.

DEFINITION. A point x^T is $\mathscr{U}$ *accessible* from x^0 at time T if there exists a bounded measurable control $u(t)$ generating a trajectory of (2.1) $x(t) \in \mathscr{U}$ for $t \in [0, T]$ such that $x(0) = x^0$ and $x(T) = x^T$. The set of all sets x^T, $\mathscr{U}$ accessible from x^0 at time T, is denoted by $\mathscr{A}(x^0, T, \mathscr{U})$. If $\mathscr{U}$ is suppressed, M is to be understood as in $\mathscr{A}(x^0, T) = \mathscr{A}(x_0, T, \mathscr{U})$. If T is suppressed, the union over all $T \geqq 0$ is understood as in $\mathscr{A}(x^0, \mathscr{U}) = \bigcup_{T>0} \mathscr{A}(x^0, T, \mathscr{U})$.

DEFINITION. The system (2.1) is *controllable* if $\mathscr{A}(x^0) = M$ for every $x^0 \in M$. The system (2.1) is *locally controllable* if restricted to every open connected subset $\mathscr{U}$ of M, (2.1) is controllable, i.e., $\mathscr{A}(x_0, \mathscr{U}) = \mathscr{U}$ for every $x^0 \in \mathscr{U} \subset M$.

It is apparent that local controllability implies controllability but not vice versa. We shall use the modifiers local and locally to mean that a property holds for (2.1) restricted to every open connected subset of the state space and hence a local property always implies that property. These definitions capture our intuitive idea of controllability and local controllability but unfortunately they are extremely difficult to work with. Deciding when a nonlinear system is controllable or locally controllable is generally a difficult task. We are more interested in controllability as one half of what constitutes a minimal realization, therefore we introduce weaker notions. The *time reversible version* of (2.1) is

$$\dot{x} = f(x, u_0, u) = g^0(x)u_0 + g(x)u, \tag{4.1a}$$
$$y = h(x), \tag{4.1b}$$
$$x(0) = x^0. \tag{4.1c}$$

DEFINITION. The system (2.1) is *reversibly controllable* if (4.1) is controllable. The system (2.1) is *locally reversibly controllable* if (4.1) locally controllable. Let $\mathscr{RA}(x_0, T, \mathscr{U})$ be the set of points accessible in $\mathscr{U}$ from x^0 along trajectories of (4.1). Equivalently the system (2.1) is (locally) reversibly controllable if for every x_0 (and $\mathscr{U}$), $\mathscr{RA}(x^0) = M(\mathscr{RA}(x^0, \mathscr{U}) = \mathscr{U})$.

Clearly (local) controllability implies (local) reversible controllability but not vice versa. Throughout we use the modifiers reversible and reversibly to mean that a property holds not for (2.1) itself but for its time reversible version (4.1), hence a property generally implies the corresponding reversible property.

These definitions emphasize one aspect of what one expects in a controllable system, the ability to steer from one point to another; but there is another: namely, that there are no uncontrollable modes, no coordinates of the state space which are unaffected by the control. The following are attempts to characterize this.

DEFINITION. The system (2.1) has the (*local*) *accessibility property* if $\mathscr{A}(x^0)$ ($\mathscr{A}(x^0, \mathscr{U})$) has nonempty interior for every $x^0 \in M$ (and open neighborhood $\mathscr{U}$). The system (2.1) has the (*local*) *reversible accessibility property* if (4.1) has the (local) accessibility property.

THEOREM 4.1. *If the system* (2.1) *has the accessibility property then it is reversibly controllable. The system has the* (*local*) *reversible accessibility property iff it is* (*locally*) *reversibly controllable. The system has the local accessibility property iff it is reversibly controllable.*

Proof. Reversible accessibility is an equivalence relation which partitions M. Suppose the system has the accessibility property so that $\mathscr{A}(x^0)$ has nonempty interior. This implies that $\mathscr{R}\mathscr{A}(x^0)$ (the set of points reversibly accessible from x^0) is an open subset of the connected manifold M, hence $\mathscr{R}\mathscr{A}(x^0) = M$. The proof of the second assertion is straightforward and the proof of the third is found in Hermann-Krener [12, Thm. 2.1]. QED

In summary the logical implications between various forms of controllability are

$$\begin{array}{lcl}
\text{local controllability} & \Rightarrow & \text{controllability} \\
\Downarrow & & \Downarrow \\
\text{local accessibility property} & \Rightarrow & \text{accessibility property} \\
\Updownarrow & & \Downarrow \\
\text{local reversible controllability} & \Rightarrow & \text{reversible controllability} \\
\Updownarrow & & \Updownarrow \\
\text{local reversible accessibility property} & \Rightarrow & \text{reversible accessibility property.}
\end{array}$$

One would like a simple criterion to decide when a system is controllable or not. Unfortunately none seems to exist. These are however relatively straightforward criteria for some of the others. We denote by $\mathscr{R}(f)$ the distribution spanned (over $\mathscr{F}(M)$) by $\{f(\cdot, u)\colon u \text{ constant}\}$. Let $\langle \mathrm{Ad}_f | \mathscr{R}(f) \rangle$ and $\langle \mathrm{ad}_f | \mathscr{R}(f) \rangle$ denote the smallest Ad_f and ad_f invariant distributions containing $\mathscr{R}(f)$. These are the Ad_f and ad_f *controllability distributions.*

By Lemmas 3.2 and 3.3 the former is spanned by terms of the form

$$\mathrm{Ad}_{f^k}^{t_k} \circ \cdots \circ \mathrm{Ad}_{f^1}^{t_1} f^0 \tag{4.2a}$$

where $k \geqq 0$ and $f^j(x) = f(x, u^j)$ for u^j constant. The latter is spanned by terms of the form

$$\mathrm{ad}_{f^k} \circ \cdots \circ \mathrm{ad}_{f^1} f^0 \tag{4.2b}$$

and by the Jacobi identity (2.8b) is involutive. Lemma 3.4 implies that the former is integrable and is the integral closure of the latter. The next result is related to a theorem of Chow [21].

THEOREM 4.2 (Sussmann [11]). *The system* (2.1) *is reversibly controllable iff*

$$\langle \mathrm{Ad}_f | \mathscr{R}(f) \rangle = \mathscr{X}(M). \tag{4.3}$$

While this is very elegant, the Ad_f controllability distribution is not always easy to compute so the following can be more useful.

THEOREM 4.3 (Hermann-Krener [12]). *The system* (2.1) *is locally reversibly controllable if*

$$\langle \mathrm{ad}_f | \mathscr{R}(f) \rangle = \mathscr{X}(M). \tag{4.4}$$

If (2.1) *is locally reversibly controllable and D is the subbundle of TM associated to the* ad_f *controllability distribution then on an open dense subset of M*

$$D(x) = T_x M. \tag{4.5}$$

Equation (4.5) is usually referred to as the *controllability rank condition* at x. For a linear system (3.1) the Ad_f and ad_f controllability distributions both equal

$$\mathscr{R}\{Ax, A^r B^j : r = 0, \cdots, n-1, j = 1, \cdots, m\}.$$

For $x = 0$ the controllability rank condition (4.5) reduces to the familiar

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

Now we turn to observability where again we follow [12] in spirit but change terminology considerably. In what follows $\mathscr{U}$ always denotes an open subset of M.

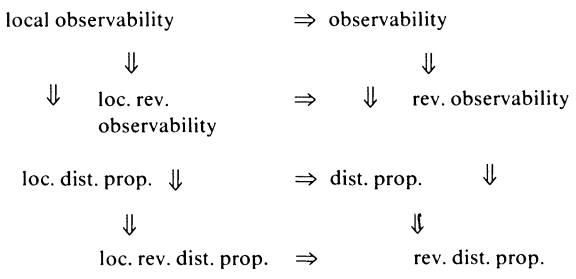
DEFINITION. Two points x^0 and x^1 are $\mathscr{U}$ *distinguishable* if there exists a bounded measurable input $u(t)$ generating solutions $x^0(t)$ and $x^1(t)$ of (2.1a) satisfying $x^i(0) = x^i$ such that $x^i(t) \in \mathscr{U}$ for all $t \in [0, T]$ and $h(x^1(t)) \neq h(x^2(t))$ for some $t \in [0, T]$. We let $\mathscr{I}(x^0, \mathscr{U})$ denote all the points $x^1 \in \mathscr{U}$ which are not $\mathscr{U}$ distinguishable from x^0. If $\mathscr{U}$ is suppressed, M is understood as in $\mathscr{I}(x^0) = \mathscr{I}(x^0, M)$.

DEFINITION. The system (2.1) is *observable* if $\mathscr{I}(x^0) = \{x^0\}$ for every x^0. The system (2.1) is *locally observable* if for every open neighborhood $\mathscr{U}$ of x^0, $\mathscr{I}(x^0, \mathscr{U}) = \{x^0\}$. The system (2.1) is (*locally*) *reversibly* observable if (4.1) is (locally) observable.

All these definitions require that x^0 be distinguishable from every other point of M. The local ones require that x^0 and x^1 be distinguishable by local experiments. Frequently it may suffice that one be able to distinguish a point from its neighbors either by local or global experiments. Therefore we introduce additional terminology which was referred to as (local) weak observability in [12].

DEFINITION. The system (2.1) has the *distinguishability property* if every x^0 has an open neighborhood $\mathscr{V}$ such that $\mathscr{I}(x^0) \cap \mathscr{V} = \{x^0\}$. The system (2.1) has the *local distinguishability property* if every x^0 has an open neighborhood $\mathscr{V}$ such that for every open $\mathscr{U}$ neighborhood of x^0, $\mathscr{I}(x^0, \mathscr{U}) \cap \mathscr{V} = \{x^0\}$. The system (2.1) has the (*local*) *reversible distinguishability property* if (4.1) has the (local) distinguishability property.

The basic implications between these definitions are as follows.



If one makes a controllability assumption more implications follow; perhaps the most interesting is

THEOREM 4.4. *If* (2.1) *is locally reversibly controllable then the local distinguishability property and the local reversible distinguishability property are equivalent.*

We defer the proof to the end of this section.

Let $\mathscr{R}(dh)$ denote the codistribution spanned by dh_i, $i=1,\cdots,p$ and let $\langle \mathrm{Ad}_f|\mathscr{R}(dh)\rangle$ and $\langle \mathrm{ad}_f|\mathscr{R}(dh)\rangle$ denote the smallest Ad_f and ad_f invariant codistributions containing $\mathscr{R}(dh)$. We refer to these as the Ad_f and ad_f *observability codistributions.* They are the spans (over $\mathscr{F}(M)$) of terms of the form

$$\mathrm{Ad}_{f^k}^{t_k}\circ\cdots\circ\mathrm{Ad}_{f^1}^{t_1}\,dh_i \tag{4.6a}$$

and

$$\mathrm{ad}_{f^k}\circ\cdots\circ\mathrm{ad}_{f^1}dh_i \tag{4.6b}$$

respectively. By (2.9a, b) the exterior differential operator d can be pulled to the front in (4.6) so that by Lemma 2.6 these codistributions are integrable.

THEOREM 4.5 (Goncalves [13]). *The system* (2.1) *has reversible distinguishability property iff*

$$\langle \mathrm{Ad}_f|\mathscr{R}(dh)\rangle = \mathscr{X}^*(M). \tag{4.7}$$

THEOREM 4.6 (Hermann-Krener [12]). *The system* (2.1) *has the local distinguishability property if*

$$\langle \mathrm{ad}_f|\mathscr{R}(dh)\rangle = \mathscr{X}^*(M). \tag{4.8}$$

If (2.1) *has the local distinguishability property and E is the subbundle of T^*M associated to the* ad_f *observability codistribution then on an open dense subset of M*

$$E(x) = T_x^*M. \tag{4.9}$$

Equation (4.9) is usually referred to as the *observability rank condition* of x. For a linear system (3.1) the Ad_f and ad_f observability codistributions are the $\mathscr{F}(M)$ span of the rows of the familiar observability matrix,

$$\begin{pmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{pmatrix}.$$

Proof of Theorem 4.4. Clearly the local distinguishability property implies the local reversible distinguishability property. To see the converse notice that the ad_f observability codistribution for (2.1) and (4.1) are the same. Hence by Theorem 4.6, the observability rank condition holds on some open dense subset $\mathscr{V}$ of M. Therefore (2.1) restricted to $\mathscr{V}$ has the local distinguishability property, every x^0 and open neighborhood $\mathscr{U}$ is such that $\mathscr{A}(x^0, \mathscr{U})$ meets $\mathscr{V}$. But this implies x^0 can be distinguished from its neighbors. QED

5. (Ad_f, g), (ad_f, g) and local (ad_f, g) invariance. In the geometric approach to linear multivariable systems, as found in Wonham [15], the concept of an (A, B) invariant subspace plays a crucial role. Recall a subspace $V\subset\mathbb{R}^n$ is (A, B) invariant if one of two equivalent conditions is satisfied,

$$AV\subset V+\mathscr{R}(B) \tag{5.1}$$

($\mathcal{R}(B)$ denotes the subspace spanned by the columns of B) or there exists an $m \times n$ matrix F such that

$$(A+BF)V \subset V. \tag{5.2}$$

For reasons that will become apparent later we refer to these as the local and global characterizations of (A, B) invariance. The global characterization (5.2) can be interpreted as modifying the dynamics (3.1a) by linear state feedback

$$u = Fx + v \tag{5.3}$$

so as to obtain the new system

$$\dot{x} = \tilde{A}x + Bv \tag{5.4}$$

where $\tilde{A} = A + BF$. The subspace V is an invariant subspace of the new dynamics.

When working with linear systems it is convenient to restrict oneself to constant distributions and linear feedback laws (5.3). We could allow a slightly more general form, say

$$u = Fx + Gv \tag{5.5}$$

but as far as (A, B) invariance is concerned it is not needed because every constant vector field leaves every constant distribution invariant. When dealing with (A, B) controllability subspaces, feedback laws such as (5.5) naturally arise.

A nonlinear feedback (or feedback) is a pair of matrix valued functions α and β on M; $\alpha(x)$ and $\beta(x)$ are $m \times 1$ and $m \times m$ matrices smoothly varying in x. They are used to define the feedback law

$$u = \alpha(x) + \beta(x)v \tag{5.6}$$

which results in the modified system

$$x = \tilde{f}(x, v) = \tilde{g}^0(x) + \tilde{g}(x)v \tag{5.7}$$

where $\tilde{g}^0(x) = g^0(x) + g(x)\alpha(x)$, $\tilde{g}(x) = g(x)\beta(x)$ and $\tilde{g}^i(x) = g(x)\beta^j(x)$ where $\beta^j(x)$ is the jth column of $\beta(x)$. It is convenient to combine these into an $(m+1) \times (m+1)$ matrix

$$\gamma = \begin{pmatrix} 1 & 0 \\ \alpha & \beta \end{pmatrix} \tag{5.8}$$

and reexpress this as

$$\tilde{f}(x) = f(x)\gamma(x) \tag{5.9}$$

where $f(x)$ and $\tilde{f}(x)$ are $n \times (m+1)$ matrices

$$f(x) = (g^0(x), g(x)), \qquad \tilde{f}(x) = (\tilde{g}^0(x), \tilde{g}(x)). \tag{5.10}$$

Hopefully this second use of the symbols f and $\tilde{f}$ will cause no confusion. If there is no drift term g^0, then the feedback γ reduces to β.

DEFINITION. A distribution $\mathcal{D}$ is (Ad_f, g) *invariant* ((ad_f, g) *invariant*) if there exists a feedback γ such that $\mathcal{D}$ is $\mathrm{Ad}\,\tilde{f}$ invariant ($\mathrm{ad}\,\tilde{f}$ invariant). If γ is invertible then the distribution is invariant *with full control* otherwise it is invariant *with partial control.*

Unless otherwise stated "invariance" means "invariance with full control". This issue does not arise in the linear theory, because for reasons mentioned above, invariance always means with full control.

DEFINITION. A distribution $\mathscr{D}$ is *locally* (ad_f, g) *invariant with full control* if and for every constant u and for every $X \in \mathscr{D}$

$$\mathrm{ad}_f(\cdot, u)(X) \in \mathscr{D} + \mathscr{R}(g)$$

where $\mathscr{R}(g)$ denotes the distribution spanned by the columns of g.

A distribution $\mathscr{D}$ is *locally* (ad_f, g) *invariant with partial control* if there exists a feedback γ which is not necessarily invertible such that $\mathscr{D}$ is locally $(\mathrm{ad}_{\tilde{f}}, \tilde{g})$ invariant where $\tilde{f} = f\gamma$ $\tilde{g} = g\beta$. Again unless otherwise stated "local (ad_f, g) invariance" assumes "with full control".

It is not hard to see that (Ad_f, g) invariance implies (ad_f, g) invariance which in turn implies local (ad_f, g) invariance. Before we delve further into this area we need some additional terminology.

DEFINITION. A family of distribution $\mathscr{D}^1, \cdots, \mathscr{D}^\mu$ separates the controls if there exists locally an invertible feedback γ where β has been partitioned into submatrices $\beta = (\beta^1, \cdots, \beta^{\mu+1})$ such that

$$D^\sigma(x) \cap G(x) = \tilde{G}^\sigma(x), \qquad \sigma = 1, \cdots, \mu, \tag{5.11a}$$

$$\left(\sum_{\sigma=1}^{\mu} D^\sigma(x) \right) \cap \tilde{G}^{\mu+1}(x) = \{0\}, \tag{5.11b}$$

where G, $\tilde{G}^\sigma$ and D^σ are the subbundles of TM associated to the distributions $\mathscr{R}(g)$, $\mathscr{R}(\tilde{g}^\sigma)$ and $\mathscr{D}^\sigma$ respectively. A family of distributions *completely separates* the controls if there exists an invertible feedback γ with $\beta = (\beta^1, \cdots, \beta^\mu)$ such that (5.11a) holds. Such feedbacks γ are said to be (*completely*) *separating* for the family of distributions $\mathscr{D}^1, \cdots, \mathscr{D}^\mu$. A distribution $\mathscr{D}$ *separates* the controls if considered as a one element family of distributions it separates.

Notice α does not play a role in these definitions.

LEMMA 5.1. *If $\mathscr{D}$ is nonsingular, involutive and locally separates the controls then the following are equivalent*:

(a) *$\mathscr{D}$ is locally (ad_f, g) invariant.*

(b) *There exist an open cover $\{\mathscr{U}^\rho\}$ of M and separating feedbacks γ^ρ defined on $\mathscr{U}^\rho$ such that $\mathscr{D}$ is $\mathrm{ad}\,\tilde{f}^\rho$ invariant on $\mathscr{U}^\rho$ where $f^\rho = \tilde{f}\gamma^\rho$ (in other words, locally $\mathscr{D}$ is (ad_f, g) invariant).*

(c) *There exist an open cover $\{\mathscr{U}^\rho\}$ of M and separating feedbacks γ^ρ defined on $\mathscr{U}^\rho$ such that $\mathscr{D}$ is $\mathrm{Ad}\,\tilde{f}^\rho$ invariant on $\mathscr{U}^\rho$ (in other words, locally $\mathscr{D}$ is (Ad_f, g) invariant).*

Proof. The equivalence of (b) and (c) follows from the nonsingularity of $\mathscr{D}$. It is trivial to verify that (b) implies (a). In [3] it is shown that (a) implies (b) using the stronger hypothesis that $\mathscr{D} \cap \mathscr{R}(g)$ and $\mathscr{R}(g)$ are nonsingular. But the proof only uses this to show that $\mathscr{D}$ separates the controls. Moreover the feedback so constructed is easily seen to be separating. Similar results are found in [4]. QED

This lemma explains our terminology, in particular why we refer to (5.1) and (5.2) as the local and global characterizations of (A, B) invariance. For a discussion of the topological obstructions to global invariance we refer the reader to [20].

LEMMA 5.2. *If $\mathscr{D}$ is (Ad_f, g) invariant or (ad_f, g) invariant then so is the involutive closure of $\mathscr{D}$. If $\mathscr{D}$ is locally (ad_f, g) invariant then so is the involutive closure of $\mathscr{D}$. If $\mathscr{D}^1$ and $\mathscr{D}^2$ are locally (ad_f, g) invariant then so is $\mathscr{D}^1 + \mathscr{D}^2$, hence the set of locally (ad_f, g) invariant distributions forms a semilattice under inclusion and addition.*

Proof. Since (Ad_f, g) or (ad_f, g) invariance is equivalent to $\mathrm{Ad}_{\tilde{f}}$ or $\mathrm{ad}_{\tilde{f}}$ invariance for some feedback modified dynamics $\tilde{f}$, the first statement follows from (2.7). The

second statement is proved in [3] and the third follows directly from the definition of local (ad_f, g) invariance. QED

Remarks. The sum and intersection of (Ad_f, g) (or (ad_f, g)) invariant distributions and the intersection of locally (ad_f, g) invariant distributions need not be invariant in the same sense. However the sum of locally (ad_f, g) invariant distributions is again locally (ad_f, g) invariant. This semilattice structure makes them convenient to work with. In particular it implies that in any distribution $\mathscr{D}$ there exists a unique maximal locally (ad_f, g) invariant distribution which we shall denote by $\mathscr{D}^*(\mathscr{D})$. If $\mathscr{D}$ is involutive then so is $\mathscr{D}^*(\mathscr{D})$. These remarks are predicated on the assumption of invariance with full control. They must be modified when considering invariance with partial control. In particular there may be distributions contained in $\mathscr{D}$ and properly containing $\mathscr{D}^*(\mathscr{D})$ which are locally (ad_f, g) invariant with partial control.

Briefly we discuss the dual formulation of the above, for it is useful in computing maximal locally (ad_f, g) invariant distributions.

DEFINITION. A codistribution $\mathscr{E}$ is (Ad_f, g) *invariant* ((ad_f, g) invariant) if there exists a feedback γ such that $\mathscr{E}$ is $\mathrm{Ad}_{\tilde{f}}$ invariant ($\mathrm{ad}_{\tilde{f}}$ invariant). A codistribution $\mathscr{E}$ is *locally* (ad_f, g) *invariant* if for every constant u and every $\omega \in \mathscr{E} \cap \mathscr{H}(g)$

$$L_{f(\cdot,u)}(\omega) \in \mathscr{E}.$$

Recall $\mathscr{H}(g)$ is the codistribution of one forms which annihilate the columns of g.

LEMMA 5.3. *If the distribution $\mathscr{D}$ is (Ad_f, g) invariant ((ad_f, g) invariant) then the codistribution $\mathscr{D}^\perp$ is (Ad_f, g) invariant ((ad_f, g) invariant). If the codistribution $\mathscr{E}$ is (Ad_f, g) invariant ((ad_f, g) invariant) then the distribution $\mathscr{E}^\perp$ is (Ad_f, g) invariant ((ad_f, g) invariant). If the distribution $\mathscr{D}$ is locally (ad_f, g) invariant then the codistribution $\mathscr{D}^\perp$ is locally (ad_f, g) invariant. If the codistribution $\mathscr{E}$ is locally (ad_f, g) invariant and $\mathscr{E}$ and $\mathscr{E} \cap \mathscr{H}(g)$ are nonsingular then the codistribution $\mathscr{E}^\perp$ is locally (ad_f, g) invariant.*

Proof. The first two assertions are almost immediate. As for the third let $\omega \in \mathscr{D}^\perp \cap \mathscr{H}(g) = (\mathscr{D} + \mathscr{R}(g))^\perp$ and $X \in \mathscr{D}'$, then

$$0 = L_{f(\cdot,u)}\langle \omega, X\rangle = \langle L_{f(\cdot,u)}(\omega), X\rangle + \langle \omega, \mathrm{ad}_{f(\cdot,u)}(X)\rangle. \tag{5.12}$$

Since $\mathscr{D}$ is locally (ad_f, g) invariant the second term on the right is zero hence $L_{f(\cdot,u)}(\omega) \in \mathscr{D}^\perp$.

The last assertion follows in a similar fashion. Let $\omega \in (\mathscr{E}^\perp + \mathscr{R}(g))^\perp = \mathscr{E} \cap \mathscr{H}(g)$ (by the nonsingularity of $\mathscr{E}$) and $X \in \mathscr{E}^\perp$. Since $\mathscr{E}$ is locally (ad_f, g) invariant the first term on the right of (5.12) is zero hence $\mathrm{ad} f(\cdot, u)(X) \in (\mathscr{E}^\perp + \mathscr{R}(g))^{\perp\perp} = \mathscr{E}^\perp + \mathscr{R}(g)$ by the nonsingularity of $(\mathscr{E}^\perp + \mathscr{R}(g))^\perp = \mathscr{E} \cap \mathscr{H}(g)$. QED

In disturbance decoupling and other problems one wishes to find $\mathscr{D}^*(\mathscr{H}(dh))$, the maximal locally (ad_f, g) invariant distribution in $\mathscr{H}(dh)$. We now present an algorithm from [1, p. 342] for the computation of $\mathscr{D}^*(\mathscr{D})$ for an arbitrary distribution $\mathscr{D}$ which works when all the distributions and codistributions involved in the calculations are nonsingular. We then specialize to compute $\mathscr{D}^*(\mathscr{H}(dh))$. When there is no possibility of confusion we shall abbreviate, $\mathscr{D}^* = \mathscr{D}^*(\mathscr{H}(dh))$.

Let $\mathscr{D}$ be an arbitrary distribution and $\mathscr{E}_*(\mathscr{D})$ be the minimal locally (ad_f, g) codistribution containing $\mathscr{D}^\perp$.

Define an increasing sequence of codistributions by

$$\mathscr{E}_0 = \mathscr{D}^\perp \quad \text{and} \quad \mathscr{E}_{k+1} = \mathscr{E}_k + L_f(\mathscr{E}_k \cap \mathscr{H}(g))$$

where the second term on the right denotes the $\mathscr{F}(M)$ span of all one forms like $L_{g^j}(\omega)$ for $j = 0, \cdots, m$ and $\omega \in \mathscr{E}_k \cap \mathscr{H}(g)$.

THEOREM 5.4 (invariant subdistribution algorithm (ISA)). *If there exists a* k_* *such that* $\mathscr{E}_{k_*} = \mathscr{E}_{k_*+1}$ *then* $\mathscr{E}(\mathscr{D}) = \mathscr{E}_{k_*}$. *If in addition* $\mathscr{E}_{k_*}$ *and* $\mathscr{E}_{k_*} \cap \mathscr{H}(g)$ *are nonsingular then* $\mathscr{D}^*(\mathscr{D}) = \mathscr{E}_{k_*}^{\perp}$.

Proof. By definition $\mathscr{E}_*(\mathscr{D})$ contains $\mathscr{E}_0 = \mathscr{D}^{\perp}$ and is locally (ad_f, g) invariant. A simple induction shows that $\mathscr{E}_*(\mathscr{D})$ contains $\mathscr{E}_k$ for all k. If $\mathscr{E}_{k_*} = \mathscr{E}_{k_*+1}$ then $\mathscr{E}_{k_*}$ is locally (ad_f, g) invariant and clearly minimal.

If $\mathscr{E}_{k_*}$ and $\mathscr{E}_{k_*} \cap \mathscr{H}(g)$ are nonsingular then $\mathscr{E}_{k_*}^{\perp}$ is a locally (ad_f, g) invariant distribution by Lemma 5.3. By duality it is the maximal such distribution contained in $\mathscr{D}$. QED

Computation of $\mathscr{D}^* = \mathscr{D}^*(\mathscr{H}(dh))$ *by* ISA.

$$\mathscr{E}_0 = \mathscr{H}(dh)^{\perp} = \mathscr{R}(dh) = \mathscr{R}(dh_1, \cdots, dh_{p_0})$$

where $p_0 = p$. Let $A_0(x)$ be the $p_0 \times m$ matrix whose ith, jth element is $\langle dh_i, g^j \rangle(x)$. Let $B_0(x)$ be the $p_0 \times 1$ vector whose ith element is $\langle dh_i, g^0 \rangle(x)$. Assume the rank of $A_0(x)$ is constant and equal to r_0. By rearranging $h_1, \cdots, h_{p_0}$ if necessary we assume that the first r_0 rows of $A_0(x)$ are linearly independent at each x. Choose $m \times 1 \alpha_0(x)$ and invertible $m \times m$ $\beta_0(x)$ such that

$$A_0(x)\alpha_0(x) + B_0(x) = \begin{pmatrix} 0 \\ \varphi_0 \end{pmatrix}, \tag{5.13a}$$

$$A_0(x)\beta_0(x) = \begin{pmatrix} I^{r_0 \times r_0} & 0 \\ \psi_0 & 0 \end{pmatrix}, \tag{5.13b}$$

where φ_0 and ψ_0 are arbitrary $1 \times (p_0 - r_0)$ and $(p_0 - r_0) \times r_0$ matrix valued functions.

Define $\tilde{g}_0^0 = g^0 + g\alpha_0$, $\tilde{g}_0 = g\beta_0 = (\tilde{g}_0^1, g_0^2)$ where $\tilde{g}_0^1$ is the first r_0 vector fields of $\tilde{g}_0$ and $\tilde{g}_0^2$ the last $m - r_0$. From the functions which are the entries of φ and ψ_0 if (5.13a, b), choose a maximal set whose differentials are linearly independent at each x mod $\mathscr{E}_0$. Call these $h_{p_0+1}, \cdots, h_{p_1}$. We claim that $\mathscr{E}_1 = \mathscr{R}\{dh_1, \cdots, dh_{p_1}\}$.

By definition $\mathscr{E}_1 = \mathscr{E}_0 + L_f(\mathscr{E}_0 \cap \mathscr{H}(g))$, but a straightforward calculation shows that this is the same as $\mathscr{E}_0 + L_{\tilde{f}_0}(\mathscr{E}_0 \cap \mathscr{H}(g))$ where

$$f_0 = f_{\gamma_0}, \qquad \gamma_0 = \begin{pmatrix} 1 & 0 \\ \alpha_0 & \beta_0 \end{pmatrix}$$

because γ_0 is invertible. From (5.13b) we see that

$$dh_i \notin \mathscr{H}(g), \qquad i = 1, \cdots, r_0, \tag{5.14a}$$

$$dh_i - \sum_{k=1}^{r_0} \psi_{0_i}^k \, dh_k \in \mathscr{H}(g), \qquad i = r_0 + 1, \cdots, p_0, \tag{5.14b}$$

so $\mathscr{E}_1$ is the sum of $\mathscr{E}_0$ and the Lie derivatives of (5.14b) by $L_{\tilde{f}_0}$, i.e.

$$L_{\tilde{f}_0}(dh_i - \Sigma \psi_{0i}^k dh_k) = L_{\tilde{f}_0}(dh_i) - \sum (L_{\tilde{f}_0}(\psi_{0i}^k) dh_k + \psi_{0i}^k L_{\tilde{f}_0}(dh_k).$$

But $dh_k \in \mathscr{E}_0$ and

$$L_{\tilde{f}_0}(dh_k) = dL_{\tilde{f}_0}(h_k) = d(0 \text{ or } 1) = 0$$

for $k = 1, \cdots, r_0$. Therefore $\mathscr{E}_1$ is spanned by $\mathscr{E}_0$ and the entries of $L_{\tilde{f}_0}(dh_i)$ for $i = r_0 + 1, \cdots, p_0$. But the latter are either zero or the differentials of the components of φ_0 and ψ_0.

$\mathscr{E}_2$ is constructed in a similar fashion. Let $A_1(x)$ be the $p_1 \times m$ matrix $\langle dh_i, g^j \rangle(x) B_1(x)$ be the $p_1 \times m$ vector $\langle dh_i, g^j \rangle(x)$. Assume $A_1(x)$ is of rank r_1 and

rearrange $h_{r_0+1}, \cdots, h_{p_0}$ so that the first r_1 rows of $A_1(x)$ are linearly independent at each x. Choose α_1, β_1, etc.

Notice that at each stage of this algorithm we obtain codistributions $\mathscr{E}_k$ spanned by exact one-forms, hence they are integrable. The new feedbacks α_{k+1}, β_{k+1} can be obtained by suitably updating α_k and β_k. Moreover α_{k_*} and β_{k_*} are feedbacks which leave $\mathscr{D}^*$ invariant. If $r_{k_*} < m$ then we can partition $\tilde{g}_{k_*} = g\beta_{k_*} = (\tilde{g}^1_{k_*}, \tilde{g}^2_{k_*})$ where $R(\tilde{g}^2_{k_*}) = \mathscr{D}^* \phi R(g)$. We shall make use of this later on.

6. Disturbance decoupling. Consider the nonlinear system

$$\dot{x} = f(x, u) + p(x)w = g^0(x) + g(x)u + p(x)w, \tag{6.1a}$$

$$y = h(x), \tag{6.1b}$$

$$x(0) = x^0. \tag{6.1c}$$

The additional input $w(t)$ represents a disturbance which can be neither controlled nor predicted. We assume it is a bounded measurable function taking values in $\mathbb{R}^l$. The way it affects the dynamics is described by the l vector fields which in local coordinates are the l columns of $p(x)$.

DEFINITION. In the system (6.1) the disturbance is *decoupled* from the output if for each bounded measurable $u(t)$, the output $y(t)$ does not depend on the disturbance $w(t)$. The *disturbance decoupling problem* (DDP) is solvable if there exists a feedback γ such that the disturbance is decoupled from the output for the feedback modified system

$$\begin{aligned} \dot{x} = \tilde{f}(x, v) + p(x)w &= \tilde{g}^0(x) + \tilde{g}(x)v + p(x)w \\ &= g^0(x) + g(x)(\alpha(x) + \beta(x)v) + p(x)w. \end{aligned} \tag{6.2}$$

The *reversible disturbance decoupling problem* (RDDP) is solvable if there exists a feedback γ such that the disturbance is decoupled from the output for the time reversible version of the feedback modified system

$$\begin{aligned} \dot{x} = \tilde{f}(x, v_0, v) + p(x)w &= \tilde{g}^0(x)v_0 + \tilde{g}(x)v + p(x)w \\ &= g^0(x)v_0 + g(x)(\alpha(x)v_0 + \beta(x)v) + p(x)w. \end{aligned} \tag{6.3}$$

Notice that in contrast with controllability and observability, reversible decoupling implies decoupling rather than vice versa. Notice also that the solvability of the RDDP for the original system implies the solvability of the DDP for the time reversible version of the original system but is not equivalent to it. This is because in the former the invertible feedback γ must be of the form $\left(\begin{smallmatrix} 1 & 0 \\ \alpha & \beta \end{smallmatrix}\right)$ while in the latter any invertible feedback is allowed.

The solvability of the DDP and generalizations involving dynamic output feedback are treated at considerable length in [1], see also [2]. We would like to review some of this work using the terminology introduced in this paper and also discuss the solvability of the RDDP. We consider only the solvability of the DDP and RDDP with full control. If a partial control solution is acceptable it can be thought of as full control solution for the system with the unneeded controls deleted.

We state the basic results and defer the proofs to the end of the section.

THEOREM 6.1. *The* RDDP *is solvable iff there exists an* (Ad_f, g) *invariant distribution* $\mathscr{D}$ *such that* $\mathscr{R}(p) \subset \mathscr{D} \subset \mathscr{H}(dh)$.

Every (Ad_f, g) invariant distribution is also (ad_f, g) invariant so the above theorem implies that if the RDDP is solvable then there must exist an (ad_f, g) invariant

distribution $\mathscr{D}$ such that $\mathscr{R}(p) \subset \mathscr{D} \subset \mathscr{H}(dh)$. Since $\mathscr{H}(dh)$ is involutive we can conclude that there must exist such a $\mathscr{D}$ which is involutive. But one can make a stronger statement.

THEOREM 6.2. *If the* DDP *is solvable then there exists an involutive* (ad_f, g) *invariant distribution* $\mathscr{D}$ *such that* $\mathscr{R}(p) \subset \mathscr{D} \subset \mathscr{H}(dh)$.

The converse of this theorem is not true as is shown by Example 6.6. Recall that the (Ad_f, g) or (ad_f, g) invariant distributions do not form a semilattice while the locally (ad_f, g) invariant ones do. We denote by $\mathscr{D}^*$ the maximal locally (ad_f, g) invariant distribution in $\mathscr{H}(dh)$, $\mathscr{D}^* = \mathscr{D}^*(\mathscr{H}(dh))$. In § 5 an algorithm for the computation of $\mathscr{D}^*$ was presented.

DEFINITION. The DDP (RDDP) is *locally solvable* if every $x^0 \in M$ has an open neighborhood $\mathscr{U}$ and a feedback γ defined on $\mathscr{U}$ which solves the DDP(RDDP) restricted to $\mathscr{U}$.

THEOREM 6.3. *If the* DDP *is locally solvable then* $\mathscr{R}(p) \subset \mathscr{D}^*$. *If* $\mathscr{R}(p) \subset \mathscr{D}^*$ *and* $\mathscr{D}^*$ *is nonsingular and separates the controls then the* RDDP *is locally solvable.*

The proofs of the above depend heavily upon the following lemmas. (These lemmas describe basic properties of (Ad_f, g) and (ad_f, g) controllability distributions, concepts which will be introduced in the next section.) Let $\langle \mathrm{Ad}_{f,p} | \mathscr{R}(p) \rangle$ denote the minimal Ad_f and Ad_p invariant distribution which contains $\mathscr{R}(p)$. By Lemmas 3.2, 3.3, 3.4 and Sussmann's Theorem 2.1, this distribution is integrable. Let $(u_0(t), u(t))$ be a bounded measurable input defined on $[0, T]$ for the time reversible version (6.4) of (6.1),

$$\dot{x} = f(x, u_0, u) + p(x)w = g^0(x)u_0 + g(x)u = p(x)w. \tag{6.4}$$

Let $\mathscr{RA}(x^0, T, u_0(t), u(t))$ denote the set of points accessible from x^0 at time T along trajectories of (6.4) with $(u_0(t), u(t))$ fixed and $w(t)$ varying over all bounded measurable disturbances. Let x^T be the endpoint of the trajectory for $w(t) = 0$.

LEMMA 6.4. *Let* L *be the leaf through* x^T *of the foliation induced by* $\langle \mathrm{Ad}_{(f,p)} | \mathscr{R}(p) \rangle$; *then* $\mathscr{RA}(x^0, T, u_0(t), u(t)) \subset L$. *Moreover for some piecewise constant control* $(u_0(t), u(t))$, $x^0 = x^T$ *and* $\mathscr{RA}(x^0, T, u_0(t), u(t))$ *is a neighborhood of* x^0 *in the topology of the leaf containing* x^0.

Let $\langle \mathrm{ad}_{(f,p)} | \mathscr{R}(p) \rangle$ denote the minimal ad_f and ad_p invariant distribution containing $\mathscr{R}(p)$. By the Jacobi identity (2.8b) this distribution is involutive. Let $\mathscr{U}$ be an open neighborhood of x^0 and $u(t)$ be a bounded measurable control defined on $[0, T]$ which generates a trajectory $x(t)$ of (2.1) from x^0 which lies in $\mathscr{U}$ for all $t \in [0, T]$. Let $\mathscr{A}(x^0, T, \mathscr{U}, u(t))$ be set of points accessible under (6.1) from x^0 in $\mathscr{U}$ at time T with $u(t)$ *fixed and* $w(t)$ varying over all bounded measurable disturbances.

LEMMA 6.5. *Let* $\mathscr{U}$ *be an open neighborhood of* x^0 *on which* $\langle \mathrm{ad}_{(f,p)} | \mathscr{R}(p) \rangle$ *is nonsingular, hence integrable. Let* L *be the leaf through* x^T; *then* $\mathscr{A}(x^0, T, \mathscr{U}, u(t)) \subset L$. *Moreover there exists a piecewise constant control* $u(t)$ *such that* $\mathscr{A}(x^0, T, \mathscr{U}, u(t))$ *has nonempty interior in the topology of this leaf.*

Next we give the proofs of these results and a counterexample to the converse of Theorem 6.2.

Proof of Lemma 6.4. Without loss of generality we can assume that $\langle \mathrm{Ad}_{(f,p)} | \mathscr{R}(f, p) \rangle = \mathscr{K}(M)$ or in other words, with $u(t)$ and $w(t)$ as controls, (6.1) is reversibly controllable. For if not, $\langle \mathrm{Ad}_{(f,p)} | \mathscr{R}(f, p) \rangle$ is an integrable distribution and by replacing the state space by the leaf of this distribution through x^0 we obtain a reversibly controllable system. The assumption of reversible controllability insures that $\mathscr{D} = \langle \mathrm{Ad}_{(f,p)} | \mathscr{R}(p) \rangle$ is nonsingular, for any x^0 and x^T can be joined by a trajectory constructed from the flows of g^j, $j = 0, \cdots, m$ and p^k, $k = 1, \cdots, l$. But if D is the subbundle of TM corresponding to $\mathscr{D}$ then the Jacobian of these composed flows is an isomorphism between $D(x^0)$ and $D(x^T)$.

Now suppose $(u_0(t), u(t))$ and $w(t)$ are bounded measurable functions on $[0, T]$. Let $x(t)$ be the solution (6.4) and $\Phi(t, s, x)$ the time dependent flow of (6.4) with $w(t) = 0$, i.e.,

$$\frac{\partial \Phi}{\partial t}(t, s, x) = g^0(\Phi(t, s, x))u_0(t) + g(\Phi(t, s, x))u(t),$$

$$\Phi(t, t, x) = x.$$

The mapping $x \to \Phi(t, s, x)$ is smooth and its Jacobian carries $D(x)$ onto $D(\Phi(t, s, x))$. Consider the trajectory $\tilde{x}(s)$ defined by

$$\tilde{x}(s) = \Phi(T, s, x(s)).$$

Clearly $\tilde{x}(0) = x^T$ (the endpoint of the solution of (6.4) with $w(t) = 0$) and $\tilde{x}(T) = x(T)$ (the endpoint of the solution of (6.4) with $w(t)$ as above). Moreover

$$\frac{d}{ds}\tilde{x}(s) = \frac{\partial \Phi}{\partial x}(T, s, x(s))(p(x(s))w(s))$$

hence is an element of $D(\tilde{x}(s))$. The nonsingularity of $\mathcal{D}$ implies $\tilde{x}(s)$ lies in the leaf L of $\mathcal{D}$ through x^T. Therefore $x(T) = \tilde{x}(T) \in L$ and $\mathcal{RA}(x^0, T, u_0(t), u(t)) \subset L$.

To prove the second assertion first we note that $\mathcal{D}$ is spanned by expressions of the form

$$\text{Ad}^{s_k}_{(f^k + p^k)} \circ \cdots \circ \text{Ad}^{s_1}_{(f^1 + p^1)} p^0 \tag{6.5}$$

where $f^j(x) = g^0(x)u_0^j + g(x)u^j$ and $p^j(x) = p(x)w^j$ for some constants u_0^j, u^j, w^j and s_j. By rescaling u_0^j, u^j, w^j we can assume $s_j < 0$ and fix the sum $\sum_{j=1}^k s_j$ arbitrarily. Choose an expression of the form (6.5) which is not zero at x^0, and define piecewise constant functions

$$u_0(t) = u_0^j, \quad u(t) = u^j, \quad w(t) = w^j \quad \text{for } t \in [t_{j-1}, t_j) \tag{6.6}$$

where $t_k = T$ and $t_{j-1} - t_j = s_j$. Assume that $\sum_{j=1}^k s_j = -T/2$. Let $x(t)$ be the solution of (6.4) satisfying the terminal condition $x(T) = x^0$. If we modify $w(t)$ to $w(t; \varepsilon)$

$$w(t; \varepsilon) = \begin{cases} w^1 + w^0 & \text{if } t \in [t_0, t_0 + |\varepsilon|) \text{ and } \varepsilon > 0, \\ w^1 - w^0 & \text{if } t \in [t_0, t_0 + |\varepsilon|) \text{ and } \varepsilon < 0, \\ w(t) & \text{otherwise,} \end{cases} \tag{6.7}$$

and let $x(t, \varepsilon)$ be the solution of (6.4) satisfying the initial condition $x(t_0, \varepsilon) = x(t_0)$ then $\partial/\partial\varepsilon\ (x(T; 0))$ is precisely (6.5) evaluated at x^0.

By reversing the order and the signs of the inputs (6.6) we can get from x^0 to $x(t_0)$ in time $T/2$ and use the original sequence of inputs (6.6) to go back to x^0. Suppose we vary ε only on the second half according to (6.7), $x(t; \varepsilon)$ is now the endpoint of the total trajectory and it sweeps out a one-dimensional C^1 submanifold containing x^0 in its interior which is contained in the integral manifold L of $\mathcal{D}$ through x^0.

If the dimension of $\mathcal{D}$ is greater than one we repeat the process, this time at $x(t_0)$ instead of x^0. We also choose a new expression (6.5) which is linearly independent of the tangent to our one-dimensional submanifold pulled back to $x(t_0)$. This is always possible since expressions of the form (6.5) span $\mathcal{D}$ at x^0.

In this way we generate a one parameter family of controls which generates a one-dimensional manifold with $x(t_0)$ in its interior. When this is pulled on to x^0 along

the original variation we get a two-dimensional submanifold of L with x^0 in its interior. We repeat the construction until the dimension of $\mathscr{D}$ is achieved. QED

Proof of Theorem 6.1. If the RDDP is solvable by a feedback, γ, let $\tilde{f}=f\gamma$ and let $\mathscr{D}=\langle \mathrm{Ad}_{(\tilde{f},p)}|\mathscr{R}(p)\rangle$. Clearly $\mathscr{D}$ is $\mathrm{Ad}_{\tilde{f}}$ invariant, hence (Ad_f, g) invariant and contains $\mathscr{R}(p)$. Suppose $\mathscr{D}$ is not contained in $\mathscr{H}(dh)$. Then at some x^0, $D(x^0)\not\subset dh(x^0)$. By Lemma 6.4 there exist T and a piecewise constant $(u_0(t), u(t))$ such that $\mathscr{R}\mathscr{A}(x^0, T, u_0(t), u(t))$ is a neighborhood of x^0 in the leaf N of $\mathscr{D}$ through x^0. This implies that $\mathscr{R}\mathscr{A}(x^0, T, u_0(t), u(t))$ is not contained in a level set of h, contradicting the solvability of the RDDP.

On the other suppose such a $\mathscr{D}$ exists. Let γ be the feedback such that $\mathscr{D}$ is $\mathrm{Ad}_{\tilde{f}}$ invariant for $\tilde{f}=f\gamma$. The integrable closure of $\mathscr{D}$ must contain $\langle \mathrm{Ad}_{(\tilde{f},p)}|\mathscr{R}(p)\rangle$ and since $\mathscr{H}(dh)$ is integrable it must contain the integrable closure of $\mathscr{D}$. Hence

$$\mathscr{R}(p)\subset\langle \mathrm{Ad}_{(\tilde{f},p)}|\mathscr{R}(p)\rangle\subset\mathscr{H}(dh).$$

Lemma 6.4 implies that for any fixed $(u_0(t), u(t))$ and T, $\mathscr{R}\mathscr{A}(x^0, T, u^0(t), u(t))$ is contained in a leaf of $\langle \mathrm{Ad}_{(\tilde{f},p)}|\mathscr{R}(p)\rangle$ which in turn is contained in a level set of h. Therefore the RDDP is strongly solvable. QED

Proof of Lemma 6.5. The first assertion follows from an application of Lemma 6.4 to the system restricted to $\mathscr{U}$, for the nonsingularity of $\langle \mathrm{ad}_{(f,p)}|\mathscr{R}(p)\rangle$ implies it equals $\langle \mathrm{Ad}_{(f,p)}|\mathscr{R}(p)\rangle$.

The proof of the second is similar to that of [1, Lemma 3.5]. Suppose $\mathscr{D}=\langle \mathrm{ad}_{(f,p)}|\mathscr{R}(p)\rangle$ is of dimension d on $\mathscr{U}$. $\mathscr{D}$ is spanned by terms of the form

$$\mathrm{ad}_{(f^k+p^k)}\circ\cdots\circ\mathrm{ad}_{(f^1+p^1)}p^0. \tag{6.8}$$

$\mathscr{D}$ is the involutive closure of $\langle \mathrm{ad}_f|\mathscr{R}(p)\rangle$ which is spanned by terms of the form

$$\mathrm{ad}_{f^k}\circ\cdots\circ\mathrm{ad}_{f^1}p^0. \tag{6.9}$$

Choose an expression (6.9) which is not 0 at x^0 define $u(t)$, $w(t,\varepsilon)$ for small $\varepsilon>0$ by

$$u(t)=u^j,\qquad t\in[t_{j-1}, t_j],$$

$$w(t,\varepsilon)=\begin{cases} w^0 & \text{if } t\in[t_0, t_0+\varepsilon),\\ 0 & \text{otherwise,}\end{cases}$$

where $t_0=0$ and $t_1,\cdots,t_n$ are to be determined. We have to take care choosing u^j, w^0, and t_j sufficiently small so that the trajectories $x(t;\varepsilon)$ of (6.1) from x^0 remain in $\mathscr{U}$ and $t_k<T$. Henceforth we shall not mention this point.

Since (6.9) is not zero for some choice of $t_1,\cdots,t_n$, as we vary ε, $x(t_k;\varepsilon)$ sweeps out a one-dimensional submanifold. Suppose that at some point on this submanifold there is an expression of the form (6.9) which is not tangent to the submanifold. Then we can repeat this process and construct a two-dimensional submanifold of points accessible at some later time. We repeat the process until we obtain a d-dimensional submanifold of accessible points such that every expression of the form (6.9) is tangent to it. Since the vector fields tangent to a manifold are trivially involutive and $\mathscr{D}$ is the involutive closure of (6.9), this manifold must be an integral manifold of $\mathscr{D}$. This shows that $\mathscr{A}(x^0, T, \mathscr{U}, u(t))$ has nonempty interior in the leaf topology. QED

Proof of Theorem 6.2. Suppose the DDP is solvable using feedback γ, let $\tilde{f}=f\gamma$ and $\mathscr{D}=\langle \mathrm{ad}_{(\tilde{f},p)}|\mathscr{R}(p)\rangle$. Clearly $\mathscr{D}$ is involutive and contains $\mathscr{R}(p)$, so all we need to show is that $\mathscr{D}\subset\mathscr{H}(dh)$.

Recall that x^0 is a regular point of $\mathscr{D}$ if $\mathscr{D}$ is nonsingular in a neighborhood of x^0. The regular points of $\mathscr{D}$ are open and dense in M hence by continuity it suffices

to verify that at each regular point x^0 the subbundle D associated to $\mathscr{D}$ satisfies $D(x^0) \perp dh(x^0)$.

Let x^0 be a regular point and $\mathscr{U}$ a neighborhood on which $\mathscr{D}$ is nonsingular. By Lemma 6.5, there exists $x^T \in \mathscr{U}$ such that $\tilde{\mathscr{A}}(x^0, T, \mathscr{U}, v(t))$ is a neighborhood of x^T in the leaf of $\mathscr{D}$ containing x^T. (We use $\tilde{\mathscr{A}}$ and $v(t)$ instead of $\mathscr{A}$ and $u(t)$ to indicate this is the $\mathscr{U}$ accessible set for fixed $v(t)$ and variable $w(t)$ of the feedback modified dynamics (6.3).) Since the feedback decouples the output from the disturbance we conclude that $D(x^T) \perp dh(x^T)$. But x^T is arbitrarily close to x^0 so $D(x^0) \perp dh(x^0)$. QED

The following example shows that the converse to Theorem 5.3 is not true. We present it as a time varying linear system

$$\dot{x} = A(t)x + B(t)u + E(t)w, \tag{6.10a}$$

$$y = C(t)x, \tag{6.10b}$$

$$x(0) = x^0, \tag{6.10c}$$

which can easily be made into an autonomous nonlinear system (6.1) by letting time be an extra state coordinate, say $x_0 = t$.

Example 6.6. Let $\rho(t)$ be a $\mathscr{C}^\infty$ function such that $\rho(t) = 0$ for $t \leqq 0$, $\rho(t) = \pi/2$ for $t \geqq 1$ and $\dot{\rho}(t) > 0$ for $t \in (0, 1)$. Define

$$A(t) = \dot{\rho}(t) \begin{pmatrix} -\sin \rho(t) & -\cos \rho(t) & 0 \\ \cos \rho(t) & -\sin \rho(t) & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

for $t \leqq 1.5$ and for $t \geqq 1.5$

$$A(t) = \dot{\rho}(t-2) \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\sin \rho(t-2) & -\cos \rho(t-2) \\ 0 & \cos \rho(t-2) & -\sin \rho(t-2) \end{pmatrix}.$$

The free dynamics (6.10a) for $u = 0$ is constant except for $t \in (0, 1) \cap (2, 3)$. On the time interval $(0, 1)$ the x_1-x_2 plane is rotated through an angle of $\pi/2$ and on $(2, 3)$ the x_2-x_3 plane is similarly rotated. Let

$$B(t) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \qquad E(t) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \qquad C(t) = (0 \quad 0 \quad 1).$$

Viewed as a nonlinear system (6.1) on the extended four-dimensional space $(x_0 = t, x_1, x_2, x_3)$ the system is not disturbance decoupled. Disturbances at small positive times affect the x_1 coordinate and are rotated to affect the x_2 coordinate. Later after $t = 2$ these disturbances are rotated to affect x_3 and hence the output. Since $B(t) = 0$ the system cannot be disturbance decoupled.

However there is an (ad_f, g) invariant distribution $\mathscr{D}$ such that $\mathscr{R}(p) \subset \mathscr{D} \subset \mathscr{H}(dh)$. Of course $\mathscr{D}$ must be singular else it would be (Ad_f, g) invariant and Theorem 6.1 would apply. Let $\sigma(t)$ be a $\mathscr{C}^\infty$ function such that $\sigma(t) = 1$ for $t \leqq 1$ and $\sigma(t) = 0$ for $t \geqq 2$. Let $\mathscr{D}$ be spanned by the vector fields

$$X^1 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \qquad X^2 = \begin{pmatrix} 0 \\ 0 \\ \sigma(x_0) \\ 0 \end{pmatrix}.$$

We leave it to the reader to verify that $\mathscr{D}$ is (ad_f, g) (in fact ad_f) invariant, as a start note that

$$g^0 = \begin{pmatrix} 1 \\ A(x_0)x \end{pmatrix}, \qquad g = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Proof of Theorem 6.3. For given x^0 and $\mathscr{U}$, let γ be the feedback which solves the DDP on $\mathscr{U}$. By Theorem 6.2 on $\mathscr{U}$ there exists an (ad_f, g) distribution $\mathscr{D}$ such that $\mathscr{R}(p) \subset \mathscr{D} \subset \mathscr{H}(dh)$. Let $\mathscr{V}$ be an open neighborhood of x^0 whose closure is contained in $\mathscr{U}$ and let φ be a $\mathscr{C}^\infty$ function 1 on $\mathscr{V}$ and 0 off $\mathscr{U}$. The distribution $\varphi\mathscr{D} = \{\varphi X : X \in \mathscr{D}\}$ is globally defined and satisfies $\varphi\mathscr{D} \subset \mathscr{H}(dh)$. Moreover $\varphi\mathscr{D}$ is (ad_f, g) invariant hence locally (ad_f, g) invariant so $\varphi\mathscr{D} \subset \mathscr{D}^*$. Therefore for each x^0 there exists neighborhood $\mathscr{V}$ such that on $\mathscr{V}$, $\mathscr{R}(p) \subset \mathscr{D} = \varphi\mathscr{D} \subset \mathscr{D}^*$, hence $\mathscr{R}(p) \subset \mathscr{D}^*$.

As for the second assertion, give x^0 Lemma 5.1 allows us to conclude that in some neighborhood of $\mathscr{U}$ and x^0, $\mathscr{D}^*$ is (Ad_f, g) invariant. Since $\mathscr{R}(p) \subset \mathscr{D}^* \subset \mathscr{H}(dh)$, Theorem 6.1 implies the RDDP is solvable on $\mathscr{U}$. QED

7. Controllability distributions. We now define the nonlinear generalizations of the concept of an (A, B) controllability subspace. These were introduced by Krener and Isidori [6], see also [14].

DEFINITION. A distribution $\mathscr{C}$ (with associated subbundle $C \subset TM$) is an (Ad_f, g) *controllability distribution* ((ad_f, g) *controllability distribution*) if there exists an invertible feedback γ with β partitioned as $(\beta^1 \beta^2)$ such that $\mathscr{C}$ separates the controls (see (5.11)), i.e., if $\tilde{g}^\sigma = g\beta^\sigma$ then for every x

$$C(x) \cap G(x) = \tilde{G}(x)^1, \qquad C(x) \cap \tilde{G}(x)^2 = \{0\} \tag{7.1a}$$

and

$$\mathscr{C} = \langle \mathrm{Ad}_{\tilde{f}} | \mathscr{R}(\tilde{g}^1) \rangle \qquad (\mathscr{C} = \langle \mathrm{ad}_{\tilde{f}} | \mathscr{R}(\tilde{g}^1) \rangle). \tag{7.1b}$$

It follows immediately from (7.1b) that any such $\mathscr{C}$ is (Ad_f, g) invariant ((ad_f, g) invariant), that (Ad_f, g) controllability distributions are integrable and that (ad_f, g) controllability distributions are involutive. Notice that (Ad_f, g) controllability does not necessarily imply (ad_f, g) controllability because the inclusion

$$\langle \mathrm{ad}_{\tilde{f}} | \mathscr{R}(\tilde{g}^1) \rangle \subset \langle \mathrm{Ad}_{\tilde{f}} | \mathscr{R}(\tilde{g}^1) \rangle$$

could be proper. However if $\mathscr{C}$ is (ad_f, g) controllable and nonsingular then the inclusion is an identity, hence $\mathscr{C}$ is (Ad_f, g) controllable.

We have already encountered several examples of such distributions. The Ad_f and ad_f controllability distributions of § 4 are (Ad_f, g) and (ad_f, g) controllability distributions for the time reversible system (4.1). (Here u_0 is an additional control and $\gamma = \beta^1 = I$.) Other important examples are the Ad_f and ad_f *exact time controllability distributions* $\langle \mathrm{Ad}_f | \mathscr{R}(g) \rangle$ and $\langle \mathrm{ad}_f | \mathscr{R}(g) \rangle$. These first appeared in the work of Sussmann and Jurdjevic [18], who considered only analytic systems, so there was no need to distinguish between the two. The first is integrable and for each x^0 and T there exists a leaf which contains $\mathscr{A}(x^0, T)$. The second is involutive; if $\mathscr{U}$ is a neighborhood of x^0 on which it is nonsingular then for each T sufficiently small, $\mathscr{A}(x^0, T, \mathscr{U})$ is contained in a leaf of $\langle \mathrm{ad}_f | \mathscr{R}(g) \rangle$ and has nonempty interior in the leaf topology. These statements follow from Lemmas 6.4 and 6.5.

These lemmas can be applied to arbitrary (Ad_f, g) or (ad_f, g) controllability distributions. The vector fields $\tilde{g}^0$, $\tilde{g}^1$ and $\tilde{g}^2$ of the feedback modified dynamics for the controllability distribution become g^0, p and g respectively in the context of these lemmas.

As one might expect there is a local version of the above concepts.

DEFINITION. Let $\mathscr{D}$ be an arbitrary distribution; we denote by $\mathscr{C}^*(\mathscr{D})$ the minimal distribution $\hat{\mathscr{D}}$ which satisfies

$$\hat{\mathscr{D}} = \mathscr{D} \cap (\mathrm{ad}_f(\hat{\mathscr{D}}) + \mathscr{R}(g)). \tag{7.2}$$

The notation $\mathrm{ad}_f(\hat{\mathscr{D}})$ denotes the $\mathscr{F}(M)$ span of all vector fields $[g^j, X]$ when $j = 0, \cdots, m$ and $X \in \hat{\mathscr{D}}$. It is not apparent that the set of distributions satisfying (7.2) is closed under intersection, hence we do not know that $\mathscr{C}^*(\mathscr{D})$ always exists. This will be shown by the controllability subdistribution algorithm.

DEFINITION. A distribution $\mathscr{C}$ is *locally* (ad_f, g) *controllable* if $\mathscr{C}$ is locally (ad_f, g) invariant and $\mathscr{C} = \mathscr{C}^*(\mathscr{C})$.

CONTROLLABILITY SUBDISTRIBUTION ALGORITHM (CSA, compare with [15, p. 110]).

Let $\mathscr{D}$ be an arbitrary distribution, $\mathscr{C}^0 = \{0\}$ and

$$\mathscr{C}^k = \mathscr{D} \cap (\mathrm{ad}_f(\mathscr{C}^{k-1}) + \mathscr{R}(g)). \tag{7.3}$$

Clearly $\mathscr{C}^0 \subset \mathscr{C}^1$, by induction $\mathscr{C}^{k-1} \subset \mathscr{C}^k$. For if $\mathscr{C}^{k-2} \subset \mathscr{C}^{k-1}$ then

$$\mathscr{C}^{k-1} = \mathscr{D} \cap (\mathrm{ad}_f(\mathscr{C}^{k-2}) + \mathscr{R}(g)) \subset \mathscr{D} \cap (\mathrm{ad}_f(\mathscr{C}^{k-1}) + \mathscr{R}(g)) = \mathscr{C}^k. \tag{7.4}$$

We claim that $\cup \mathscr{C}^k$ is the minimal distribution satisfying (7.2), i.e.

$$\mathscr{C}^*(\mathscr{D}) = \bigcup_{k \geq 0} \mathscr{C}^k. \tag{7.5}$$

Clearly $\cup \mathscr{C}^k$ satisfies (7.2). If $\hat{\mathscr{D}}$ is any distribution satisfying (7.2) then $\mathscr{C}^0 \subset \hat{\mathscr{D}}$ and an induction similar to (7.4) shows that $\mathscr{C}^k \subset \hat{\mathscr{D}}$ for all k. Hence $\cup \mathscr{C}^k$ is the minimal distribution satisfying (7.2).

It is very important to note that for an arbitrary distribution $\mathscr{D}$, $\mathscr{C}^*(\mathscr{D})$ is *not* invariant and hence not controllable in any of the above senses.

LEMMA 7.1. *Let $\mathscr{D}$ be locally (ad_f, g) invariant; then $\mathscr{C}^*(\mathscr{D})$ is locally (ad_f, g) invariant and in fact is the unique maximal locally (ad_f, g) controllability distribution contained in $\mathscr{D}$.*

Proof. Suppose $X \in \mathscr{C}^*(\mathscr{D})$ as defined by (7.5). Since $\mathscr{D}$ is locally (ad_f, g) invariant there exists $Y \in \mathscr{R}(g)$ such that

$$\mathrm{ad}_f(X) + Y \in \mathscr{D}. \tag{7.6}$$

(A word of caution regarding notation is in order. By the above we mean that for $i = 0, \cdots, m$ there exists $Y^i \in \mathscr{R}(g)$ such that

$$\mathrm{ad}_g i(X) + Y^i \in \mathscr{D}.$$

In (7.6) Y is a matrix whose columns are $Y^0, \cdots, Y^m$. Without mentioning it again we will continue to abuse notation in this fashion.) Since $\mathscr{C}^*(\mathscr{D})$ satisfies (7.2) it follows that

$$\mathrm{ad}_f(X) + Y \in \mathscr{C}^*(\mathscr{D})$$

so $\mathscr{C}^*(\mathscr{D})$ is locally (ad_f, g) invariant.

Let $\mathscr{C}^k$ be defined (7.3). Then $\mathscr{C}^k \subset \mathscr{C}^*(\mathscr{D})$ so

$$\mathscr{C}^k = \mathscr{C}^*(\mathscr{D}) \cap \mathscr{C}^k = \mathscr{C}^*(\mathscr{D}) \cap (\mathrm{ad}_f(\mathscr{C}^{k-1}) + \mathscr{R}(g)),$$

$$\mathscr{C}^k = \mathscr{C}^*(\mathscr{D}) \cap (\mathrm{ad}_f(\mathscr{C}^{k-1}) + \mathscr{R}(g)),$$

hence $\mathscr{C}^*(\mathscr{D})$ is locally (ad_f, g) controllable.

Suppose $\mathscr{C}$ is any other local (ad_f, g) controllability distribution in $\mathscr{D}$, define $\hat{\mathscr{C}}^0 = \{0\}$ and

$$\hat{\mathscr{C}}^k = \hat{\mathscr{C}} \cap (\mathrm{ad}_f(\hat{\mathscr{C}}^{k-1}) + \mathscr{R}(g)).$$

Since $\hat{\mathscr{C}}^0 = \mathscr{C}^0$ and $\hat{\mathscr{C}} \subset \mathscr{C}$, a simple induction similar to (7.4) shows that $\hat{\mathscr{C}}^k \subset \mathscr{C}^k$ and hence $\hat{\mathscr{C}} = \cup\, \hat{\mathscr{C}}^k \subset \cup\, \mathscr{C}^k = \mathscr{C}^*(\mathscr{D})$. Therefore $\mathscr{C}^*(\mathscr{D})$ is maximal. QED

From this lemma we see that every distribution $\mathscr{D}$ contains a unique maximal locally (ad_f, g) controllable distribution. The argument proceeds in two steps. Since the locally (ad_f, g) invariant distributions form a semilattice under addition, every distribution contains a unique maximal locally (ad_f, g) invariant distribution $\mathscr{D}^*(\mathscr{D})$. By the above lemma this distribution contains an unique maximal locally (ad_f, g) controllability distribution $\mathscr{C}^*(\mathscr{D}^*(\mathscr{D}))$. Note that $\mathscr{C}^*(\mathscr{D}^*(\mathscr{D})) \subset \mathscr{C}^*(\mathscr{D})$ but generally this is a proper inclusion. Frequently we shall wish to compute $\mathscr{C}^*(\mathscr{D}^*(\mathscr{H}(dh)))$ which we shall abbreviate $\mathscr{C}^*$ when there is no possibility of confusion. At the end of this section we discuss the computation of $\mathscr{C}^*$ by extending the algorithm for $\mathscr{D}^*$ of § 5.

The above remarks are predicted on the assumption of full control. There may exist distributions $\hat{\mathscr{D}}$ which are locally (ad_f, g) invariant with partial control such that $\mathscr{D}^*(\mathscr{D}) \subseteq \hat{\mathscr{D}} \subset \mathscr{D}$. On the other hand from the CSA we see that if $\hat{\mathscr{D}}$ is any locally (ad_f, g) controllability distribution with partial control that is contained in $\mathscr{D}$ then $\hat{\mathscr{D}} \subset \mathscr{C}^*(\mathscr{D})$.

The set of locally (ad_f, g) controllability distributions is a semilattice under addition.

LEMMA 7.2. *Suppose $\mathscr{C}_1$ and $\mathscr{C}_2$ are locally (ad_f, g) controllable. Then so is $\mathscr{C} = \mathscr{C}_1 + \mathscr{C}_2$.*

Proof. Of course $\mathscr{C}$ is locally (ad_f, g) invariant. Let $\mathscr{C}_i^k$ and $\mathscr{C}^k$ be defined by the controllability subdistribution algorithm (7.3) applied to $\mathscr{C}_i$ and $\mathscr{C}$. Clearly $\mathscr{C}_i^0 = \mathscr{C}^0$ and $\mathscr{C}_i \subset \mathscr{C}$ so by induction $\mathscr{C}_i^k \subset \mathscr{C}^k$. But

$$\mathscr{C} = \mathscr{C}_1 + \mathscr{C}_2 = \left(\bigcup_{k\geqq 0} \mathscr{C}_1^k\right) + \left(\bigcup_{k\geqq 0} \mathscr{C}_2^k\right) = \bigcup_{k\geqq 0} (\mathscr{C}_1^k + \mathscr{C}_2^k) \subset \bigcup_{k\geqq 0} \mathscr{C}^k \subset \mathscr{C}. \qquad \text{QED}$$

The next lemma is important for it shows that if $\mathscr{D}$ is (ad_f, g) invariant then $\langle \mathrm{ad}_{\tilde{f}} | \mathscr{D} \cap \mathscr{R}(g) \rangle$ is independent of the choice of feedback so long as it leaves $\mathscr{D}$ invariant.

LEMMA 7.3. *Suppose $\mathscr{D}$ is (ad_f, g) invariant under γ. Let $\mathscr{C}^k$ and $\mathscr{C}^*(\mathscr{D})$ be defined by the* CSA (7.3), (7.5) *applied to $\mathscr{D}$. Then for $k \geqq 1$*

$$\mathscr{C}^k = \sum_{j=0}^{k} \mathrm{ad}_{\tilde{f}}^{j-1} (\mathscr{D} \cap \mathscr{R}(g))$$

and

$$\mathscr{C}^*(\mathscr{D}) = \langle \mathrm{ad}_{\tilde{f}} | \mathscr{D} \cap \mathscr{R}(g) \rangle.$$

Proof. The second assertion follows from the first which follows by induction. For $k = 1$ it is clearly true. Suppose it holds for $k-1$. Let $X \in \mathscr{C}^{k-1}$. Then

$$\mathrm{ad}_{\tilde{f}}(X) = \mathrm{ad}_f(X)\gamma - fL_X(\gamma).$$

Since $\gamma = \begin{pmatrix} 1 & 0 \\ \alpha & \beta \end{pmatrix}$ it follows that $L_X(\gamma) = \begin{pmatrix} 0 & 0 \\ * & * \end{pmatrix}$ and $fL_X(\gamma)$. Moreover γ is invertible so

$$\mathrm{ad}_{\tilde{f}}(X) + \mathscr{R}(g) = \mathrm{ad}_f(X) + \mathscr{R}(g).$$

This allows us to express $\mathscr{C}^k$ as

$$\mathscr{C}^k = \mathrm{ad}_{\tilde{f}}(\mathscr{C}^{k-1}) + (\mathscr{D} \cap \mathscr{R}(g)).$$

$$= \sum_{j=0}^{k} \mathrm{ad}_{\tilde{f}}^{j-1}(\mathscr{D} \cap \mathscr{R}(g)). \qquad \text{QED}$$

COROLLARY 7.4. *If $\mathscr{C}$ is (ad_f, g) controllable then $\mathscr{C}$ is locally (ad_f, g) controllable.*

Proof. If $\mathscr{C}$ is (ad_f, g) controllable then it is (ad_f, g) invariant, hence locally (ad_f, g) invariant. Let γ be a feedback which leaves $\mathscr{C}$ invariant and separates the controls, so that $C_x \cap G_x = \tilde{G}_x^1$ then $\mathscr{C}^*(\mathscr{C}) = \langle \mathrm{ad}_{\tilde{f}} | \mathscr{C} \cap \mathscr{R}(g) \rangle = \langle \mathrm{ad}_{\tilde{f}} | \mathscr{R}(\tilde{g}^1) \rangle = \mathscr{C}$ so $\mathscr{C}$ is locally (ad_f, g) controllable. QED

COROLLARY 7.5. *If $\mathscr{C}$ is nonsingular, involutive and separates the controls then the following are equivalent.*

(a) *$\mathscr{C}$ is locally (ad_f, g) controllable.*

(b) *There exist an open cover $\{\mathscr{U}^\rho\}$ of M and separating feedbacks γ^ρ such that $\mathscr{C}$ is (ad_f, g) controllable on $\mathscr{U}^\rho$ under γ^ρ (in other words, locally $\mathscr{C}$ is (ad_f, g) controllable).*

(c) *There exist an open cover $\{\mathscr{U}^\rho\}$ of M and separating feedbacks γ^ρ such that $\mathscr{C}$ is (Ad_f, g) controllable on γ^ρ under γ^ρ (locally $\mathscr{C}$ is (Ad_f, g) controllable).*

Proof. This follows directly from Lemmas 5.1 and 7.3. QED

Computation of $\mathscr{C}^* = \mathscr{C}^*(\mathscr{D}^*(\mathscr{H}(dh))$. One could apply the CSA to $\mathscr{D}^* = \mathscr{D}^*(dh)$ computed by the ISA of § 5. A more convenient approach is to apply Lemma 7.3 so that in the notation of the end of § 5,

$$\mathscr{C}^* = \langle \mathrm{ad}_{\tilde{f}_{k_*}} | \mathscr{R}(\tilde{g}^2_{k_*}) \rangle.$$

Acknowledgments. I am greatly indebted to my coauthors of [7] and [12]; the opportunity to work with them has fundamentally shaped my views on nonlinear systems. I am also indebted to my colleagues G. Meyer and J. Lewis of NASA Ames Research Center for their assistance, encouragement and feedback in the writing of this paper.

REFERENCES

[1] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI AND S. MONACO, *Nonlinear decoupling via feedback, a differential geometric approach*, Report 79-19, Istituto di Automatica, Universita di Roma, 1979; IEEE Trans. Automat. Control, AC-26 (1981), pp. 331-345.

[2] R. M. HIRSCHORN, *(A,B) invariant distributions and disturbance, decoupling of nonlinear systems*, this Journal, 19 (1981), pp. 1-19.

[3] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI AND S. MONACO, *Locally (f, g) invariant distributions*, System. Control Lett., 7 (1981), pp. 12-16.

[4] H. NIJMEIJER, *Controlled invariance for affine control systems*, Report BW 133/80, Mathematics Centre, Amsterdam, 1980.

[5] A. J. KRENER, *A heuristic view of nonlinear decoupling*, Proc. 13th Asilomar Conference on Circuits, Systems and Computers, 1979.

[6] A. J. KRENER AND A. ISIDORI, *(Adf, G), invariant and controllability distributions*, in Feedback Control of Linear and Nonlinear Systems, D. Hinrichsen and A. Isidori, eds., Springer-Verlag, Berlin, 1982.

[7] A. ISIDORI, A. J. KRENER, C. GORI-GIORGI AND S. MONACO, *The observability of cascade connected nonlinear systems*, Proc. IFAC World Congress, Tokyo, 1981.

[8] H. NIJMEIJER AND A. VAN DER SCHAFT, *Controlled invariance for nonlinear systems*, Report BW 136/81, Mathematics Centre, Amsterdam, 1981.

[9] A. J. KRENER AND A. ISIDORI, *Nonlinear zero distributions*, Proc. 19th IEEE Conference on Decision and Control, Albuquerque NM, 1980, pp. 665-668.
[10] H. SUSSMANN, *Orbits of families of vector fields and integrability of distributions*, Trans. AMS, 180 (1973), pp. 171-188.
[11] ——, *Existence and uniqueness of minimal realizations of nonlinear systems*, Math. Systems Theory, 10 (1977), pp. 263-284.
[12] R. HERMANN AND A. J. KRENER, *Nonlinear controllability and observability*, IEEE Trans. Automat. Control, AC-22 (1977), pp. 720-740.
[13] J. GONCALVES, *Nonlinear controllability and observability with applications to gradient systems*, Ph.D. Thesis, Warwick, 1981.
[14] H. NIJMEIJER, *Controllability distributions for nonlinear control systems*, Report BW 140/81, Mathematics Centre, Amsterdam, 1981.
[15] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Springer-Verlag, New York, 1974.
[16] R. HERMANN, *The differential geometry of foliations*, II, J. Math. and Mech., 11 (1962), pp. 303-316.
[17] T. NAGANO, *Linear differential systems with singularities and an application to transitive Lie algebras*, J. Math. Soc. Japan, 18 (1966), pp. 388-404.
[18] H. SUSSMANN AND V. JURDJEVIC, *Controllability of nonlinear systems*, J. Differential Equations, 12 (1972), pp. 95-116.
[19] W. M. BOOTHBY, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.
[20] A. J. KRENER, *(f, g) invariant distributions connections and Pontryagin classes*, Proc. 20th IEEE Conference on Decision and Control, San Diego, 1981.
[21] W. L. CHOW, *Über Systems von linearen partiellen Differential Gleichungen ersten Ordinung*, Math. Ann., 117 (1939), pp. 98-105.
[22] M. SPIVAC, *A Comprehensive Introduction to Differential Geometry, Vol.* 1, Publish or Perish Press, Berkeley, CA, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

004

A TRANSFER-FUNCTION APPROACH TO LINEAR TIME-VARYING DISCRETE-TIME SYSTEMS*

E. W. KAMEN†, P. P. KHARGONEKAR†‡ AND K. R. POOLLA†§

Abstract. In the first part of the paper a transfer-function approach is developed for the class of linear time-varying discrete-time systems. The theory is specified in terms of skew (noncommutative) rings of polynomials and formal power series, both with coefficients in a ring of time functions. The transfer-function matrix is defined to be a matrix whose entries belong to a skew ring of formal power series. It is shown that various system properties, such as asymptotic stability, can be characterized in terms of the skew-ring framework. In the last part of the paper, the transfer-function framework is applied to the study of feedback control. New results are obtained on assignability of system dynamics by using dynamic output feedback and dynamic state feedback. The results are applied to the control of an armature-controlled dc motor with a variable loading.

Key words. time-varying systems, difference equations with time varying coefficients, generalized z-transform, transfer function representation, feedback control, pole assignability

1. Introduction.

The polynomial matrix-fraction approach to linear time-invariant systems has turned out to be very useful in studying several system and control theoretic problems, such as realization, dynamic compensation, regulation in the presence of disturbances, etc. For details on this work, the reader may refer to Rosenbrock [1970], Wolovich [1974], Fuhrmann [1976], Rosenbrock and Hayton [1978], Cheng and Pearson [1978], Antsaklis [1979], Desoer et al. [1980], Emre and Silverman [1981], to mention a few.

Given the power of the polynomial matrix-fraction approach, it would be very desirable to have a corresponding theory for linear time-varying systems. In order to generalize the matrix fraction approach to the time-varying case, we need to incorporate time variance into a transform type description of the behavior of the system. Attempts have been made to develop a transfer-function theory for linear time-varying systems (e.g., the system function defined by Zadeh [1950]), but until recently there was no theory which closely resembles the time-invariant case.

In the first part of this paper we develop a transfer-function approach for the class of linear time-varying discrete-time systems. Our framework is given in terms of skew (noncommutative) rings of polynomials, formal power series, and formal Laurent series, all with coefficients in a ring of time functions. The basic elements of the transfer-function approach we are considering here can be found in an unpublished paper of Kamen [1974]; however, a full development of this approach is not carried out in that paper. We also note that our approach differs significantly from the one considered recently by Murray [1982], in which transfer operators are constructed in terms of a crossed product.

After the preliminaries in the next section, in § 3 we define the transfer-function matrix of a linear time-varying discrete-time system. Our framework is developed in

* Received by the editors May 31, 1983, and in revised form July 15, 1984. A preliminary version of the first part of the paper was presented to the 21st IEEE Conference on Decision and Control, Orlando, December 1982. This work was supported in part by the National Science Foundation under grant ECS-8200607.

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

‡ Present address, Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota 55455.

§ Present address, Department of Electrical Engineering, University of Illinois, Urbana, Illinois 61801.

terms of a "generalized z-transform." Using this construct, we show that the z-transform of the output is equal to the product of the z-transform of the input and the transfer function matrix. Hence our setup corresponds very closely to the transfer-function representation of a linear time-invariant system. We also show that the transfer-function matrix of a system specified by state equations has exactly the same form as in the time-invariant case.

In § 4 we show that uniform asymptotic stability can be characterized in terms of the skew-polynomial ring framework. This leads to a spectral radius criterion for uniform asymptotic stability. In § 5 we apply the results of §§ 3 and 4 to the study of dynamic output feedback. Our approach is based on polynomial matrix-fraction representations for the transfer-function matrices of the given system and the feedback compensator. If the polynomial matrix-fraction representation of the given system satisfies a Bezout-type identity, we show that it is possible to assign (up to unimodular matrices) the closed-loop system dynamics. This result is applied to the control of an armature-controlled dc motor with a time-varying motor constant resulting from variable loading.

In § 6 we consider dynamic state feedback for the class of time-varying systems which are reachable in a finite number of steps. It is shown that by using dynamic state feedback, we are able to construct a closed-loop system whose system matrix is algebraically equivalent to a constant matrix with arbitrary assignable eigenvalues. Previously, assignability results for time-varying systems were available only for the very special class of index-invariant time-varying systems (see Wolovich [1968] and Morse and Silverman [1972]), or for the class of cyclizable time-varying systems (see Kamen and Hafez [1979]).

2. Preliminaries. With $\mathbb{Z}$ = set of integers and $\mathbb{R}$ = field of real numbers, let A denote the $\mathbb{R}$-linear space of all functions from $\mathbb{Z}$ into $\mathbb{R}$. With the operations of pointwise addition and pointwise multiplication, it is easy to see that A is a commutative ring with identity 1, where $1(k) = 1$ for all $k \in \mathbb{Z}$. Let σ denote the *right-shift operator* on A defined by

$$(\sigma\alpha)(k) = \alpha(k-1) \quad \text{for all } k \in \mathbb{Z}.$$

Since the shift operator σ is a ring automorphism on A, the ring A is called a *difference ring.*

For any positive integer n, we let $\mathbb{R}^n$ denote the space of n-element column vectors with entries in $\mathbb{R}$. For any vector $x \in \mathbb{R}^n$, the norm of x will be denoted by $\|x\|$, where $\|x\|$ is defined in any one of the usual ways. Given an $n \times n$ matrix M over $\mathbb{R}$, we define the norm $\|M\|$ of M by

$$\|M\| = \{\sup \|Mx\| : x \in \mathbb{R}^n, \|x\| = 1\}.$$

Let A_+ denote the subring of A consisting of all functions with support bounded on the left; that is, for any $\alpha \in A_+$ there is an integer k_α (depending on α in general) such that $\alpha(k) = 0$ for all $k \leqq k_\alpha$. For any positive integer n, let A_+^n denote the $\mathbb{R}$-linear space of all n-element column vectors with entries in A_+.

DEFINITION 2.1. Let m and p be positive integers. An m-input p-output linear time-varying causal input/output map f is an $\mathbb{R}$-linear map

$$f: A_+^m \to A_+^p$$

such that if $u(k) = 0$, $k \leqq k_u$, for some $u \in A_+^m$, then $f(u)(k) = 0$ for all $k \leqq k_u$.

It is well known that for any input/output map f as defined above, there exists a $p \times m$ matrix function $W_f(i,j)$ such that for any $u \in A_+^m$,

$$f(u)(i) = \sum_{j=-\infty}^{i} W_f(i,j)u(j).$$

The matrix function W_f is the *unit-pulse response function* associated with input/output map f. Note that by causality, $W_f(i,j)$ is not defined for $i<j$.

Our next concept is the notion of a system.

DEFINITION 2.2. Let B be a fixed difference subring (i.e., $\sigma(B) = B$) of A containing 1. An m-input p-output n-dimensional linear time-varying system Σ over B is a quadruple $\Sigma = (F, G, H, J)$ of matrices over B where F is $n \times n$, G is $n \times m$, H is $p \times n$, and J is $p \times m$.

With a system $\Sigma = (F, G, H, J)$, we shall associate the dynamical equations

$$x(k+1) = F(k)x(k) + G(k)u(k),$$

$$y(k) = H(k)x(k) + J(k)u(k),$$

where $u(k)$, $x(k)$, $y(k)$ have the usual interpretations. We shall sometimes assume that $J = 0$, in which case we shall write $\Sigma = (F, G, H)$. The assumption that $J = 0$ does not result in any loss of generality.

In the above definition of system, it is important to note that by selecting the difference subring B, we can restrict attention to a particular class of systems. For example, we can study the class of linear time-varying systems with bounded coefficients by taking B to be equal to the set $l^\infty(\mathbb{Z})$ of all bounded functions from $\mathbb{Z}$ into the reals $\mathbb{R}$. It is easy to see that with the induced pointwise operations, $l^\infty(\mathbb{Z})$ is a difference subring of A containing 1.

A system $\Sigma = (F, G, H, J)$ or the pair (F, G) over A is said to be *reachable in N steps* if there is a positive integer N such that for any $j \in \mathbb{Z}$ and any $x \in \mathbb{R}^n$, there exists an input sequence $u(j-N), u(j-N+1), \cdots, u(j-1)$ which drives Σ from the zero state at time $j-N$ to the state x at time j. The system Σ or the pair (F, H) is *observable in $\bar{N}$ steps* if there is a positive integer $\bar{N}$ such that for any $j \in \mathbb{Z}$ and any $x \in \mathbb{R}^n$, the output response $y(\cdot)$ resulting from initial state $x \neq 0$ at time j with $u(k) = 0$ for $k \geqq j$ has the property that $y(k) \neq 0$ for at least one $k \in \{j, j+1, \cdots, j+\bar{N}-1\}$. The system Σ is *canonical* if it is both reachable and observable in N steps.

Weiss [1972] showed that reachability can be characterized in terms of the rank of a matrix function constructed from the matrices F and G: Let $M_0, M_1, \cdots, M_{N-1}$ denote the matrices defined by $M_j = F(\sigma M_{j-1})$, $j = 1, 2, \cdots, N-1$, $M_0 = G$, and let U_k denote the k-step reachability matrix defined by $U_k = [M_0 \; M_1 \cdots M_{k-1}]$. Then the system $\Sigma = (F, G, H, J)$ is reachable in N steps if and only if rank $U_N(k) = n$ for all k in $\mathbb{Z}$. Note that this condition is equivalent to right invertibility of U_N over the ring A. There are similar ("dual") criteria for observability. See Weiss [1972] for the details.

In some cases one may be interested in a slightly different notion of reachability: Suppose that Σ is defined over a difference subring B of A. For example, we could take $B = l^\infty(\mathbb{Z})$. Then Σ is said to be *reachable over B in N steps* for some positive integer N if and only if U_N is right invertible over B. In the example of $B = l^\infty(\mathbb{Z})$, this notion of reachability is equivalent to uniform boundedness of the inputs $u(j-N)$, $u(j-N+1), \cdots$, etc. (uniform with respect to j) in the definition of reachability given above. In fact, in this case, Σ is reachable over $l^\infty(\mathbb{Z})$ in N steps if and only if

$$\det(U_N U_N'(j)) \geqq \varepsilon > 0 \quad \text{for some } \varepsilon > 0 \text{ all } j \text{ in } \mathbb{Z}.$$

The *state transition matrix* of a system $\Sigma = (F, G, H, J)$ is the $n \times n$ matrix function $\Phi(i,j)$ defined by

$$\Phi(i,j) = \begin{cases} F(i-1)F(i-2)\cdots F(j), & i > j, \\ I, & i = j, \\ \text{not defined}, & i < j. \end{cases}$$

As is well known, the *unit-pulse response function* W_Σ of the system Σ is given by

$$W_\Sigma(i,j) = \begin{cases} H(i)\Phi(i,j+1)G(j), & i > j, \\ J(j), & i = j, \\ \text{not defined}, & i < j. \end{cases}$$

The input/output behavior of a system Σ is described by its input/output map f_Σ where

$$f_\Sigma(u)(i) = \sum_{j=-\infty}^{i} W_\Sigma(i,j)u(j), \qquad u \in A_+^m.$$

Given an input/output map $f: A_+^m \to A_+^p$, a *realization* of f is a system $\Sigma = (F, G, H, J)$ over A such that $f_\Sigma = f$.

It is easily seen (and well known) that the system $\Sigma = (F, G, H, J)$ is a realization of f and only if

$$W_f(i,j) = \left.\begin{cases} H(i)\Phi(i,j+1)G(j), & i > j \\ J(j), & i = j, \end{cases}\right\} = W_\Sigma(i,j),$$

where W_f is the unit-pulse response function associated with f. For results on realizability, we refer the reader to Weiss [1972], Evans [1972], Ferrer and Kamen [1984].

3. The transfer-function framework. The commutative ring of polynomials, the commutative ring of formal power series, and the commutative ring of formal Laurent series, all with coefficients in the reals $\mathbb{R}$, play a central role in the transfer-function approach to linear time-invariant systems. For linear time-varying systems, the analogous algebraic objects are noncommutative rings of polynomials, formal power series, and formal Laurent series, all with coefficients in a commutative ring of time functions. The definitions of these rings are given below.

As before, let B be a fixed difference subring of A containing 1. With z equal to a symbol (or indeterminate), let $B((z^{-1}))$ denote the set of all formal Laurent series of the form

$$\sum_{r=-N}^{\infty} z^{-r}\alpha_r, \quad N \in \mathbb{Z}, \quad \alpha_r \in B.$$

With the usual addition, and with multiplication defined by

$$z^r z^t = z^{r+t}, \qquad r, t \in \mathbb{Z},$$

$$\alpha z^r = z^r(\sigma^r \alpha), \qquad r \in \mathbb{Z}, \quad \alpha \in B,$$

where $(\sigma^r\alpha)(k) = \alpha(k-r)$, $B((z^{-1}))$ is a noncommutative ring with identity. The ring $B((z^{-1}))$ is called the *skew ring of formal Laurent series over B with coefficients written on the right.* Here the term "skew" means that the coefficients of the series do not commute with the indeterminate. There are two important subrings of $B((z^{-1}))$: the skew ring of polynomials $B[z]$ defined by

$$B[z] = \left\{ \alpha \text{ in } B((z^{-1})) \colon \alpha = \sum_{r=-N}^{0} z^{-r}\alpha_r, N \geqq 0 \right\},$$

and the skew ring of formal power series $B[[z^{-1}]]$ defined by

$$B[[z^{-1}]] = \left\{ \alpha \text{ in } B((z^{-1})) : \alpha = \sum_{r=0}^{\infty} z^{-r}\alpha_r \right\}.$$

Define a projection map

$$\pi : B((z^{-1})) \to B[[z^{-1}]] : \sum_{r=-N}^{\infty} z^{-r}\alpha_r \to \sum_{r=1}^{\infty} z^{-r}\alpha_r.$$

We denote $\alpha - \pi(\alpha)$ by $(\alpha)_+$ for any α in $B((z^{-1}))$. By $(\alpha)_0$, we shall mean the constant coefficient of α in $B((z^{-1}))$. For an element α in $B((z^{-1}))$, α is said to be *strictly proper* if $\pi(\alpha) = \alpha$, and $\pi(\alpha)$ is called the *strictly proper part* of α.

Given a positive integer r, we shall let $B^{r\times r}[z]$ (resp. $B^{r\times r}((z^{-1}))$) denote the ring of $r \times r$ matrices over the skew polynomial ring $B[z]$ (resp. the skew Laurent series ring $B((z^{-1}))$). Any Q in $B^{r\times r}[z]$ can be written in the form

$$Q = \sum_{i=0}^{q} z^i Q_i,$$

where the Q_i are $r \times r$ matrices over B. The *degree of* Q, denoted by deg Q, is the largest integer i such that $Q_i \neq 0$. If deg $Q = q$, Q is said to be *monic* if $Q_q = I = r \times r$ identity matrix. We have the following result on invertibility.

PROPOSITION 3.1. *The matrix Q in $A^{r\times r}[z]$ is right-invertible over $A^{r\times r}((z^{-1}))$ if and only if there exists a matrix T in $A^{r\times r}[z]$ such that QT is a monic polynomial matrix in $A^{r\times r}[z]$. Further, T can be chosen such that* deg Q = deg QT.

Proof. Suppose Q is right-invertible, i.e., there exists a Ψ in $A^{r\times r}((z^{-1}))$ such that $Q\Psi = I$. Let deg $Q = d$. We can now write

$$z^d I = Q\Psi z^d = Q(\Psi z^d)_+ + Q\pi(\Psi z^d).$$

Notice that deg $(Q\pi(\Psi z^d)) <$ deg $Q = d$. Therefore the highest degree term of $Q(\Psi z^d)_+$ is $z^d I$. Choosing $T = (\Psi z^d)_+$ which is *polynomial*, proves the necessity. Notice that deg QT = deg $Q = d$. Now assume that there exists a T in $A^{r\times r}[z]$ such that QT is a monic $r \times r$ polynomial matrix. We can then perform right division of I by QT and find a ϕ in $A^{r\times r}((z^{-1}))$ such that $QT\phi = I$, which implies that Q is right invertible over $A^{r\times r}((z^{-1}))$. This proves the converse. □

An analogous result holds for left-invertibility of Q over $A^{r\times r}((z^{-1}))$: Q is left-invertible if and only if there exists a T_1 in $A^{r\times r}[z]$ such that $T_1 Q$ is a monic polynomial matrix, in which case T_1 can be chosen such that deg $(T_1 Q)$ = deg (Q). In general, left-invertibility of Q over $A^{r\times r}((z^{-1}))$ is not equivalent to right-invertibility of Q over $A^{r\times r}((z^{-1}))$, the pathology being due to the skew nature of our rings.

We shall almost always deal with polynomial matrices that are both left- and right-invertible, in which case we shall call them *invertible* and avoid the use of cumbersome prefixes.

Our transfer-function approach is derived in terms of a generalized z-transform defined as follows. As in the previous section, let A_+ denote the subring of A consisting of all functions $\alpha : \mathbb{Z} \to \mathbb{R}$ with support bounded on the left. Let Δ denote the unit-pulse function concentrated at the origin, that is, $\Delta(k) = 1$ when $k = 0$ and $\Delta(k) = 0$ for all other k. Now given $f \in A_+$, the *generalized z-transform* of f is defined to be the series $F(z) \in A((z^{-1}))$ given by

$$F(z) = \sum_r z^{-r}\gamma_r \tag{3.2}$$

where $\gamma_r = f(r)\Delta$; that is, γ_r is the pulse at the origin with magnitude $f(r)$. Note that we can rewrite (3.2) in the form

$$F(z) = \left[\sum_r f(r)z^{-r}\right]\Delta,$$

where $\sum_r f(r)z^{-r} \in \mathbb{R}((z^{-1}))$ is the (ordinary) z-transform of f. In other words, we can view the generalized z-transform as the product in the skew-ring structure of the ordinary z-transform and the unit-pulse function Δ.

The transform map $f \to F(z)$ has the following fundamental properties (the easy proof is omitted).

PROPOSITION 3.3. *The map $A_+ \to A((z^{-1})) : f \to F(z)$ is a left-A-module homomorphism; that is, if $f \in A_+$ and $g \in A_+$ have transforms $F(z)$ and $G(z)$, then for any $\alpha, \beta \in A$, the transform of $\alpha f + \beta g$ is equal to $\alpha F(z) + \beta G(z)$. In addition, for any positive integer r, the transform of $\sigma^r f$ is $z^{-r}F(z)$ and the transform of $\sigma^{-r}f$ is $z^r F(z)$.*

Now let f be a m-input p-output map as defined in the previous section, and let W_f denote the unit-pulse response function associated with f. For each integer $r \geqq 0$, let W_r denote the $p \times m$ matrix function on $\mathbb{Z}$ defined by

$$W_r(j) = W_f(r+j, j), \qquad j \in \mathbb{Z}.$$

We assume that each W_r is over a difference subring B of A containing 1 (we could have $B = A$).

DEFINITION 3.4. The (formal) transfer-function matrix $W_f(z)$ associated with the input/output map f is the $p \times m$ matrix over $B[[z^{-1}]]$ defined by

$$W_f(z) = \sum_{r=0}^{\infty} z^{-r} W_r.$$

The input/output equation $y = f(u)$ can be characterized in terms of the transfer matrix $W_f(z)$ as follows.

PROPOSITION 3.5. *Let $y = f(u)$ be the output resulting from input u, and let $Y(z)$ and $U(z)$ denote the (generalized) z-transforms of y and u taken component-by-component. Then*

$$Y(z) = W_f(z)U(z). \tag{3.6}$$

Proof. By definition of multiplication in $A((z^{-1}))$, we have

$$W_f(z)U(z) = \sum_r \sum_k z^{-r-k}(\sigma^{-k}W_r)u_k.$$

By a change of variables, we get

$$W_f(z)U(z) = \sum_r z^{-r}\gamma_r,$$

where

$$\gamma_r = \sum_k (\sigma^{-k}W_{r-k})u_k.$$

By definition of W_r and u_r, we have that

$$\gamma_r(j) = \begin{cases} \sum_k W(r,k)u(k), & j = 0, \\ 0, & j \neq 0. \end{cases}$$

Thus $\gamma_r(0) = f(u)(r) = y(r)$, and the result is proved. □

Note how closely (3.6) corresponds to the transfer-function relationship in the time-invariant case. This analogy with the time-invariant case is also seen when one computes the transfer function of a system. To illustrate this we first need the following result on the invertibility of $(zI-F)$.

PROPOSITION 3.7. *Let $\Sigma=(F, G, H, J)$ be a system over B with state transition matrix $\Phi(i,j)$. Then the matrix $(zI-F)$ has a (unique) inverse $(zI-F)^{-1}$ over the skew ring $B[[z^{-1}]]$ given by*

$$(zI-F)^{-1}=\sum_{i=1}^{\infty} z^{-i}(\sigma^{-1}\Phi_{i-1}), \tag{3.8}$$

where $\Phi_i(j)=\Phi(i+j,j)$.

Proof. By definition of multiplication in $B[[z^{-1}]]$, we have

$$(zI-F)\left(\sum_{i=1}^{\infty} z^{-i}(\sigma^{-1}\Phi_{i-1})\right)=I+\sum_{i=1}^{\infty} z^{-i}(\sigma^{-1}\Phi_i-(\sigma^{-i}F)(\sigma^{-1}\Phi_{i-1})).$$

It is easily shown that $\Phi_i=(\sigma^{-i+1}F)\Phi_{i-1}$ for $i\geqq 1$, and thus

$$(zI-F)\left(\sum_{i=1}^{\infty} z^{-i}(\sigma^{-1}\Phi_{i-1})\right)=I.$$

Since F is over B, all the Φ_i are over B, and thus $(zI-F)$ has an inverse over $B[[z^{-1}]]$. □

It is important to note that by definition of multiplication in the skew ring $B((z^{-1}))$, $(zI-F)^{-1}$ is not in general equal to the power series $\sum_i z^{-i}F^{i-1}$. In fact, this power series is equal to the inverse of $(zI-F)$ if and only if F is constant (i.e., $F(k_1)=F(k_2)$ for all $k_1, k_2\in\mathbb{Z}$).

PROPOSITION 3.9. *Let $\Sigma=(F, G, H, J)$ be a system over B with input/output map f_Σ. Then the transfer-function matrix $W_\Sigma(z)$ of f_Σ is defined over the skew ring $B[[z^{-1}]]$ and is given by*

$$W_\Sigma(z)=J+H(zI-F)^{-1}G. \tag{3.10}$$

Proof. Using Proposition 3.7, we have

$$\begin{aligned} J+H(zI-F)^{-1}G &= J+\sum_{i=1}^{\infty} Hz^{-i}(\sigma^{-1}\Phi_{i-1})G \\ &= J+\sum_{i=1}^{\infty} z^{-i}(\sigma^{-i}H)(\sigma^{-1}\Phi_{i-1})G. \end{aligned}$$

An easy computation reveals that $W_i=(\sigma^{-i}H)(\sigma^{-1}\Phi_{i-1})G$ for $i\geqq 1$, and since $W_0=J$ we have the desired result. □

By Proposition 3.9 we see that our notion of transfer function in the time-varying case has exactly the same form as the transfer function in the time-invariant case. However, it is important to note that the evaluation of the transfer-function expression is not the same for both the time-varying and time-invariant cases; for example, here the right side of (3.10) is computed using the noncommutative multiplication in the skew ring $B[[z^{-1}]]$.

We shall now relate the transfer-function framework to a representation consisting of the following collection of input/output difference equations with time-varying coefficients:

$$\sum_{i=0}^{q} Q_i(k+i)w(k+i)=\sum_{i=0}^{r} R_i(k+i)u(k+i), \tag{3.11a}$$

$$y(k)=\sum_{i=0}^{s} P_i(k+i)w(k+i). \tag{3.11b}$$

In (3.11a, b), $u\in A_+^m$, $w\in A_+^r$, $y\in A_+^p$, and the Q_i, R_i, P_i are $r\times r$, $r\times m$, $p\times r$ matrices over A. Taking the generalized z-transform of both sides of (3.11a, b) and using Proposition 3.3, we get the following polynomial matrix representation:

$$Q(z)W(z)=R(z)U(z), \tag{3.12a}$$

$$Y(z)=P(z)W(z), \tag{3.12b}$$

where $Q(z)$, $R(z)$, and $P(z)$ are polynomial matrices over the skew ring $A[z]$ given by

$$Q(z)=\sum_{i=0}^{q} z^iQ_i, \quad R(z)=\sum_{i=0}^{r} z^iR_i, \quad P(z)=\sum_{i=0}^{s} z^iP_i.$$

This leads to the following result.

PROPOSITION 3.13. *Suppose that $Q(z)$ has a left inverse $Q^{-1}(z)\in A^{r\times r}((z^{-1}))$. Then for any $u\in A_+^m$ with zero initial conditions before the application of u (i.e., if $u(k)=0$, $k<k_0$, then we assume that $w(k)=0$ and $y(k)=0$, $k<k_0$), (3.11a, b) has a unique solution $y\in A_+^p$ with the z-transform $Y(z)$ of y given by*

$$Y(z)=P(z)Q^{-1}(z)R(z)U(z). \tag{3.14}$$

Further, if $P(z)Q^{-1}(z)R(z)$ is over the skew ring $A[[z^{-1}]]$ of formal power series, the map $A_+^m\to A_+^p: u\to y$, where $Y(z)$ is given by (3.14), is a causal input/output map.

Proof. The result follows easily from the properties of the transform, and thus will be omitted. □

If $Q(z)$ has left inverse $Q^{-1}(z)$ and $P(z)Q^{-1}(z)R(z)$ is over $A[[z^{-1}]]$, by Proposition 3.13 we see that the input/output difference equation representation (3.11a, b) defines an input/output map f with transfer-function matrix

$$W_f(z)=P(z)Q^{-1}(z)R(z).$$

Conversely, suppose that we are given an input/output map f with transfer-function matrix $W_f(z)$. Then if $W_f(z)$ can be written in the form

$$W_f(z)=P(z)Q^{-1}(z)R(z) \tag{3.15}$$

for some polynomial matrices $P(z)$, $Q(z)$, $R(z)$ defined over the skew ring $A[z]$, we see that f defines a collection of input/output difference equations given by (3.11a, b). Thus there is a direct correspondence between input/output difference equations of the form (3.11a, b) and matrix-fraction representations of the form (3.15). For linear time-invariant systems, this was observed by Rosenbrock [1970].

Example 3.16. Consider an armature-controlled dc motor given by the input/output differential equation

$$\frac{d^2\theta(t)}{dt^2}+\alpha(t)\frac{d\theta(t)}{dt}=\beta(t)u(t). \tag{3.17}$$

In (3.17), the input $u(t)$ is the applied armature voltage, the output $\theta(t)$ is the angular position of the motor shaft, and the time-varying coefficients $\alpha(t)$ and $\beta(t)$ are given by

$$\alpha(t)=0.74+0.3k_M(t), \qquad \beta(t)=5.56k_M(t),$$

where $k_M(t)$ is the (normalized) effective motor constant. The nominal value of k_M is 1, but during operation k_M may vary significantly due to a variable loading applied

to the motor shaft. We assume that the time variation of $k_M(t)$ is known prior to system operation; for example, the motor may be part of a machining operation whose characteristics can be measured (e.g., in test runs). By sampling with a suitably small sampling period T, we get the following approximation to a sampled-data state model for the motor:

$$\begin{bmatrix} \theta(kT+T) \\ \dot{\theta}(kT+T) \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1-\alpha(kT)T \end{bmatrix} \begin{bmatrix} \theta(kT) \\ \dot{\theta}(kT) \end{bmatrix} + \begin{bmatrix} \beta(kT)(T^2/2) \\ \beta(kT)T \end{bmatrix} u(kT). \tag{3.18}$$

Taking the generalized z-transform of both sides of (3.18), we get

$$z\Theta(z) = \Theta(z) + T\Gamma(z) + \beta\frac{T^2}{2}U(z),$$

$$z\Gamma(z) = (1-\alpha T)\Gamma(z) + \beta T U(z),$$

where $\Theta(z)$, $\Gamma(z)$, and $U(z)$ are the generalized z-transforms of the angular position $\theta(kT)$, angular velocity $\dot{\theta}(kT)$, and input $u(kT)$, respectively. Combining these two transformed equations and keeping in mind the skew multiplication in the ring $A[z]$, we have

$$Q(z)\Theta(z) = R(z)U(z),$$

where

$$Q(z) = z^2 + (\alpha T - 2)z + (1-\alpha T),$$

$$R(z) = z\beta\frac{T^2}{2} + \beta\frac{T^2}{2}(1+\alpha T).$$

Thus, the transfer-function relationship for the motor is

$$\Theta(z) = Q^{-1}(z)R(z)U(z), \tag{3.19}$$

where $W(z) = Q^{-1}(z)R(z)$ is the (generalized) transfer function of the motor. In § 5, we shall use the transfer-function representation (3.19) to design a stabilizing dynamic output feedback compensator for the motor.

4. Stability. In this section we show that it is possible to characterize uniform asymptotic stability in terms of the skew-ring structure defined in § 3. We begin with the definition of stability.

Given the system $\Sigma = (F, G, H, J)$ defined over the difference ring A, consider the free behavior of the system described by the vector difference equation

$$x(k+1) = F(k)x(k). \tag{4.1}$$

The system Σ is said to be *uniformly asymptotically stable* (u.a.s.) if for every real number $\varepsilon > 0$, there exists a positive integer n_ε such that for all j in $\mathbb{Z}$

$$\|x(j)\| \leqq 1 \text{ implies that } \|x(j+i)\| \leqq \varepsilon \text{ for all } i \geqq n_\varepsilon,$$

where $x(j+i)$ is the solution of (4.1) at time $j+i$ starting from initial state $x(j)$. It is well known that Σ is u.a.s. if and only if for any $\varepsilon > 0$ there is a positive integer n_ε such that

$$\|\Phi(j+i, j)\| \leqq \varepsilon \quad \text{for all } j \in \mathbb{Z} \text{ and all } i \geqq n_\varepsilon,$$

where $\Phi(i, j)$ is the state transition matrix.

As in Proposition 3.7, for $i = 0, 1, 2, \cdots$, let Φ_i denote the $n \times n$ matrix with entries in A defined by $\Phi_i(j) = \Phi(i+j, j)$, $j \in \mathbb{Z}$. By definition of $\Phi(i, j)$, we have that

$$\Phi_0 = I \quad \text{and} \quad \Phi_i = (\sigma^{-i+1}F)(\sigma^{-i+2}F) \cdots (\sigma^{-1}F)(F) \quad \text{for } i \geqq 1. \tag{4.2}$$

Now suppose that F is over the difference subring $l^\infty(\mathbb{Z})$ consisting of all bounded functions from $\mathbb{Z}$ into $\mathbb{R}$. Then by (4.2), Φ_i is over $l^\infty(\mathbb{Z})$ for all $i \geqq 0$. Defining the norm $\|\Phi_i\|$ by

$$\|\Phi_i\| = \sup_{j \in \mathbb{Z}} \|\Phi_i(j)\|,$$

we have the following result.

PROPOSITION 4.3. *Suppose that F is an $n \times n$ matrix over $l^\infty(\mathbb{Z})$. Then Σ is u.a.s. if and only if $\|\Phi_i\|$ converges (in $\mathbb{R}$) to zero as $i \to \infty$.*

The proof of this result follows immediately from the above characterization of u.a.s. given in terms of the state transition matrix. (We should note that it is possible to remove the constraint that F be bounded. For details, see Green and Kamen [1984].)

Now suppose that F is constant, that is, F is an $n \times n$ matrix over the reals $\mathbb{R}$. In this case it is well known that Σ is u.a.s. if and only if the matrix norm of the coefficients of the formal power series

$$(zI - F)^{-1} = \sum_{i=1}^{\infty} z^{-i}F^{i-1} = \sum_{i=1}^{\infty} F^{i-1}z^{-i}$$

converges to zero. In other words, $\|F^i\| \to 0$ as $i \to \infty$. As we now show, there is a generalization of this result for time-varying systems. First we need to define what is meant by a stable power series.

Let $P(z)$ denote an $n \times n$ matrix over the skew ring $l^\infty(\mathbb{Z})((z^{-1}))$; that is, $P(z)$ is a matrix Laurent series given by

$$P(z) = \sum_{i=-N}^{\infty} z^{-i}P_i,$$

where the P_i are $n \times n$ matrices over $l^\infty(\mathbb{Z})$. The matrix Laurent power series $P(z)$ is said to be *stable* if $\|P_i\| \to 0$ as $i \to \infty$. In terms of this notion, we have the following criterion for u.a.s.

PROPOSITION 4.4. *Suppose that F is over $l^\infty(\mathbb{Z})$. Then Σ is u.a.s. if and only if the matrix power series $(zI - F)^{-1}$ is stable.*

Proof. Combine Propositions 3.7 and 4.3. □

In the remainder of this section, we show that u.a.s. is equivalent to a spectral radius criterion for a bounded linear operator on a Banach space.

Let $l^\infty(\mathbb{Z})^n$ denote the $\mathbb{R}$-linear space of n-element column vectors with entries in $l^\infty(\mathbb{Z})$. With the norm,

$$\|v\| = \sup_{j \in \mathbb{Z}} \|v(j)\|,$$

$l^\infty(\mathbb{Z})^n$ is a Banach space. Now given an $n \times n$ matrix F over $l^\infty(\mathbb{Z})$, let S_F denote the $\mathbb{R}$-linear map from $l^\infty(\mathbb{Z})^n$ into itself defined by

$$S_F(v)(k) = (\sigma F)(\sigma v)(k) = F(k-1)v(k-1).$$

It is easy to show that S_F is a bounded linear map; in fact, the norm $\|S_F\|$ of S_F is equal to $\|F\|$, where $\|F\| = \sup_{j \in \mathbb{Z}} \|F(j)\|$.

The map S_F arises in the algebraic theory of linear time-varying discrete-time systems as shown by Kamen and Hafez [1979]. Here our objective is to show that u.a.s.

can be characterized in terms of the spectral radius $\rho(S_F)$ of S_F defined by

$$\rho(S_F) = \lim_{i\to\infty} \|S_F^i\|^{1/i} = \inf_i \|S_F^i\|^{1/i}.$$

THEOREM 4.5. *Suppose that F is over $l^\infty(\mathbb{Z})$. Then Σ is u.a.s. if and only if $\rho(S_F) < 1$.*

Proof. By induction on i, it can be shown that

$$S_F^i(v) = (\sigma^i \Phi_i)(\sigma^i v) \quad \text{for all } i \geqq 1 \text{ and any } v \in l^\infty(\mathbb{Z})^n,$$

where Φ_i is given by (4.2). Since the right-shift operator σ is an isometry, it follows that

$$\|S_F^i\| = \|\Phi_i\| \quad \text{for all } i \geqq 1.$$

Now suppose that Σ is u.a.s. Then there exists a positive integer q such that $\|\Phi_q\| < 1$, and thus $\|S_F^q\| < 1$, which implies that $\|S_F^q\|^{1/q} < 1$. Hence $\rho(S_F) < 1$. Conversely, suppose that $\rho(S_F) < 1$. Then there exists a positive integer q such that $\|S^q\|^{1/q} < 1$, which implies that $\|S_F^q\| < 1$. Thus, $\|S_F^q\|^i \to 0$ as $i \to \infty$, which implies that $\|S_F^i\| \to 0$ as $i \to \infty$. Hence $\|\Phi_i\| \to 0$, and by Proposition 4.3 the system is u.a.s. □

A very interesting consequence of Theorem 4.5 is that uniform asymptotic stability in the time-varying finite-dimensional case is equivalent to asymptotic stability of the *time-invariant* infinite-dimensional system

$$\gamma(k+1) = S_F \gamma(k), \quad k \geqq 0, \quad \gamma(0) \in l^\infty(\mathbb{Z})^n.$$

This correspondence has been exploited by Green and Kamen [1984] to obtain new results on the stability of linear time-varying systems.

5. Application to feedback control. In this section we apply the transfer-function framework to the study of output feedback. In particular, we obtain a result on a type of "assignability" by using dynamic output feedback.

Given an m-input p-output system $\Sigma = (F, G, H)$ over $l^\infty(\mathbb{Z})$, recall from § 3 that the input/output transfer-function relationship of the system is given by

$$Y(z) = W_\Sigma(z) U(z),$$

where $W_\Sigma(z) = H(zI - F)^{-1}G$ is the system's transfer-function matrix, and $U(z)$, $Y(z)$ are the generalized z-transforms of the system's input $u(k)$ and output $y(k)$. Suppose that $W_\Sigma(z)$ has the matrix-fraction representation $W_\Sigma(z) = Q^{-1}(z)R(z)$, where $Q(z)$ is an invertible element of $l^\infty(\mathbb{Z})^{p\times p}[z]$ and $R(z)$ is an element of $l^\infty(\mathbb{Z})^{p\times m}[z]$. Now consider the closed-loop system given by

$$\begin{aligned} Y(z) &= W_\Sigma(z) U(z), \\ U(z) &= -W_c(z) Y(z) + V(z), \end{aligned} \tag{5.1}$$

where $W_c(z)$ is the transfer-function matrix of a (possibly) time-varying feedback compensator and $V(z)$ is the generalized z-transform of a possible external input $v(k)$. We assume that $W_c(z)$ has the matrix-fraction representation $W_c(z) = P_c(z)Q_c^{-1}(z)R_c(z)$, where $P_c(z)$ and $R_c(z)$ are $m \times q$, $q \times p$ matrices over $l^\infty(\mathbb{Z})[z]$, and $Q_c(z)$ is a $q \times q$ invertible matrix with entries in $l^\infty(\mathbb{Z})[z]$. We can then represent the closed-loop system given by (5.1) in the form

$$\begin{bmatrix} Q(z) & -R(z)P_c(z) \\ -R_c(z) & Q_c(z) \end{bmatrix} \begin{bmatrix} Y(z) \\ X(z) \end{bmatrix} = \begin{bmatrix} R(z) \\ 0 \end{bmatrix} V(z), \tag{5.2}$$

where

$$X(z) = Q_c^{-1}(z) R_c(z) Y(z).$$

From (5.2), we see that the closed-loop system dynamics are determined by the inverse of the matrix

$$\Pi(z) := \begin{bmatrix} Q(z) & -R(z)P_c(z) \\ -R_c(z) & Q_c(z) \end{bmatrix}.$$

In particular, from the results in § 4, it follows that the closed-loop system given by (5.2) is internally u.a.s. if $\Pi(z)$ has an inverse $\Pi^{-1}(z)$ over $l^\infty(\mathbb{Z})((z^{-1}))$ and $\Pi^{-1}(z)$ is a stable matrix series (as defined in § 4). Thus, a sufficient condition (which turns out to be necessary also) for the compensator $W_c(z) = P_c(z)Q_c^{-1}(z)R_c(z)$ to be stabilizing is that $\Pi^{-1}(z)$ be a stable matrix series.

Instead of asking for a compensator $W_c(z)$ which results in a stable matrix $\Pi^{-1}(z)$, we can attempt to answer the question as to the extent to which the matrix $\Pi^{-1}(z)$ (or the matrix $\Pi(z)$) can be "assigned" by choosing the compensator $W_c(z)$. In the time-invariant case (with $R_c(z) = I$), we have that

$$\det \Pi(z) = \det (Q(z)Q_c(z) - R(z)P_c(z)),$$

and we can consider assigning the coefficients of the polynomial $\det \Pi(z)$ by choosing $P_c(z)$ and $Q_c(z)$ (with the constraint that $Q_c(z)$ is invertible and $P_c(z)Q_c^{-1}(z)$ is proper). This problem was solved by Rosenbrock and Hayton [1978]. But in the time-varying case, there is no known definition of $\det \Pi(z)$. In the time-varying case, we can still consider the extent to which $\Pi^{-1}(z)$ can be assigned by choosing $W_c(z)$. One such result is given below.

THEOREM 5.3. *Suppose that $W_\Sigma(z)$ has a polynomial matrix-fraction representation $W_\Sigma(z) = Q^{-1}(z)R(z)$ which satisfies the Bezout-type identity $Q(z)Y_1(z) + R(z)Y_2(z) = I$ for some $p \times p$, $m \times p$ matrices $Y_1(z)$, $Y_2(z)$ with entries in $l^\infty(\mathbb{Z})[z]$. Let $V(z)$ be a monic $p \times p$ matrix over $l^\infty(\mathbb{Z})[z]$ with* $\deg V(z) \geqq 2 \deg Q(z) + \deg Y_2(z)$. *Then there exists a compensator with strictly proper transfer-function matrix $W_c(z) = P_c(z)Q_c^{-1}(z)R_c(z)$ and unimodular polynomial matrices $T_1(z)$, $T_2(z)$ such that*

$$\Pi^{-1}(z) = T_1(z) \begin{bmatrix} I & 0 \\ 0 & V^{-1}(z) \end{bmatrix} T_2(z). \tag{5.4}$$

Since $T_1(z)$ and $T_2(z)$ are unimodular matrices (having inverses over $l^\infty(\mathbb{Z})[z]$), the expression (5.4) for $\Pi^{-1}(z)$ implies that the closed-loop system dynamics are determined by $V^{-1}(z)$, which can be arbitrarily assigned. Note that if we choose $V(z)$ so that $V^{-1}(z)$ is a stable matrix series, the resulting closed-loop system will be u.a.s. In particular, if we choose $V(z) = z^r I_p$, the closed-loop system response resulting from any nonzero initial state will become zero after a finite number of steps (assuming there is no external input); in other words, we have a "dead-beat" control system.

We shall sketch the key steps for a constructive proof of Theorem 5.3. Here we assume that $Q(z)$ is a monic polynomial matrix; for the nonmonic case, see Khargonekar and Poolla [1984].

Step 1. First find polynomial matrices $Q(z)$, $R(z)$, $Y_1(z)$, $Y_2(z)$ such that

$$\begin{aligned} &W_\Sigma(z) = Q^{-1}(z)R(z), \\ &Q(z)Y_1(z) + R(z)Y_2(z) = I. \end{aligned} \tag{5.5}$$

It can be shown that $W_\Sigma(z)$ admits a polynomial matrix-fraction representation $Q^{-1}(z)R(z)$ which satisfies the Bezout-type identity (5.5) if and only if $W_\Sigma(z)$ has a canonical realization (i.e., a realization which is both reachable and observable in N steps for some positive integer N). For a proof of this fact, see Khargonekar and

Poolla [1984]. It can also be shown that the polynomial matrices $Y_1(z)$ and $Y_2(z)$ can be computed by solving a system of linear equations with coefficients in $l^\infty(\mathbb{Z})$, or one could follow the procedure given in Khargonekar and Poolla [1984]. From here on, we assume that $Q(z)$ is monic.

Step 2. Choose any $p \times p$ monic polynomial matrix $V(z)$ such that $\deg V(z) \geqq 2 \deg Q(z) + \deg Y_2(z)$ and such that $V^{-1}(z)$ is stable. Divide $V(z)$ by $Q(z)$ to obtain

$$V(z) = M(z)Q(z) + N(z),$$

with $\deg N(z) < \deg Q(z)$.

Step 3. Take the compensator transfer function $W_c(z)$ to be

$$W_c(z) = -Y_2(z)(M(z) + N(z)Y_1(z))^{-1}N(z).$$

With $W_c(z)$ as defined above, it follows that $\Pi^{-1}(z)$ can be expressed in the form (5.4) for some unimodular matrices $T_1(z)$, $T_2(z)$.

Step 4. Realize the compensator by a state model $\Sigma_c = (F_c, G_c, H_c)$ as follows. First, we can write

$$M(z) + N(z)Y_1(z) = \sum_{i=0}^{r} z^i Q_i, \qquad Q_r = I,$$

$$z^r(M(z) + N(z)Y_1(z))^{-1} = \sum_{i=0}^{\infty} z^{-i} A_i,$$

$$-N(z) = \sum_{i=0}^{r-1} z^i R_i,$$

$$Y_2(z) = \sum_{i=0}^{r-1} P_i z^i,$$

where the Q_i, A_i, R_i, P_i are matrices over $l^\infty(\mathbb{Z})$. Then $\Sigma_c = (F_c, G_c, H_c)$ is a realization of $W_c(z)$ where

$$F_c = \begin{bmatrix} -Q_{r-1} & I & 0 & \cdots & 0 & 0 \\ -Q_{r-2} & 0 & I & & \vdots & \vdots \\ \vdots & \vdots & & & & \\ -Q_1 & 0 & 0 & \cdots & & I \\ -Q_0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \qquad G_c = \begin{bmatrix} R_{r-1} \\ R_{r-2} \\ \vdots \\ R_0 \end{bmatrix}, \tag{5.6}$$

$$H_c = [P_0 \quad P_1 \quad \cdots \quad P_{r-1}] \begin{bmatrix} I & 0 & \cdots & 0 & 0 \\ \sigma^r(A_1) & I & & 0 & 0 \\ \sigma^r(A_2) & \sigma^{r-1}(A_1) & & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ \sigma^r(A_{r-2}) & \sigma^{r-1}(A_{r-3}) & \cdots & I & 0 \\ \sigma^r(A_{r-1}) & \sigma^{r-1}(A_{r-2}) & \cdots & \sigma(A_1) & I \end{bmatrix}. \tag{5.7}$$

Example 5.8. Again consider the armature-controlled dc motor whose transfer function $W_c(z) = Q^{-1}(z)R(z)$ was derived in Example 3.16. By inspection, we have

$$Q(z)Y_1(z) + R(z)Y_2(z) = 1,$$

where

$$Y_1(z) = (\sigma\beta)\frac{T^2}{2c}, \qquad Y_2(z) = -z\left(\frac{1}{c}\right) + 3\frac{\sigma\beta}{\beta c}, \qquad c = (\sigma\beta)T^2(2 + T\alpha).$$

Carrying out Step 2 in the above procedure, if we choose $V(z)=z^4$, we have

$$z^4 = M(z)Q(z)+N(z),$$

where

$$M(z) = z^2 - z(\sigma^{-1}(\alpha T-2))+\lambda,$$
$$N(z) = (\sigma^{-2}(\alpha T-2)\sigma^{-1}(1-\alpha T)-(\alpha T-2)\lambda)z-(1-\alpha T)\lambda,$$
$$\lambda = \sigma^{-2}(\alpha T-2)\sigma^{-1}(\alpha T-2)-\sigma^{-2}(1-\alpha T).$$

The compensator (dead-beat controller) is then given by

$$W_c(z) = -Y_2(z)(M(z)+N(z)Y_1(z))^{-1}N(z).$$

A state model for this compensator can be computed directly from (5.6), (5.7). We leave the details to the interested reader.

6. State feedback. In this section we show that by using dynamic state feedback it is possible to assign the coefficients of a characteristic polynomial which determines the closed-loop system dynamics. This result is a time-varying analogue of the famous result that reachability implies coefficient assignability in linear time-invariant systems.

Let $\Sigma=(F, G, H, J)$ be reachable in N steps over B, an arbitrary difference subring of A. Then, there exist matrices $P_0, P_1, \cdots, P_{N-1}$ over B such that

$$\sum_{i=0}^{N-1} M_i P_i = I$$

where the matrices M_i are defined recursively in § 2. Let $S_i=\sigma^{-i}(P_i)$. Consider the linear time-varying system $\Sigma_1=(A, B, C)$ defined as follows:

$$z(k+1) = \begin{bmatrix} 0 & I & 0 & \cdots & 0 \\ 0 & 0 & I & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & I \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} z(k) + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ I \end{bmatrix} v(k), \tag{6.1}$$

$$u(k) = [S_0 \quad S_1 \quad \cdots \quad S_{N-1}](k)z(k). \tag{6.2}$$

Here, $z(k)$ is in $\mathbb{R}^{nN}$, each block submatrix is $n\times n$, $u(k)$ is the input to Σ, and $v(k)$ is (yet) unspecified input to Σ_1. Combining the state equations for Σ with those of Σ_1, and using the control law (6.2), we have

$$\begin{bmatrix} x(k+1) \\ z(k+1) \end{bmatrix} = \underbrace{\begin{bmatrix} F & GC \\ 0 & A \end{bmatrix}}_{\hat{F}}(k)\begin{bmatrix} x(k) \\ z(k) \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ B \end{bmatrix}}_{\hat{G}} v(k). \tag{6.3}$$

Let us consider the $(N+1)$ step reachability matrix U_{N+1} of the pair $(\hat{F}, \hat{G})$. By direct calculation, it is easily seen that

$$U_{N+1} = \begin{bmatrix} * & * & \cdots & * & I \\ 0 & 0 & \cdots & I & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & I & \cdots & 0 & 0 \\ I & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

It follows that the pair $(\hat{F}, \hat{G})$ is *index-invariant.* (Recall that a pair (F, G) is called index-invariant by Morse and Silverman [1972] if and only if rank $U_k(j) = n_k$ for all j in $\mathbb{Z}$.) We can now apply the discrete-time versions of the results of Wolovich [1968], Brunovsky [1970], and Morse and Silverman [1972] on state-feedback control of index-invariant systems to the system (6.3). In particular, given any real numbers $a_0, a_1, \cdots, a_{n(N+1)-1}$, there exist a feedback matrix $\hat{L}$ over B and an $(N+1)n \times (N+1)n$ matrix T over B with T^{-1} also over B, such that

$$F^* := (\sigma^{-1}T^{-1})(\hat{F} + \hat{G}\hat{L})T$$

is time-invariant and the characteristic polynomial of F is

$$\det(\lambda I - F^*) = \lambda^{(N+1)n} + \sum_{i=0}^{(N+1)n-1} a_i \lambda^i. \tag{6.4}$$

Let us partition the matrix

$$\hat{L} = [L | L_c],$$

where L is an $n \times n$ matrix and L_c is an $n \times Nn$ matrix. Then, the corresponding state-feedback control law for (6.3) can be written as

$$v(k) = L(k)x(k) + L_c(k)z(k). \tag{6.5}$$

Now the overall system equations may be written as follows:

$$x(k+1) = F(k)x(k) + G(k)C(k)z(k),$$

$$z(k+1) = (A + BL_c)(k)z(k) + BL(k)x(k),$$

$$u(k) = [S_0 \quad S_1 \quad \cdots \quad S_{N-1}](k)z(k).$$

Thus, (6.5) represents a dynamic state-feedback scheme for Σ. Further, the "characteristic polynomial" (in the sense of (6.4)) of (6.5) has arbitrarily assignable coefficients $a_0, a_1, \cdots, a_{n(N+1)-1}$. We can summarize the above discussion in the following:

THEOREM 6.6. *Let $\Sigma = (F, G, H, J)$ be reachable in N steps over B at all times. Let $a_0, a_1, \cdots, a_{(N+1)n-1}$ be a given set of real numbers. Then, there exist a dynamic feedback compensator $\Sigma_c = (\hat{A}, \hat{B}, \hat{C})$ and an $(N+1)n \times (N+1)n$ matrix T over B, with inverse T^{-1} over B such that with*

$$\bar{F} := \begin{bmatrix} F & G\hat{C} \\ \hat{A} & \hat{B} \end{bmatrix},$$

$F^* := (\sigma^{-1}T)^{-1}\bar{F}T$ is time-invariant and

$$\det(\lambda I - F^*) = \lambda^{(N+1)n} + \sum_{i=0}^{(N+1)n-1} a_i \lambda^i.$$

By letting $B = l^\infty(\mathbb{Z})$, and using Lyapunov transformations, Theorem 6.5 gives us a stabilization result with arbitrarily assignable dynamics.

7. Discussion. In this paper we have developed a transfer-function approach to linear time-varying discrete-time systems. In this framework, we considered representation, stability and feedback control of linear time-varying systems. In the application to feedback control, we derived results on the assignability of closed-loop system dynamics by using a dynamic output feedback or state feedback compensator.

There are a number of topics that one can pursue using the transfer-function framework. One of these is an in-depth study of matrix-fraction descriptions of linear

time-varying systems. (There has already been a good deal of progress on this—see Khargonekar and Poolla [1984].) Another topic is the application of the transfer-function construct to the study of tracking and disturbance rejection (see Poolla [1984]).

REFERENCES

B. D. O. ANDERSON AND J. B. MOORE [1981], *Detectability and stabilizability of time-varying discrete-time linear systems*, this Journal, 19, pp. 20-32.

P. J. ANTSAKLIS [1979], *Some relations satisfied by prime polynomial matrices and their role in linear multivariable system theory*, IEEE Trans. Automat. Control, AC-25, pp. 611-616.

P. BRUNOVSKY [1970], *A classification of linear controllable systems*, Kybernetika, 3, pp. 173-188.

L. CHENG AND J. B. PEARSON [1978], *Frequency domain synthesis of multivariable linear regulators*, IEEE Trans. Automat. Control, AC-23, pp. 3-15.

C. A. DESOER, R. W. LIU, J. MURRAY AND R. SAEKS [1980], *Feedback system design: the fractional representation approach to analysis and synthesis*, IEEE Trans. Automat. Control, AC-25, pp. 399-412.

E. EMRE AND L. M. SILVERMAN [1981], *The equation $XR+QY=\Phi$: a characterization of solutions*, this Journal, 19, pp. 33-38.

D. S. EVANS [1972], *Finite-dimensional realizations of discrete-time weighting patterns*, SIAM J. Appl. Math., 22, pp. 45-67.

J. J. FERRER AND E. W. KAMEN [1984], *Realization of linear time-varying discrete-time systems*, in preparation.

P. A. FUHRMANN [1976], *Algebraic system theory: an analyst's point of view*, J. Franklin Institute, 301, pp. 521-540.

——— [1977], *On strict system equivalence and similarity*, Internat. J. Control, 25, pp. 5-10.

W. L. GREEN AND E. W. KAMEN [1984], *On stability of linear difference equations with time-varying coefficients*, in preparation.

E. W. KAMEN [1974], *A new algebraic approach to linear time-varying systems*, Technical report, Georgia Institute of Technology, Atlanta.

E. W. KAMEN AND K. M. HAFEZ [1979], *Algebraic theory of linear time-varying systems*, this Journal, 17, pp. 500-510.

P. P. KHARGONEKAR [1982], *On matrix fraction representations for linear systems over commutative rings*, this Journal, 20, pp. 172-197.

P. P. KHARGONEKAR AND A. B. OZGULER [1984], *Regulator problem with internal stability: a frequency domain solution*, IEEE Trans. Automat. Control, AC-29, pp. 332-343.

P. P. KHARGONEKAR AND K. POOLLA [1984], *On matrix-fraction representations for linear time-varying systems*, submitted for publication.

P. P. KHARGONEKAR AND E. D. SONTAG [1982], *On the relation between stable matrix fraction factorizations and regulable realizations of linear systems over rings*, IEEE Trans. Automat. Control, AC-27, pp. 627-638.

A. S. MORSE AND L. M. SILVERMAN [1972], *Structure of index-invariant systems*, this Journal, 11, pp. 215-225.

J. MURRAY [1982], *Time-varying systems and crossed products*, Math. Systems Theory, 17, pp. 217-241.

M. NEWMAN [1973], *Integral Matrices*, Academic Press, New York.

K. POOLLA [1984], *Linear time-varying systems: Representation and control via transfer-function matrices*, Ph.D. Dissertation, Univ. Florida, Gainesville.

H. H. ROSENBROCK [1970], *State Space and Multivariable Theory*, John Wiley, New York.

H. H. ROSENBROCK AND G. E. HAYTON [1978], *The general problem of pole-assignment*, Internat. J. Control, 27, pp. 837-852.

L. WEISS [1972], *Controllability, realization, and stability of discrete-time systems*, this Journal, 10, pp. 230-251.

W. A. WOLOVICH [1968], *On the stabilization of controllable systems*, IEEE Trans. Automat. Control, AC-13, pp. 569-572.

——— [1974], *Linear Multivariable Systems*, Springer-Verlag, New York.

L. A. ZADEH [1950], *Frequency analysis of variable networks*, Proc. IRE, 38, pp. 291-299.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

005

DIFFERENTIAL GAMES, OPTIMAL CONTROL AND DIRECTIONAL DERIVATIVES OF VISCOSITY SOLUTIONS OF BELLMAN'S AND ISAACS' EQUATIONS*

P.-L. LIONS† AND P. E. SOUGANIDIS‡

Abstract. Recent work by the authors and others has demonstrated the connections between the dynamic programming approach to optimal control theory and to two-person, zero-sum differential games problems and the new notion of "viscosity" solutions of Hamilton-Jacobi PDE's introduced by M. G. Crandall and P.-L. Lions. In particular, it has been proved that the dynamic programming principle implies that the value function is the viscosity solution of the associated Hamilton-Jacobi-Bellman and Isaacs equations. In the present work, it is shown that viscosity super- and subsolutions of these equations must satisfy some inequalities called super- and subdynamic programming principle respectively. This is then used to prove the equivalence between the notion of viscosity solutions and the conditions, introduced by A. Subbotin, concerning the sign of certain generalized directional derivatives.

AMS (MOS) subject classifications. 35F30, 35L60, 90D25, 49C20

Key words. differential games, optimal control, Hamilton-Jacobi equations, directional derivatives, viscosity solutions

Introduction. Recent work by the authors and others has demonstrated the connections between the dynamic programming approach to optimal control theory problems and to two-person, zero-sum differential games and the new notion of "viscosity" solutions of Hamilton-Jacobi partial differential equations introduced by M. G. Crandall and P.-L. Lions [6].

The formal relationships here are (cf. W. H. Fleming and R. Rishel [15], R. Isaacs [18]): if the values of various optimal control problems and differential games are regular, then they solve certain first order partial differential equations with "min", "max", "max-min" or "min-max" type nonlinearity. The problem is that usually the value functions are not smooth enough to make sense of the above in any obvious way. Many papers in the subject have worked around this difficulty: see Fleming [12], [13], [14], Friedman [15], [16], Elliott-Kalton [8], [9], Krassovski-Subbotin [20], Subbotin [28], etc.

Recently, however, the new notion of "viscosity" solution for first order partial differential equations was introduced by M. G. Crandall and P.-L. Lions [6]. (Also see M. G. Crandall, L. C. Evans and P.-L. Lions [5].) This solution was proved to be unique under some very general assumptions. Moreover, it was observed by P.-L. Lions [21] that the dynamic programming condition for the value in control theory problems implies that this value is the viscosity solution of the associated Hamilton-Jacobi-Bellman partial differential equation. These considerations extend to the theory of differential games. It follows, in particular, that the dynamic programming conditions imply that the values are viscosity solutions of the associated Hamilton-Jacobi-Isaacs partial differential equations. See P. Souganidis [26], [27] for a proof of this based on both the Fleming and the Friedman definitions of upper and lower values for a

* Received by the editors December 6, 1983, and in revised form July 2, 1984.

† Université Paris IX-Dauphine, Place de Lattre de Tassigny, 75775, Paris, Cedex 16, France. This work was done while the author was visiting the Mathematics Research Center, University of Wisconsin-Madison.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported in part by the National Science Foundation under grant MCS-8002946 and by the U.S. Army under grant DAAG 29-80-C-0041. This work was done while the author was a Predoctoral Fellow at the Mathematics Research Center, University of Wisconsin-Madison.

differential game, N. Barron, L. C. Evans and R. Jensen [1] for a different proof for the Friedman definition, L. C. Evans and P. Souganidis [11] for the Elliott-Kalton values in $\mathbb{R}^n$, L. C. Evans and H. Ishii [10] for the Elliott-Kalton values in bounded domains. Some related papers are: P.-L. Lions [23], P.-L. Lions and M. Nisio [25], I. Capuzzo Dolcetta and L. C. Evans [3], I. Capuzzo Dolcetta [2], I. Capuzzo Dolcetta and H. Ishii [4], H. Ishii [19], etc.

The present paper is concerned with the relation between the notion of viscosity sub- and super-solutions of first-order, dynamic programming PDE and the optimality principle of dynamic programming, as well as the directional derivatives of viscosity solution of the above equations at an arbitrary point. In particular, we show that continuous sub- and super-solutions of the Hamilton-Jacobi-Bellman and Isaacs equations satisfy certain inequalities related to the optimality principle of dynamic programming. Under some assumptions this implies a particular sign for certain generalized directional derivatives. Finally, this sign suffices to characterize functions as viscosity sub- and super-solutions of the appropriate equations. For purely technical reasons this is interesting, since it shows how the definition of the viscosity solution, which restricts the behavior of the solution only at points of upper- or lower-differentiability, forces conditions on the (Dini) directional derivatives at an arbitrary point. Moreover, it provides one with an easy way to check whether a smooth function satisfies the equations at every point. From the point of view of the applications these questions relate to dynamic programming for control or game problems without smoothness of the value function and presumably have some bearing on the synthesis of (generalized) optimal controls, but we have not attempted to work out the details. Finally, they may be useful for the investigation of the structure of singularities of solutions of Hamilton-Jacobi type equations (e.g., the eiconal equation in optics).

The work is motivated by a paper of A. Subbotin [28]. In [28], Subbotin gives a necessary and sufficient condition for a function to be the value of a differential game. This condition, which is not within the context of the viscosity solution, roughly says that at every point certain generalized derivatives must have a particular sign. L. C. Evans and H. Ishii [10], using a "blow-up" argument, showed that the value of an infinite horizon control problem satisfies Subbotin's condition, as it applies to control problems. The techniques used here are different than the ones in [10]. One direction of the equivalence claimed above is straightforward. The other is closely related to the principle of dynamic programming and requires some arguments of P.-L. Lions [22], [24], which treat optimal control problems of diffusion processes.

The paper is organized as follows: The rest of the introduction recalls the definition of the viscosity solution. Section 1 is devoted to optimal control problems. Section 2 deals with differential games. In the Appendix we make some observations concerning the existence of directional derivatives of the value function. All the definitions and results from other papers are recalled when necessary.

We conclude the introduction with the definition of viscosity solutions.

DEFINITION 0.1 [5], [6]. Let $H:\Omega\times\mathbb{R}\times\mathbb{R}^N\to\mathbb{R}$ and $z:\partial\Omega\to\mathbb{R}$ be continuous functions, where Ω is an open subset of $\mathbb{R}^N$. A continuous function $u:\bar{\Omega}\to\mathbb{R}$ is a viscosity subsolution of

$$\begin{aligned} H(x, u, Du) &= 0 \quad \text{in } \Omega, \\ u(x) &= z(x) \qquad \text{on } \partial\Omega \end{aligned} \tag{0.1}$$

if $u(x)\leqq z(x)$ on $\partial\Omega$ and, moreover, for every $\phi\in C^\infty(\Omega)$,[1] if $x_0\in\Omega$ is a local max of

[1] $C^\infty_{(0)}(\mathcal{O})$ denotes the set of real valued infinitely many times continuously differentiable functions (of compact support) defined on $\mathcal{O}$.

$u-\varphi$, then

$$H(x_0, u(x_0), D\varphi(x_0)) \leqq 0. \tag{0.2}$$

A continuous function $u:\bar{\Omega}\to\mathbb{R}$ is a viscosity supersolution of (1.1), if $u(x)\geqq z(x)$ on $\partial\Omega$ and, moreover, for every $\phi\in C^\infty(\Omega)$, if $x_0\in\Omega$ is a local min of $u-\varphi$, then

$$H(x_0, u(x_0), D\phi(x_0)) \geqq 0. \tag{0.3}$$

A continuous function $u:\bar{\Omega}\to\mathbb{R}$ is a viscosity solution of (1.1), if it is both sub- and supersolution of (1.1).

DEFINITION 0.2 [5], [6]. Let $H:\Omega\times[0, T]\times\mathbb{R}\times\mathbb{R}^N\to\mathbb{R}$, $z:\partial\Omega\times[0, T]\to\mathbb{R}$ and $u_0:\bar{\Omega}\to\mathbb{R}$ be continuous functions, where Ω is an open subset of $\mathbb{R}^N$. A viscosity subsolution (respectively supersolution) of

$$\begin{aligned} &\frac{\partial u}{\partial t}+H(t, x, u, Du)=0 && \text{in } \Omega\times(0, T],\\ &u(x, t)=z(x, t) && \text{on } \partial\Omega\times(0, T],\\ &u(x, 0)=u_0(x) && \text{on } \Omega \end{aligned} \tag{0.4}$$

is a function $u\in C(\bar{Q}_T)$[2] such that:

$$u\leqq z \quad \text{on } \partial\Omega\times[0, T], \qquad u(x, 0)\leqq u_0(x) \quad \text{in } \Omega \tag{0.5}$$

(respectively

$$u\geqq z \quad \text{on } \partial\Omega\times[0, T], \qquad u(x, 0)\geqq u_0(x) \quad \text{in } \Omega\,), \tag{0.6}$$

and, for every $\phi\in C^\infty(Q_T)$,

(0.7) if $(x_0, t_0)\in Q_T$ is a local max of $u-\varphi$, then

$$\frac{\partial\phi}{\partial t}(x_0, t_0)+H(t_0, x_0, u(x_0, t_0), D\phi(x_0, t_0))\leqq 0$$

(respectively

if $(x_0, t_0)\in Q_T$ is a local min of $u-\varphi$, then

(0.8)
$$\frac{\partial\varphi}{\partial t}(x_0, t_0)+H(t_0, x_0, u(x_0, t_0), D\phi(x_0, t_0))\geqq 0\,).$$

A function $u\in C(\bar{Q}_T)$ is a viscosity solution of (0.4), if it is both sub- and supersolution of (0.4).

For a detailed account of the recent developments in the theory of viscosity solutions as well as references, we refer to the book by P.-L. Lions [21] and the article by M. G. Crandall and P. Souganidis [7].

1. In this section we consider Hamilton-Jacobi-Bellman equations associated with optimal control problems. In particular, we look at problems of the form

$$\begin{aligned} &u+\sup_{y\in Y}\{-f(x, y)\cdot Du-l(x, y)\}=0 && \text{in } \Omega,\\ &u=g && \text{on } \partial\Omega \end{aligned} \tag{1.1}$$

where Ω is an open subset of $\mathbb{R}^N$, $g\in C(\partial\Omega)$, Y is some separable metric space and

[2] $C(\mathcal{O})$ is the set of continuous real valued functions defined on $\mathcal{O}$. $Q_T=\Omega\times(0, T]$, $\bar{Q}_T=\bar{\Omega}\times[0, T]$.

$f:\bar{\Omega}\times Y\to\mathbb{R}^N$, $l(x,y):\bar{\Omega}\times Y\to\mathbb{R}$ are continuous functions such that

there exists a constant $C>0$ such that

$$|f(x,y)|,\ |l(x,y)|\leqq C \quad \text{for every } (x,y)\in\bar{\Omega}\times Y$$

(1.2) and

$$|f(x,y)-f(\hat{x},y)|,\ |l(x,y)-l(\hat{x},y)|\leqq C|x-\hat{x}| \quad \text{for every } (x,\hat{x},y)\in\bar{\Omega}\times\bar{\Omega}\times Y.$$

Problem (1.1) corresponds to an infinite horizon control problem (for the details we refer to [21] and W. Fleming and R. Rishel [15]) with dynamics given by

$$\frac{dx}{d\tau}=f(x(\tau),y(\tau)) \quad \text{for } 0<\tau, \qquad x(0)=x\in\Omega \tag{1.3}$$

where $y:[0,\infty)\to Y$ is measurable. For notational simplicity in what follows let

$$M=\{y:[0,\infty)\to Y,\ y(\cdot)\ \text{measurable}\}. \tag{1.4}$$

Let u be the unique viscosity solution of (1.1) if it exists. It is known (P.-L. Lions [21], I. Capuzzo Dolcetta and L. C. Evans [3], L. C. Evans and H. Ishii [10]) that u satisfies the *optimality principle* of dynamic programming, that is

$$u(x)=\inf_M\left\{e^{-(t\wedge t_x)}u(x(t\wedge t_x))+\int_0^{t\wedge t_x}e^{-s}l(x(s),y(s))\,ds\right\} \tag{1.5}$$

for every $t>0$ and $x\in\Omega$[3]

where, for $x\in\Omega$ and $y\in M$, $t_x=t_x(y)$ is the *exit time from* Ω of the solution of (1.3) for the particular x, y, i.e.,

$$t_x=\inf\{t>0:\ x(t)\in\mathbb{R}^N-\Omega\}.$$

The first result of this section concerns viscosity supersolutions of (1.1). In particular, we show that every viscosity supersolution of (1.1) satisfies some inequality, which, in view of (1.5), may be called *superoptimality principle* of dynamic programming. This was first proved by P.-L. Lions in [22], [24], in the general context of optimal stochastic control. Here we give two proofs related to those given in [22], [24], but slightly more adapted to the special situation at hand. The first proof uses the fact that a viscosity supersolution of (1.1) is a viscosity supersolution of an appropriately defined time-dependent problem. The second proof is based on the fact that a viscosity supersolution of (1.1) is a viscosity solution of an obstacle problem, which can be solved easily using differential games. The first step in both proofs is a localization argument introduced in [24]. It consists of multiplying by suitable cut-off function. This allows us to reduce to the case $\Omega=\mathbb{R}^N$.

We have

PROPOSITION 1.1. *Let $v\in C(\bar{\Omega})$ be a viscosity supersolution of* (1.1). *Then, for every $t>0$ and $x\in\Omega$, we have*

$$v(x)\geqq\inf_M\left\{e^{-(t\wedge t_x)}v(x(t\wedge t_x))+\int_0^{t\wedge t_x}e^{-s}l(x(x),y(s))\,ds\right\}. \tag{1.6}$$

[3] $r\wedge s=\min\{r,s\}$.

Proof 1. The first step in the proof is to modify the problem so that it is defined in $\mathbb{R}^N$. To this end, for $\delta > 0$ let Ω_δ be defined by

$$\Omega_\delta = \{x \in \Omega : |x| < 1/\delta \text{ and } \operatorname{dist}(x, \partial\Omega) > \delta\}.$$

Moreover, choose $\xi, \varphi \in C^\infty(\mathbb{R}^N)$ such that $\xi \equiv 1$ on $\bar{\Omega}_\delta$, $0 \leqq \xi \leqq 1$, $\xi \equiv 0$ on $\mathbb{R}^N \setminus \bar{\Omega}_{\delta/2}$, $\varphi \equiv 1$ on a neighborhood of supp ξ, $0 \leqq \varphi \leqq 1$ and $\varphi \equiv 0$ on $\mathbb{R}^N \setminus \bar{\Omega}_{\delta/4}$. Then the function $\tilde{v} : \mathbb{R}^N \to \mathbb{R}$ which is defined by

$$\tilde{v}(x) = \begin{cases} (\varphi v)(x) & \text{in } \Omega, \\ 0 & \text{in } \mathbb{R}^N \setminus \Omega \end{cases}$$

is a viscosity supersolution of the problem

$$\xi(x)\tilde{u} + \sup_{y \in Y} \{-\xi(x) f(x, y) \cdot D\tilde{u} - \xi(x) l(x, y)\} = 0 \quad \text{in } \mathbb{R}^N. \tag{1.7}$$

Next for $t > 0$ fixed consider the initial value problem

$$\begin{aligned} &\partial w/\partial s + \sup_{y \in Y} \{-\xi(x) f(x, y) \cdot Dw - \xi(x) l(x, y)\} + \xi(x) w = 0 \quad \text{in } Q_t, \\ &w(x, 0) = \tilde{v}(x) \quad \text{in } \mathbb{R}^N. \end{aligned} \tag{1.8}$$

In view of the results of [6] and [21], (1.8) has a unique viscosity solution $w \in C(\mathbb{R}^N \times [0, t])$ given by

$$\begin{aligned} w(x, s) = \inf_M \Bigg\{ & v(\tilde{x}(s)) \exp\left(-\int_0^s \xi(\tilde{x}(\tau))\, d\tau\right) \\ & + \int_0^s \exp\left(\int_0^\tau \xi(\tilde{x}(\lambda))\, d\lambda\right) l(\tilde{x}(\tau), y(\tau))\, d\tau \Bigg\} \end{aligned} \tag{1.9}$$

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

$$\frac{d\tilde{x}}{d\tau} = -\xi(\tilde{x}(\tau)) f(\tilde{x}(\tau), y(\tau)) \quad \text{for } 0 < \tau < t,$$

$$\tilde{x}(0) = x.$$

$\tilde{v}$, however, is a viscosity supersolution of (1.8). The uniqueness estimates of [6] imply

$$\tilde{v}(x) \geqq w(x, s) \quad \text{for every } (x, s) \in \mathbb{R}^N \times [0, t].$$

Next observe that for $x \in \Omega$ and $y \in M$, if $t < t_x$, then $\tilde{x}(s) = x(s)$ for $0 \leqq s \leqq t$, where $x(\cdot)$ is the solution of (1.3), provided that δ is sufficiently small. Moreover, $x(\cdot) \in \{x \in \Omega : \xi(x) = 1\}$. These observations together with (1.9) imply (1.6) for $t < t_x$. If $t \geqq t_x$, choose $t_n \uparrow t_x$. Then

$$v(x) \geqq \inf_{y \in M} \left\{ v(x(t_n))\, e^{-t_n} + \int_0^{t_n} e^{-s} l(x(s), y(s))\, ds \right\}.$$

As $n \to \infty$ we obtain (1.6), since $v \in C(\bar{\Omega})$.

Proof 2 (*obstacle problem method*). Here, in order to exhibit the main ideas, for simplicity, we are going to assume $\Omega = \mathbb{R}^N$. The general case follows by appropriate use of the localization technique explained in Proof 1.

It is easy to see that v is the unique viscosity solution of the problem

$$v + \min\{\sup_{y \in Y} \{-f(x, y) \cdot Dv - l(x, y)\}, -v\} = 0 \quad \text{in } \mathbb{R}^N$$

which can be rewritten as

$$v+\min_{z\in Z}\sup_{y\in Y}\{-\tilde{f}(x,y,z)\cdot Dv-\tilde{h}(x,y,z)\}=0 \quad \text{in } \mathbb{R}^N \tag{1.10}$$

with $Z=\{1,2\}$ and

$$\tilde{f}(x,y,z)=\begin{cases}0, & \text{if } z=1,\\ f(x,y), & \text{if } z=2,\end{cases}$$

$$\tilde{h}(x,y,z)=\begin{cases}v(x), & \text{if } z=1,\\ -l(x,y), & \text{if } z=2.\end{cases}$$

Equation (1.10) corresponds to an infinite horizon differential game, thus v must satisfy the dynamic programming principle, as it is shown in the first part of L. C. Evans and H. Ishii [10].

We need some more notation. In particular, let

$$N=\{z:[0,\infty)\to Z,\ z(\cdot)\ \text{measurable}\}.$$

Moreover, denote by Γ the set of mappings $\alpha: N\to M$, which, for every $t>0$, satisfy the following condition:

If $z(x)=\hat{z}(s)$ for a.e. $0\leqq s\leqq t$, then $\alpha[z](s)=\alpha[\hat{z}](s)$ for a.e. $0\leqq s\leqq t$.

In view of [10, Thm. 3.1], for every $t>0$, we obtain

$$v(x)=\inf_{\alpha\in\Gamma}\sup_{z\in N}\left\{\int_0^t e^{-s}\tilde{h}(\tilde{x}(s),\alpha[z](s),z(s))\,ds+e^{-t}v(\tilde{x}(t))\right\} \tag{1.11}$$

where for $x\in\mathbb{R}^N$, $z\in N$ and $\alpha\in\Gamma$, $\tilde{x}(\cdot)$ is the unique solution of

$$\frac{d\tilde{x}}{ds}=-f(\tilde{x}(s),\alpha[z](s),z(s)) \quad \text{for } 0<s,$$

$$\tilde{x}(0)=x.$$

Choose $\bar{z}\in N$ such that $\bar{z}\equiv 2$. Then (1.11) implies

$$v(x)\geqq\inf_{\alpha\in\Gamma}\left\{v(x(t))\,e^{-t}+\int_0^t e^{-\tau}l(x(\tau),a[\bar{z}](\tau))\,d\tau\right\}$$

since, in this case, $\tilde{x}(\cdot)$ is the solution of (1.3). But

$$\{\alpha[\bar{z}]:\ \alpha\in\Gamma\}\subset M;$$

thus the result.

The next proposition deals with viscosity subsolutions of (1.1). In particular, we show that a viscosity subsolution of (1.1) satisfies an inequality, which we call the *suboptimality principle* of dynamic programming. The proof relies on the fact that viscosity subsolutions of (1.1) are viscosity subsolutions of an appropriately defined time-dependent problem.

We have

PROPOSITION 1.2. *Let $w\in C(\bar{\Omega})$ be a viscosity subsolution of* (1.1). *Then, for every $x\in\mathbb{R}^N$ and $t\geqq 0$,*

$$w(x)\leqq\inf_{M}\left\{e^{-(t\wedge t_x)}w(x(t\wedge t_x))+\int_0^{(t\wedge t_x)}e^{-s}l(x(s),y(s))\,ds\right\}. \tag{1.12}$$

Proof. Here we give the proof in the case $\Omega=\mathbb{R}^N$. For the general case, one has to use first a localization argument as in Proof 1 of Proposition 1.1.

For $t>0$ consider the problem

$$\frac{\partial z}{\partial s}+\max_{y\in Y}\{-f(x,y)\cdot Dz+l(x,y)\}+z=0 \quad \text{in } Q_t,$$

$$z(x,0)=w(x) \quad \text{in } \mathbb{R}^N.$$

w is a viscosity subsolution of this problem; therefore, for every $x\in\mathbb{R}^N$,

$$w(x)\leqq z(x,t)=\inf_M\left\{w(x(t))\,e^{-t}+\int_0^t e^{-s}l(x(s),y(s))\,ds\right\}.$$

The above are justified as in Proof 1 of Proposition 1.1.

Next we want to use Propositions 1.1 and 1.2 to obtain a kind of infinitesimal version of the super- and suboptimality principle of the dynamic programming, satisfied by viscosity super- and subsolutions of (1.1). To do this, we have to assume that we work with sub- and supersolutions which are locally Lipschitz. Moreover, we need to introduce the following notation.

$$(\mathrm{FL})(x)=\overline{\mathrm{co}}\,\{(f(x,y),l(x,y)):y\in Y\}. \tag{1.13}$$

We have:

THEOREM 1.3. *Let $v\in C^{0,1}_{\mathrm{loc}}(\Omega)$[4] be a viscosity supersolution of* (1.1). *Then, for every $x\in\Omega$,*

$$v(x)+\lim_{\delta\downarrow 0}\sup_{(f,l)\in(\mathrm{FL})(x)}\left\{\frac{v(x)-v(x+\delta f)}{\delta}-l\right\}\geqq 0 \tag{1.14}$$

and the inequality is achieved as $\delta\downarrow 0$ uniformly on compact sets.

Proof. For $x\in\Omega$ fixed let K be the Lipschitz constant of v in a ball of radius C centered at x, where C is given by (1.2). Moreover, assume that $0<\delta<1$ is small enough so that $\delta<t_x$. In view of Proposition 1.1, we have:

$$\sup_M\left\{\frac{v(x)-e^{-\delta}v(x+\int_0^\delta f(x(s),y(s))\,ds)}{\delta}-\frac{1}{\delta}\int_0^\delta e^{-s}l(x(s),y(s))\,ds\right\}\geqq 0.$$

Therefore

$$\frac{1-e^{-\delta}}{\delta}v(x)+e^{-\delta}\sup_M\left\{\frac{v(x)-v(x+\delta(1/\delta)\int_0^\delta f(x,y(s))\,ds)}{\delta}-\frac{1}{\delta}\int_0^\delta l(x,y(s))\,ds\right\}$$

$$\geqq-(K+1)C^2\delta-C\left(1+\frac{e^{-\delta}-1}{\delta}\right).$$

But

$$\left(\frac{1}{\delta}\int_0^\delta f(x,y(x))\,ds,\frac{1}{\delta}\int_0^\delta l(x,y(s))\,ds\right)\in(\mathrm{FL})(x).$$

The above inequality implies

$$\frac{1-e^{-\delta}}{\delta}v(x)+e^{-\delta}\sup_{(f,l)\in(\mathrm{FL})(x)}\left\{\frac{v(x)-v(x+\delta f)}{\delta}-l\right\}\geqq-(K+1)C^2\delta-C\left(1+\frac{e^{-\delta}-1}{\delta}\right).$$

Letting $\delta\downarrow 0$ we obtain the result. The uniformity claimed in the statement is an

[4] $C^{0,1}_{\mathrm{loc}}(\mathcal{O})$ denotes the set of real valued (locally) Lipschitz continuous functions defined on $\mathcal{O}$.

immediate consequence of the fact that the above also holds for every $y \in \mathbb{R}^N$ in a neighborhood of x of radius $C/2$.

As a consequence of Theorem 1.3 we have

COROLLARY 1.4. *Let $v \in C^{0,1}_{\text{loc}}(\Omega)$ be a viscosity supersolution of* (1.1). *Then, for every $x \in \Omega$,*

$$v(x) + \sup_{(f,l)\in(\mathrm{FL})(x)} \left\{ \overline{\lim_{\delta \downarrow 0}} \frac{v(x) - v(x+\delta f)}{\delta} - l \right\} \geqq 0 \tag{1.15}$$

and the inequality is achieved as $\delta \downarrow 0$ uniformly on compact sets.

Remark 1.5. In the second part of [10] L. C. Evans and H. Ishii proved that if $\{f(x, y): y \in Y\}$ is convex and $l(x, y) \equiv 0$, then a locally Lipschitz viscosity supersolution of (1.1) satisfies

$$\inf_{y\in Y} \lim_{\delta \downarrow 0} \frac{e^{-\delta} v(x+\delta f) - v(x)}{\delta} \leqq 0$$

which, under their assumptions, is equivalent to (1.15). As mentioned in the Introduction, they used a "blow-up" argument. The proof we give here is based completely on Proposition 1.1 and Theorem 1.3.

Proof of Corollary 1.4. Relation (1.14) implies that there is a subsequence $\delta_k \downarrow 0$ as $k \to \infty$ such that

$$v(x) + \lim_{k\to\infty} \sup_{(f,l)\in(\mathrm{FL})(x)} \left\{ \frac{v(x) - v(x+\delta_k f)}{\delta_k} - l \right\} \geqq 0.$$

Then for $\varepsilon > 0$ fixed but arbitrary there is a $k_0 = k_0(\varepsilon) > 0$ such that for $k \geqq k_0$

$$v(x) + \sup_{(f,l)\in(\mathrm{FL})(x)} \left\{ \frac{v(x) - v(x+\delta_k f)}{\delta_k} - l \right\} \geqq -\varepsilon.$$

Next for each $k \geqq k_0$ choose $(f_k, l_k) \in (\mathrm{FL})(x)$ such that

$$\frac{v(x) - v(x+\delta_k f_k)}{\delta_k} - l_k = \sup_{(f,l)\in(\mathrm{FL})(x)} \left\{ \frac{v(x) - v(x+\delta_k f)}{\delta_k} - l \right\}.$$

The compactness of $(\mathrm{FL})(x)$ implies that along some subsequence of $\delta_k \downarrow 0$ (which again for simplicity is denoted by δ_k) we have

$$(f_k, l_k) \to (f, l) \in (\mathrm{FL})(x).$$

This, together with the Lipschitz property of v, implies

$$v(x) + \overline{\lim_{k\to\infty}} \left\{ \frac{v(x) - v(x+\delta_k f)}{\delta_k} - l \right\} \geqq 0$$

and, therefore, the result.

Remark 1.6. It is of some interest to know whether, in the case at hand, (1.15) holds with $\overline{\lim}$ replaced by $\underline{\lim}$.

For the case of a viscosity subsolution of (1.1) we have:

THEOREM 1.7. *Let $w \in C^{0,1}_{\text{loc}}(\Omega)$ be a viscosity subsolution of* (1.1). *Then, for every $x \in \Omega$,*

$$w(x) + \overline{\lim_{\delta \downarrow 0}} \sup_{(f,l)\in(\mathrm{FL})(x)} \left\{ \frac{w(x) - w(x+\delta f)}{\delta} - l \right\} \leqq 0 \tag{1.16}$$

and

$$w(x)+\sup_{(f,l)\in(\mathrm{FL})(x)}\overline{\lim_{\delta\downarrow 0}}\left\{\frac{w(x)-w(x+\delta f)}{\delta}-l\right\}\leqq 0 \tag{1.17}$$

and the inequality is achieved as $\delta\downarrow 0$ *uniformly on compact sets.*

Proof. Relation (1.17) follows immediately from (1.16). To prove (1.16) observe that, in view of Proposition 1.2, we have

$$\frac{w(x)-e^{-\delta}w(x+\int_0^\delta f(x(s),y(s))\,ds}{\delta}-\int_0^\delta e^{-s}l(x(s),y(s))\,ds\leqq 0$$

and, therefore

$$\begin{aligned}&\frac{1-e^{-\delta}}{\delta}w(x)+e^{-\delta}\frac{w(x)-w(x+\delta(1/\delta)\int_0^\delta f(x,y(s))\,ds)}{\delta}-\frac{1}{\delta}\int_0^\delta l(x,y(s))\,ds\\ &\leqq(K+1)C^2\delta+C\left(1+\frac{e^{-\delta}-1}{\delta}\right)\end{aligned} \tag{1.18}$$

for every $y\in M$, where K is the Lipschitz constant of w in the ball of radius C centered at x.

In view of the general geometrical fact

$$\left\{\left(\frac{1}{\delta}\int_0^\delta f(x,y(s))\,ds,\frac{1}{\delta}\int_0^\delta l(x,y(s))\,ds\right)\colon y\in M\right\}=\overline{\mathrm{co}}\,\{(f(x,y),l(x,y))\colon y\in Y\},$$

relation (1.18) implies

$$\frac{1-e^{-\delta}}{\delta}w(x)+e^{-\delta}\sup_{(f,l)\in(\mathrm{FL})(X)}\left\{\frac{w(x)-w(x+\delta f)}{\delta}-l\right\}\leqq O(1)$$

where $O(1)\to 0$ as $\delta\downarrow 0$ and thus (1.16). The uniformity follows from the fact that all the above hold with the same constants for all points in an appropriate neighborhood of x.

Combining Corollary 1.4 and Theorem 1.6 we obtain

COROLLARY 1.8. *Let* $u\in C^{0,1}_{\mathrm{loc}}(\Omega)$ *be a viscosity solution of* (1.1). *Then*

$$u(x)+\sup_{(f,l)\in(\mathrm{FL})(x)}\left\{\overline{\lim_{\delta\downarrow 0}}\left\{\frac{u(x)-u(x+\delta f)}{\delta}\right\}-l\right\}=0\quad\forall x\in\Omega. \tag{1.19}$$

We continue with a result which is the inverse of Corollary 1.5 and Theorem 1.7. In particular, it says that (1.15) and (1.17) together with appropriate boundary conditions characterize continuous functions as viscosity super-(respectively sub-)solutions of (1.1). We have

PROPOSITION 1.9. (a) *Let* $v\in C(\bar\Omega)$ *satisfy* (1.15) *for every* $x\in\Omega$. *Then* v *satisfies* (0.3) *with* H *as in* (1.1).

(b) *Let* $w\in C(\bar\Omega)$ *satisfy* (1.17) *for every* $x\in\Omega$. *Then* w *satisfies* (0.2) *with* H *as in* (1.1).

Proof. (a) For $\phi\in C^\infty(\Omega)$ let $x_0\in\Omega$ be a local minimum of $v-\phi$. We want to show that

$$v(x_0)+\sup_{y\in Y}\{-f(x_0,y)\cdot D\phi(x_0,y)+l(x_0,y)\}\geqq 0.$$

But for δ sufficiently small we have

$$\frac{\phi(x_0)-\phi(x_0+\delta f)}{\delta}-l\geqq\frac{v(x_0)-v(x_0+\delta f)}{\delta}-l\quad\text{for all }(f,l)\in(\mathrm{FL})(x_0).$$

This inequality and (1.15) imply

$$v(x_0)+\sup_{(f,l)\in(\mathrm{FL})(x_0)}\{-f\cdot D\phi(x_0)-l\}\geqq 0.$$

Finally, since

$$\sup_{\lambda\in\Lambda}\lambda=\sup_{\lambda\in\overline{\mathrm{co}}\,\Lambda}\lambda,$$

we have the result.

(b) The proof is similar to the one of part (a), therefore we omit it.

Remark 1.10. All the results of this section extend to several other cases including time-dependent problems. The type of statements that one obtains are similar to the ones of § 2.

We conclude this section with an observation concerning smooth functions which satisfy (1.15) and (1.17). Since the proof is similar to the proof of Proposition 1.9, we omit it. We have:

PROPOSITION 1.11. *Let* $v\in C(\bar{\Omega})$ *satisfy* (1.15) *and* (1.17) *for every* $x\in\Omega$. *If* v *is differentiable at* $x_0\in\Omega$, *then*

$$v(x_0)+\sup_{y\in Y}\{-f(x_0,y)\cdot Dv(x_0)-l(x_0,y)\}=0. \tag{1.20}$$

Remark 1.12. Using the properties of viscosity solutions one can obtain Proposition 1.11 directly from Proposition 1.12. The point here is that (1.20) follows immediately from the conditions on the directional derivatives of v.

2. In this section we consider Hamilton–Jacobi equations which are related to the theory of two-player, zero-sum differential games. Since in § 1 we looked at stationary problems, here to show the generality of the arguments involved, we work with time-dependent ones. In particular, we consider the following problems:

$$\begin{aligned}
&\frac{\partial U}{\partial t}+\inf_{y\in Y}\sup_{z\in Z}\{-f(t,x,y,z)\cdot DU-l(t,x,y,z)\}=0 \quad \text{in } \Omega\times(0,T],\\
&U(x,t)=g(x,t) \quad \text{on } \partial\Omega\times[0,T], \qquad (2.1)\\
&U(x,0)=u_0(x) \quad \text{in } \Omega,
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial V}{\partial t}+\sup_{z\in Z}\inf_{y\in Y}\{-f(t,x,y,z)\cdot DV-l(t,x,y,z)\}=0 \quad \text{in } \Omega\times(0,T],\\
&V(x,t)=g(x,t) \quad \text{on } \partial\Omega\times[0,T], \qquad (2.2)\\
&V(x,0)=u_0(x) \quad \text{in } \Omega,
\end{aligned}$$

where Y, Z are compact sets and $f:[0,T]\times\bar{\Omega}\times Y\times Z\to\mathbb{R}^N$, $l:[0,T]\times\bar{\Omega}\times Y\times Z\to\mathbb{R}$, $g:\partial\Omega\times[0,T]\to\mathbb{R}$, $u_0:\bar{\Omega}\to\mathbb{R}$ are bounded continuous functions. Moreover, they satisfy

There exists a constant $C>0$ such that

$$|f(t,x,y,z)|,\ |l(t,x,y,z)|\leqq C \quad \text{for every } (t,x,y,z)\in[0,T]\times\bar{\Omega}\times Y\times Z \tag{2.3}$$

and

$$|f(t,x,y,z)-f(\hat{t},\hat{x},y,z)|,\ |l(t,x,y,z)-l(\hat{t},\hat{x},y,z)|\leqq C(|t-\hat{t}|+|x-\hat{x}|)$$
$$\text{for every } (t,x,y,z),(\hat{t},\hat{x},y,z)\in[0,T]\times\bar{\Omega}\times Y\times Z.$$

Problems (2.1) and (2.2) correspond to a finite horizon two-player, zero-sum differential game (for details we refer to W. Fleming [12], [13], [14], Elliott and Kalton [8], A. Friedman [16], [17]) with dynamics given by

$$(2.4)_{T-t} \qquad \begin{aligned} &\frac{dx}{d\tau} = f(\tau, x(\tau), y(\tau), z(\tau)) \quad \text{for } T-t<\tau<T, \\ &x(T-t) = x \in \Omega \end{aligned}$$

where $y:[t, T]\to Y$, $z:[t, T]\to Z$ are measurable functions. Before we continue we need to introduce some notation. In particular, for $0 \leqq t \leqq T$ define

$$M(t) = \{y:[t, T]\to Y \text{ measurable}\},$$

$$N(t) = \{z:[t, T]\to Z \text{ measurable}\}.$$

Moreover, denote by $\Gamma(t)$, $\Delta(t)$ the sets of mappings $\alpha: N(t)\to M(t)$, $\beta: M(t)\to N(t)$ respectively with the following property

For each s such that $t \leqq s \leqq T$

if $z(\tau) = \hat{z}(\tau)$ for a.e. $t \leqq \tau \leqq s$, then $\alpha[z](\tau) = \alpha[\hat{z}](\tau)$ for a.e. $t \leqq \tau \leqq s$

and

if $\hat{y}(\tau) = \hat{y}(\tau)$ for a.e. $t \leqq \tau \leqq s$, then $\beta[y](\tau) = \beta[y](\tau)$ for a.e. $t \leqq \tau \leqq s$.

Let U, V be the unique viscosity solutions of (2.1), (2.2) respectively if they exist. It is known (L. C. Evans and P. Souganidis [11] for $\Omega = \mathbb{R}^N$, L. C. Evans and H. Ishii [10] for stationary problems) and it follows from the results of this section for other cases that U, V satisfy the *optimality principle* of dynamic programming, that is

For $(x, t) \in \Omega \times (0, T)$ and $\delta > 0$ such that $\delta \leqq t$

$$U(x, t) = \inf_{\beta \in \Delta(T-t)} \sup_{y \in M(T-t)} \left\{ \int_{T-t}^{(T-t+\delta)\wedge t_x} l(s, x(s), y(s), \beta[y](s))\, ds \right.$$

$$\left. + U(x((T-t+\delta)\wedge t_x), T-((T-t+\delta)\wedge t_x)) \right\}$$

(2.5) and

$$V(x, t) = \sup_{\alpha \in \Gamma(T-t)} \inf_{z \in N(T-t)} \left\{ \int_{T-t}^{(T-t+\delta)\wedge t_x} l(s, x(s), \alpha[z](s), z(s))\, ds \right.$$

$$\left. + V(x((T-t+\delta)\wedge t_x), T-((T-t+\delta)\wedge t_x)) \right\}$$

where, for $x \in \Omega$, $x(\cdot)$ is the solution of $(2.3)_{T-t}$ with the appropriate $y(\cdot)$, $z(\cdot)$ functions and t_x is the exit time from $\Omega \times (0, T)$ of $x(\cdot)$.

The first result of this section concerns viscosity supersolutions and subsolutions of (2.1) and (2.2). In particular, we show that they satisfy some inequalities, which, in view of (2.5), may be called the *super-* and *suboptimality principle* of dynamic programming. All the results are going to be stated as they apply to the general problems (2.1) and (2.2); the proofs, however, for simplicity are going to be given only for the special case $\Omega = \mathbb{R}^N$. To obtain the most general results, one has to use the localization argument, which was described in the course of the proof of Proposition 1.1.

We have

PROPOSITION 2.1. *Let $v, w \in C(\bar{Q}_T)$ be viscosity super-(respectively sub-)solutions of* (2.1) (*respectively* (2.2)). *For every $(x, t) \in \Omega \times (0, T)$ and $\delta > 0$ such that $\delta \leqq t$, we have*

$$v(x,t) \geqq \inf_{\beta \in \Delta(T-t)} \sup_{y \in M(T-t)} \left\{ \int_{T-t}^{(T-t+\delta)\wedge t_x} l(s, x(s), y(s), \beta[y](s))\, ds + v(x((T-t+\delta)\wedge t_x), T-((T+\delta)\wedge t_x)) \right\} \tag{2.6}$$

and

$$w(x,t) \leqq \sup_{\alpha \in \Gamma(T-t)} \inf_{z \in N(T-t)} \left\{ \int_{T-t}^{(T-t+\delta)\wedge t_x} l(s, x(s), \alpha[z](s), z(s))\, ds + w(x((T-t+\delta)\wedge t_x), T-((T-t+\delta)\wedge t_x)) \right\} \tag{2.7}$$

Proof. Here we prove only (2.6), since (2.7) is proved in exactly the same way. As mentioned above we are going to assume $\Omega = \mathbb{R}^N$.

For $\varepsilon > 0$ choose $\xi, \phi \in C^\infty(\mathbb{R})$ such that $0 \leqq \xi \leqq 1, 0 \leqq \phi \leqq 1$, $\xi \equiv 1$ on $[\varepsilon, T-\varepsilon]$, $\xi \equiv 0$ on $(-\infty, \varepsilon/2] \cup (T-\varepsilon/2, \infty)$, $\phi \equiv 1$ on $[\varepsilon/4, T-\varepsilon/4]$, $\phi \equiv 0$ on $(-\infty, \varepsilon/8] \cup [T-\varepsilon/8, \infty)$. Moreover, let $\tilde{v}: \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ be defined by

$$\tilde{v}(x,s) = \begin{cases} \phi(s) v(x,s) & \text{if } T \geqq s \geqq 0, \\ 0 & \text{if } s < 0 \text{ or } s > T. \end{cases}$$

It is easy to check that $\tilde{v}$ is a viscosity supersolution of the problem

$$\xi(s) \frac{\partial \tilde{U}}{\partial t} + \inf_{y \in Y} \sup_{z \in Z} \{ -\xi(s) f(s, x, y, z) \cdot D\tilde{U} - \xi(s) l(s, x, y, z) \} = 0 \quad \text{in } \mathbb{R}^{N+1}.$$

Next let $T > t > \delta > 0$ be fixed. Then $\tilde{v}$ is also a viscosity supersolution of

$$\begin{aligned} &\frac{\partial W}{\partial \tau} + \inf_{y \in Y} \sup_{z \in Z} \left\{ -\xi(s) f(s, x, y, z) \cdot DW + \xi(s) \frac{\partial W}{\partial s} - \xi(s) l(s, x, y, z) \right\} = 0 \\ &\qquad \text{in } \mathbb{R}^{N+1} \times (0, T-t+\delta), \\ &W(x, s, 0) = \tilde{v}(x, s) \quad \text{in } \mathbb{R}^{N+1}. \end{aligned} \tag{2.8}$$

If $W \in C(\mathbb{R}^{N+1} \times [0, T-t+\delta])$ is the unique viscosity solution of (2.8), the uniqueness estimates of [6] imply

$$\tilde{v}(x,t) \geqq W(x, t, \delta).$$

Moreover, the results of L. C. Evans and P. Souganidis [11] give

$$\begin{aligned} &W(x, t, \delta) \\ &= \inf_{\beta \in \Delta(T-t)} \sup_{y \in M(T-t)} \left\{ \int_{T-t}^{T-t+\delta} \xi(T-t+\delta-t(\rho)) l(T-t+\delta-t(\rho), \tilde{x}(\rho), \beta[y](\rho))\, d\rho \right. \\ &\qquad \left. + \tilde{v}(\tilde{x}(T-t+\delta), t(T-t+\delta)) \right\} \end{aligned}$$

where for $y \in M(T-t)$ and $\beta \in \Delta(T-t)$, $\tilde{x}(\cdot)$, $t(\cdot)$ are the solution of

$$\frac{d\tilde{x}}{d\rho} = \xi(T-t+\delta-t(\rho))f(T-t+\delta-t(\rho), \tilde{x}(\rho), y(\rho), \beta[y](\rho)) \quad \text{for } T-t<\rho<T-t+\delta,$$

$$\frac{dt}{d\rho} = -\xi(T-t+\delta-t(\rho)) \quad \text{for } T-t<\rho<T-t+\delta,$$

$$\tilde{x}(T-t) = x, \qquad t(T-t) = t.$$

As $\varepsilon \downarrow 0$ the above observations imply the result, since

$$\tilde{x}(0) \to x(0) \quad \text{uniformly on } [T-t, T-t+\delta]$$

where $x(\cdot)$ is the solution of $(2.4)_{T-t}$

The next proposition considers subsolutions of (2.1) and supersolutions of (2.2). Since the proof of the results is exactly the same as the proof of Proposition 2.1, we omit it.

PROPOSITION 2.2. *Let $v, w \in C(\bar{Q}_T)$ be viscosity sub-(respectively super-)solutions of* (2.1) (*respectively* (2.2)). *For every $(x, t) \in \Omega \times (0, T)$ and $\delta > 0$ such that $\delta \leqq t$, we have*

$$v(x,t) \leqq \inf_{\beta \in \Delta(T-t)} \sup_{y \in M(T-t)} \left\{ \int_{T-t}^{(T-t+\delta)\wedge t_x} l(s, x(s), y(s), \beta[y](s))\, ds + v(x((T-t+\delta)\wedge t_x), T-((T-t+\delta)\wedge t_x)) \right\} \tag{2.9}$$

and

$$w(x,t) \geqq \sup_{\alpha \in \Gamma(T-t)} \inf_{z \in N(T-t)} \left\{ \int_{T-t}^{(T-t+\delta)\wedge t_x} l(s, x(s), \alpha[z](s), z(s))\, ds + w(x((T-t+\delta)\wedge t_x), T-((T-t+\delta)\wedge t_x)) \right\}. \tag{2.10}$$

Next we want to use Proposition 1.1 to obtain a kind of infinitesimal version of the super- and suboptimality principle of dynamic programming. To do this we have to assume, as in § 1, that we deal with locally Lipschitz viscosity super- and subsolutions. Before we state the results we need the following notation:

$$\text{for } (t, x, y) \in (0, T) \times \Omega \times Y, \quad (\mathrm{FL})(t, x, y) = \overline{\mathrm{co}}\{(f(t, x, y, z), l(t, x, y, z)) : z \in Z\} \tag{2.11}$$

and

$$\text{for } (t, x, z) \in (0, T) \times \Omega \times Z, \quad (\mathrm{FL})(t, x, z) = \overline{\mathrm{co}}\{(f(t, x, y, z), l(t, x, y, z)) : y \in Y\}. \tag{2.12}$$

The result is

PROPOSITION 2.3. *Let $v, w \in C^{0,1}_{\mathrm{loc}}(\Omega \times (0, T)) \cap C(\bar{Q}_T)$ be super-(respectively sub-)solutions of* (2.1) (*respectively* (2.2)). *For every $(x, t) \in \Omega \times (0, T)$ we have*

$$\underline{\lim}_{\delta \downarrow 0} \inf_{y \in Y} \sup_{(f, l) \in (\mathrm{FL})(T-t, x, y)} \left\{ \frac{w(x,t) - w(x+\delta f, t-\delta)}{\delta} - l \right\} \geqq 0 \tag{2.13}$$

and

$$\overline{\lim_{\delta\downarrow 0}} \sup_{z\in Z} \inf_{(f,l)\in(\mathrm{FL})(T-t,x,z)} \left\{\frac{w(x,t)-w(x+\delta f,t-\delta)}{\delta}-l\right\}\leqq 0 \tag{2.14}$$

with the inequalities being achieved as $\delta\downarrow 0$ uniformly on compact sets.

Proof. Here we show only (2.13), since (2.14) follows in a similar way. For a fixed $(x,t)\in\Omega\times(0,T)$ let K be the Lipschitz constant of v in a neighborhood of (x,t). For $\delta>0$ sufficiently small we have

$$T-t+\delta<t_{\bar{x}}$$

for every $y\in M(T-t)$, $\beta\in\Delta(T-t)$, and this uniformly for every $\bar{x}$ in a neighborhood of x. (2.6) then implies

$$\sup_{\beta\in\Delta(T-t)} \inf_{y\in M(T-t)} \left\{\frac{v(x,t)-v(x(T-t+\delta),t-\delta)}{\delta} - \frac{1}{\delta}\int_{T-t}^{T-t+\delta} l(s,x(s),y(s),\beta[y](s))\,ds\right\}\geqq 0.$$

But

$$\sup_{\beta\in\Delta(T-t)} \inf_{y\in M(T-t)} \leqq \inf_{y\in M(T-t)} \sup_{\beta\in\Delta(T-t)} \leqq \inf_{y\in Y} \sup_{\beta\in\Delta(T-t)}.$$

Therefore, in view of (2.3), (2.4), we have

$$\inf_{y\in Y} \sup_{(f,l)\in(\mathrm{FL})(T-t,x,y)} \left\{\frac{v(x,t)-v(x+\delta f,t-\delta)}{\delta}-l\right\}\geqq O(1)$$

where $O(1)\to 0$ as $\delta\downarrow 0$ uniformly for (x,t), in a compact set. Here we used the fact that for $y\in Y$

$$\left(\frac{1}{\delta}\int_{T-t}^{T-t+\delta} f(T-t,x,y,\beta[y](s))\,ds, \frac{1}{\delta}\int_{T-t}^{T-t+\delta} l(T-t,x,y,\beta[y](s))\,ds\right) \in(\mathrm{FL})(T-t,x,y)$$

for every $\beta\in\Delta(T-t)$.

Letting $\delta\downarrow 0$, we obtain the result.

As a consequence of Proposition 2.2 we have

COROLLARY 2.4. *Let $v,w\in C^{0,1}_{\mathrm{loc}}(\Omega\times(0,T))\cap C(\bar{Q}_T)$ be super-(respectively sub-)solutions of* (2.1) *(respectively* (2.2)*). For every $(x,t)\in\Omega\times(0,T)$ we have*

$$\inf_{y\in Y} \sup_{(f,l)\in(\mathrm{FL})(T-t,x,y)} \overline{\lim_{\delta\downarrow 0}} \left\{\frac{v(x,t)-v(x+\delta f,t-\delta)}{\delta}-l\right\}\geqq 0 \tag{2.15}$$

and

$$\sup_{z\in Z} \inf_{(f,l)\in(\mathrm{FL})(T-t,x,z)} \underline{\lim_{\delta\downarrow 0}} \left\{\frac{w(x,t)-w(x+\delta f,t-\delta)}{\delta}-l\right\}\leqq 0 \tag{2.16}$$

with the inequalities being achieved as $\delta\downarrow 0$ uniformly on compact sets.

Since Corollary 2.3 follows from Proposition 2.2 in the same way that Corollary 1.4 follows from Theorem 1.3 we omit its proof.

We continue with a proposition and a corollary concerning viscosity sub- and supersolutions of (2.1) and (2.2) respectively. Since these results follow from Proposition 2.2 the same way as Proposition 2.3 and Corollary 2.4 follow from Proposition

2.1 we omit their proof. We should also remark, however, that one can obtain these results directly from Proposition 2.3 and Corollary 2.4, by observing that a viscosity subsolution (supersolution) of (2.1) ((2.2)) is a viscosity subsolution (supersolution) of (2.2) ((2.1)). We have:

PROPOSITION 2.5. *Let* $v, w \in C^{0,1}_{\text{loc}}(\Omega \times (0, T)) \cap C(\bar{Q}_T)$ *be sub-(respectively super-) solutions of* (2.1) *(respectively* (2.2)). *For every* $(x, t) \in \Omega \times (0, T)$ *we have*

$$\overline{\lim_{\delta \downarrow 0}} \sup_{z \in Z} \inf_{(f, l) \in (\mathrm{FL})(T-t, x, z)} \left\{ \frac{v(x, t) - v(x + \delta f, t - \delta)}{\delta} - l \right\} \leqq 0 \tag{2.17}$$

and

$$\underline{\lim_{\delta \downarrow 0}} \inf_{y \in Y} \sup_{(f, l) \in (\mathrm{FL})(T-t, x, y)} \left\{ \frac{w(x, t) - w(x + \delta f, t - \delta)}{\delta} - l \right\} \geqq 0 \tag{2.18}$$

with the inequalities being achieved as $\delta \downarrow 0$ *uniformly on compact sets.*

COROLLARY 2.6. *Let* $v, w \in C^{0,1}_{\text{loc}}(\Omega \times (0, T)) \cap C(\bar{Q}_T)$ *be sub-(respectively super-) solutions of* (2.1) *(respectively* (2.2)). *For every* $(x, t) \in \Omega \times (0, T)$ *we have*

$$\sup_{z \in Z} \inf_{(f,l) \in (\mathrm{FL})(T-t, x, z)} \underline{\lim_{\delta \downarrow 0}} \left\{ \frac{v(x, t) - v(x + \delta f, t - \delta)}{\delta} - l \right\} \leqq 0 \tag{2.19}$$

and

$$\inf_{y \in Y} \sup_{(f, l) \in (\mathrm{FL})(T-t, x, y)} \overline{\lim_{\delta \downarrow 0}} \left\{ \frac{w(x, t) - w(x + \delta f, t - \delta)}{\delta} - l \right\} \geqq 0 \tag{2.20}$$

with the inequalities being achieved as $\delta \downarrow 0$ *uniformly on compact sets.*

The next result is the inverse of Corollary 2.4 and Corollary 2.5. In particular, it says that (2.15), (2.16), (2.19) and (2.20) together with appropriate boundary conditions characterize continuous functions as viscosity super- and subsolutions of (2.1) and (2.2).

We have:

PROPOSITION 2.7. (a) *Let* $v \in C(\Omega \times (0, T))$ *satisfy* (2.15). *Then* v *also satisfies* (0.9) *with* H *as in* (2.1).

(b) *Let* $w \in C(\Omega \times (0, T))$ *satisfy* (2.16). *Then* w *also satisfies* (0.7) *with* H *as in* (2.2).

(c) *Let* $v \in C(\Omega \times (0, T))$ *satisfy* (2.19). *Then* v *also satisfies* (0.8) *with* H *as in* (2.1).

(d) *Let* $w \in C(\Omega \times (0, T))$ *satisfy* (2.20). *Then* w *also satisfies* (0.9) *with* H *as in* (2.2).

Since the proof is similar to the proof of Proposition 1.9(a), we omit it.

We conclude this section which is an immediate consequence of Corollary 2.4 and Proposition 2.7. We have

COROLLARY 2.8. *Suppose that for every* $(t, x, p) \in [0, T] \times \bar{\Omega} \times \mathbb{R}^N$,

$$\begin{aligned} &\sup_{z \in Z} \inf_{y \in Y} \{-f(t, x, y, z) \cdot p - l(t, x, y, z)\} \\ &\qquad = \inf_{y \in Y} \sup_{z \in Z} \{-f(t, x, y, z) \cdot p - l(t, x, y, z)\}. \end{aligned} \tag{2.21}$$

Then a function $u \in C(\bar{Q}_T) \cap C^{0,1}_{\text{loc}}(\Omega \times (0, T))$ *is a viscosity solution of*

$$\frac{\partial u}{\partial t} + \sup_{z \in Z} \inf_{y \in Y} \{-f(t, x, y, z) \cdot Du - l(t, x, y, z)\} = 0 \quad \text{in } \Omega \times (0, T),$$

$$u(x, t) = g(x, t) \quad \text{on } \partial\Omega \times [0, T],$$

$$u(x, 0) = u_0(x) \text{ in } \Omega$$

if and only if u *satisfies* (2.15), (2.16) *and the correct boundary conditions.*

Remark 2.9. A result analogous to Corollary 2.8 is proved by Subbotin [28] but not in the context of viscosity solutions. In particular, in [28] (2.15) and (2.16) are necessary and sufficient conditions for a locally Lipschitz continuous function to be the value of a positional differential game, under the assumption that $l \equiv 0$ and $\Omega = \mathbb{R}^N$. Corollary 2.8 also implies in view of the results of [26], [27], [10], [11], [1], that the notion of the value of a positional differential game is the same as the value of differential game introduced by W. Fleming, A. Friedman and N. Elliott and J. Kalton.

Remark 2.10. A remark analogous to Remark 1.6 holds here too.

Remark 2.11. A result analogous to Proposition 1.11 holds here too.

Appendix. In view of Remark 1.6 and Remark 2.10, we want to make some (classical) observations concerning the existence of directional derivatives of the value function of optimal control and differential games problems. For simplicity here we investigate the case of an infinite horizon optimal control problem in $\mathbb{R}^N$. In particular, we deal with the existence of

$$\lim_{h \downarrow 0} \frac{v(x+h\chi) - v(x)}{h}$$

for all $x, \chi \in \mathbb{R}^N$, where v is the value function. Using the notation of § 1, let us also assume:

For every $x \in \mathbb{R}^N$, $y \in Y$ and $h \in \mathbb{R}$

$$|f(x+h, y) - f(x, y) - D_x f(x, y) \cdot h| \leqq |h| \varepsilon(|h|)$$

(1) and

$$|l(x+h, y) - l(x, y) - D_x l(x, y) \cdot h| \leqq |h| \varepsilon(|h)$$

where $\varepsilon(|h|) \to 0$ as $|h| \to 0$.

For every $y(\cdot) \in M$, let

$$J(x, y) = \int_0^\infty e^{-s} l(x(s), y(s))\, ds \tag{2}$$

where $x(\cdot)$ is the solution of (1.3) with $x(0) = x$. Moreover, let

$$v(x) = \inf_{y \in M} J(x, y). \tag{3}$$

In view of the discussion in § 1 and the references given there, v is the value function of the associated optimal control problem.

We have

PROPOSITION A.1. *Assume that* (1.2) *and* (1) *hold with*

$$1 > \sup_{(x, y) \in \mathbb{R}^N \times Y} |D_x f(x, y)|. \tag{4}$$

Let v be given by (3). *Then*

$$\lim_{h \downarrow 0} \frac{v(x+h\chi) - v(x)}{h}$$

exists for every $x, \chi \in \mathbb{R}^N$ and

$$\lim_{h \downarrow 0} \frac{v(x+h\chi) - v(x)}{h} = \inf \left\{ \lim_{n \to \infty} \frac{\partial J(x_n, y_n)}{\partial \chi} : y_n \in M, J(x_n, y_n) \xrightarrow[u \to \infty]{} v(x) \right\}. \tag{5}$$

Proof. The proof is a consequence of the following lemma.

Lemma A.2. *Let* $w(x)=\inf_i w^i(x)$ *with* w, w^i *equibounded, equicontinuous and satisfying*:

$$\forall \chi \in \mathbb{R}^N, |\chi|=1 \text{ there exist } \partial w^i(x)/\partial \chi \text{ such that}$$

$$\left| \frac{w^i(x+h\chi)-w^i(x)}{h} - \frac{\partial w^i}{\partial \chi}(x) \right| \leqq \delta(h) \underset{h\downarrow 0}{\longrightarrow} 0. \tag{6}$$

Then $\lim_{h\downarrow 0^+}(w(x+h\chi)-w(x))/h$ *exists for all* χ *and is equal to*

$$\lim_{h\downarrow 0^+} \frac{w(x+h\chi)-w(x)}{h} = \inf\left\{ \underline{\lim}_{n\to\infty} \frac{\partial w^{i_n}}{\partial \chi}(x) : w^i(x) \underset{n\to\infty}{\longrightarrow} w(x) \right\}. \tag{7}$$

In view of our hypotheses, v and $J(\cdot, y)$ satisfy the assumptions of Lemma 2. Therefore here we only prove the lemma. We have:

Proof of Lemma 2. Let i_n be a sequence such that

$$w^{i_n}(x) \to w(x) \quad \text{as } n \to \infty.$$

Then

$$\begin{aligned} \frac{w(x+h\chi)-w(x)}{h} &\leqq \frac{w^{i_n}(x+h\chi)-w^{i_n}(x)}{h} + \frac{w^{i_n}(x)-w(x)}{h} \\ &\leqq \frac{\partial w^{i_n}}{\partial \chi}(x) + \delta(h) + \frac{|w^{i_n}(x)-w(x)|}{h} \\ &\leqq \underline{\lim}_{n\to\infty} \frac{\partial w^{i_n}}{\partial \chi}(x) + \delta(h). \end{aligned}$$

Therefore

$$\overline{\lim_{h\downarrow 0}} \frac{w(x+h\chi)-w(x)}{h} \leqq \alpha$$

where α is the right-hand side of (7). For the other direction, let $h_n>0$ be such that $h_n \downarrow 0$ as $n\to\infty$. Choose i_n such that

$$v(x+h\chi) \leqq v^{i_n}(x+h_n\chi) \leqq v(x+h_n\chi) + \frac{h_n}{n} \quad \text{as } n\to\infty.$$

Then, in view of the assumptions,

$$v^{i_n}(x) \to v(x) \quad \text{as } n\to\infty.$$

We have

$$\frac{v^{i_n}(x+h_n\chi)-v^{i_n}(x)}{h_n} \leqq \frac{v(x+h_n\chi)-v(x)}{h_n} + \frac{1}{n}$$

which implies

$$\alpha \leqq \underline{\lim_{h\downarrow 0}} \frac{v(x+h\chi)-v(x)}{h}$$

and thus the result.

Remark 3. Results analogous to the above also hold for finite horizon control problems and differential games. In the finite horizon case, one does not have to assume (4).

REFERENCES

[1] N. E. BARRON, L. C. EVANS AND R. JENSEN, *Viscosity solutions of Isaacs' equations and differential games with Lipschitz controls*, J. Differential Equations, to appear.

[2] I. CAPUZZO DOLCETTA, *On a discrete approximation of the Hamilton-Jacobi equation of dynamic programming*, Appl. Math. Optim., to appear.

[3] I. CAPUZZO DOLCETTA AND L. C. EVANS, *Optimal switching for ordinary differential equations*, this Journal, to appear.

[4] I. CAPUZZO DOLCETTA AND H. ISHII, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., to appear.

[5] M. G. CRANDALL, L. C. EVANS AND P.-L. LIONS, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 282 (1984), pp. 487-502.

[6] M. G. CRANDALL AND P.-L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[7] M. G. CRANDALL AND P. E. SOUGANIDIS, *Developments in the theory of nonlinear first-order partial differential equations*, Proc. International Symposium on Differential Equations, Birmingham, AL, 1983, Knowles and Lewis, eds., North-Holland, Amsterdam.

[8] R. J. ELLIOTT AND N. J. KALTON, *The existence of value in differential games*, Mem. Amer. Math. Soc., 126, 1972.

[9] ———, *Cauchy problems for certain Isaacs-Bellman equations and games of survival*, Trans. Amer. Math. Soc., 198 (1974), pp. 45-72.

[10] L. C. EVANS AND H. ISHII, *Nonlinear first order* PDE *on bounded domains*, to appear.

[11] L. C. EVANS AND P. E. SOUGANIDIS, *Differential games and representation formulas for solutions of Hamilton-Jacobi-Isaacs equations*, Indiana Univ. Math. J., to appear.

[12] W. H. FLEMING, *The convergence problem for differential games*, J. Math. Anal. Appl., 3 (1961), pp. 102-116.

[13] ———, *The convergence problem for differential games* II, in Advances in Game Theory, Ann. Math. Studies 52, Princeton Univ., Princeton, NJ.

[14] ———, *The Cauchy problem for degenerate parabolic equations*, J. Math. Mech., 13 (1964), pp. 987-1008.

[15] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[16] A. FRIEDMAN, *Differential Games*, John Wiley, New York, 1971.

[17] ———, *Differential Games*, CBMS Regional Conference Series in Applied Mathematics 18, American Mathematical Society, Providence, RI, 1974.

[18] R. ISAACS, *Differential Games*, John Wiley, New York, 1965.

[19] H. ISHII, *Viscosity solutions of Hamilton-Jacobi equations with discontinuous Hamiltonian and differential games*, in preparation.

[20] N. N. KRASSOVSKII AND A. I. SUBBOTIN, *Positional Differential Games*, Nauka, Moscow, 1974. (In Russian.)

[21] P.-L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, Boston, 1982.

[22] ———, *Some recent results in the optimal control of diffusion processes*, to appear in Stochastic Analysis, Proceedings of the Tanigachi International Symposium on Stochastic Analysis, Katata and Kysto, 1982, Kinokuniya, Tokyo, 1983.

[23] ———, Articles on Hamilton-Jacobi-Bellman and Isaacs equations in the Encyclopaedia of Systems and Control, Pergamon, Oxford, 1984.

[24] ———, *Optimal control of diffusion processes and Hamilton-Jacobi-Bellman equations, Parts* 1, 2, Comm. PDE, 8 (1983), pp. 1101-1174, 1229-1276.

[25] P.-L. LIONS AND M. NISIO, *A uniqueness result for the semigroup associated with the Hamilton-Jacobi-Bellman operator*, Proc. Japan Acad., 58 (1982), pp. 273-276.

[26] P. E. SOUGANIDIS, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, Mathematics Research Center TSR 2511, Univ. Wisconsin, Madison.

[27] ———, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations with applications to differential games*, J. Nonlinear Anal. TMA, to appear.

[28] A. I. SUBBOTIN, *A generalization of the basic equation of the theory of differential games*, Soviet Math. Dokl., 22 (1980), pp. 358-362.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

006

SOME EXAMPLES OF REACHABLE SETS AND OPTIMAL COST FUNCTIONS THAT FAIL TO BE SUBANALYTIC*

S. LOJASIEWICZ, JR.† AND H. J. SUSSMANN‡

Abstract. We give examples of: a) a control system $\dot{x} = f(x) + Bu$, with f a vector field whose components are quadratic polynomials, and with the controls taking values in the unit cube, such that the time T reachable sets from the origin are not subanalytic; b) a system $\dot{x} = g(x) + Bu$, with the components of g cubic polynomials, such that the system is completely controllable but the optimal time function is not subanalytic; c) a linear system with a compact, convex, semialgebraic control constraint for which the time T reachable sets are not subanalytic.

Key words. time-optimal control, value function, subanalytic sets

1. Introduction. An important question in control theory is that of the smoothness, or at least piecewise smoothness, of reachable sets. It is natural to conjecture that, under fairly general conditions, reachable sets are finite or locally finite unions of smooth embedded submanifolds. Another intimately related question is that of the piecewise smoothness of optimal cost functions. So far, the only general theorems that have been proved in this direction are those based on a method introduced by Brunovsky in [1]: one applies the maximum principle or some other necessary conditions for optimality to deduce that all points that can be reached at all can actually be reached by means of a trajectory of a particularly simple kind (e.g. bang-bang, or a finite concatenation of bang-bang and singular arcs), and one uses this to prove that the reachable sets must be subanalytic. (For the definition and basic properties of subanalytic sets, cf. [7] or [9].) Once this is established, piecewise smoothness follows. A similar approach has been used to prove piecewise regularity of optimal cost functions and existence of regular synthesis (cf. [1], [2], [3], [4], [8], [9], [10] and especially [7]).

The purpose of this paper is to present some negative results that point to the limitations of the above method. We will give examples of very simple systems for which the reachable sets or the optimal cost function fail to be subanalytic.

This does not yet settle the question whether such sets and functions are piecewise smooth under general conditions, but it establishes that, if the answer is going to be positive, then new ideas will necessarily have to be involved, perhaps using classes of sets more general than that of subanalytic sets, but for which reasonable stratification theorems can still be proved.

We shall not attempt to give a general definition of what is meant by a "simple system," but the systems to be exhibited here are clearly very simple in almost any sense of the word, since they are "almost" like linear systems with a polyhedral constraint, except for the fact that, in two of our examples, the linear term Ax is replaced by a polynomial, and in the third example the polyhedral constraint is replaced by a set of inequalities involving quadratic polynomials.

The systems considered here will be of the general form

$$\dot{x} = f(x) + \sum_{i=1}^{m} u_i g_i(x) \tag{1.1}$$

* Received by the editors January 10, 1984, and in final form July 17, 1984.

† Institute of Mathematics, Polish Academy of Sciences, 31-027 Kraków, ul. Solskiego 30, Poland.

‡ Mathematics Department, Rutgers University, New Brunswick, New Jersey 08903. The work of this author was partially supported by the National Science Foundation under grant MCS78-02442-03.

where

(a) the state x is in $\mathbb{R}^n$,

(b) $f, g_1, \cdots, g_m$ are real analytic vector fields on $\mathbb{R}^n$,

(c) the control $u = (u_1, \cdots, u_m)$ takes values in a compact convex subset K of $\mathbb{R}^m$,

(d) the admissible controls are arbitrary measurable K-valued functions defined on an arbitrary interval $[a, b] \subseteq \mathbb{R}$,

(e) for any given admissible $u(\cdot):[a, b] \to K$ and any initial condition $x(a)$, the corresponding trajectory $\{x(t)\}$ is defined for all $t \in [a, b]$.

For a system (1.1) as above, we can define the *attainable sets* (or *reachable sets*) $A_p(T)$, $A_p(\leqq T)$ as follows: $A_p(T)$ is the set of all points $q \in \mathbb{R}^n$ that can be reached from p in time T, and

$$A_p(\leqq T) = \bigcup_{0 \leqq t \leqq T} A_p(T). \tag{1.2}$$

It is well known that, under the above hypotheses (especially condition (e)) the sets $A_p(T)$, $A_p(\leqq T)$ are compact.

It is natural to ask whether the sets $A_p(T)$, $A_p(\leqq T)$ have a "nice" structure, and it is known that the answer to this question is "yes" in some interesting cases. For instance, if (1.1) is a linear system (i.e. $f(x) = Ax$, $g_i(x) = b_i$) and K is a polyhedron, then it is known that the sets $A_p(T)$, $A_p(\leqq T)$ are subanalytic. This implies, in particular, that $A_p(T)$ and $A_p(\leqq T)$ are finite unions of connected embedded analytic submanifolds of $\mathbb{R}^n$.

In § 2 of this paper we present an example of a system of the form (1.1) for which it is not true that the sets $A_p(T)$, $A_p(\leqq T)$ are subanalytic. Moreover, the system will still have a polyhedral control set K, and will not be too far from linear. Precisely, our system will be of the form

$$\dot{x} = f(x) + Bu \tag{1.3}$$

where $B = (b_1, \cdots, b_m)$ is a constant matrix and the components $f_1, \cdots, f_n$ of the map $f: \mathbb{R}^n \to \mathbb{R}^n$ are polynomials of degree $\leqq 2$. We will show that the sets $A_0(2)$ and $A_0(\leqq 2)$ are not subanalytic.

In the second part of the paper we use a slightly modified version of the system (1.3) to exhibit an example of another unpleasant phenomenon. We construct a system

$$\dot{z} = g(z) + Dv, \tag{1.4}$$

which is *completely controllable* (i.e. for every pair p, q of states there exists a trajectory that goes from p to q) and for which the *optimal time function* V is not subanalytic. (Recall that a function $\phi: \mathbb{R}^k \to \mathbb{R}$ is said to be *subanalytic* if its graph $\{(x, y): \phi(x) = y\}$ is a subanalytic subset of $\mathbb{R}^k \times \mathbb{R}$. For the precise definition of V, see § 3.) In this example, the components of the vector field g are polynomials of degree not greater than three.

We do not know whether the reachable sets discussed in our first example are finite unions of submanifolds. Nor do we know whether the function V of our second example is piecewise smooth in some reasonable sense. It would be desirable to know whether or not this is the case, or to construct other examples where V fails to be "piecewise smooth" but the control system itself is real analytic. In particular, it seems reasonable to expect that some modification of the well-known Fuller example might be used for that purpose, but it is not clear to us how to do it. (For Fuller's example, cf. [6].)

When the control constraint set K is not a polyhedron, the reachable sets may

fail to be subanalytic even for a *linear* system and a "very nice" K. In fact, we give an example of a linear system in $\mathbb{R}^3$, with two inputs, and a control constraint set $K=\{(u_1, u_2)\colon u_1^2-1\leqq 2u_2\leqq 1-u_1^2\}$, for which the reachable sets are not subanalytic. This example is described in § 4.

The remainder of this section is devoted to outlining the idea that underlies the construction of the example of § 2. Although the example itself appears to be extremely complicated (21 variables, 10 controls), the idea is quite simple, as we now explain.

Our point of departure is the initial value problem

$$c\ddot{x}+x=0,\quad x(0)=c,\quad \dot{x}(0)=0,\quad 0<c\leqq 1 \tag{1.5}$$

whose solution is

$$x(t)=c\cos\left(\frac{t}{\sqrt{c}}\right). \tag{1.6}$$

We can rewrite the equation $c\ddot{x}+x=0$ as a system in three variables x, y, c:

$$\dot{x}=y,\quad c\dot{y}=-x,\quad \dot{c}=0. \tag{1.7}$$

The initial conditions of (1.5) single out the half-open segment

$$L_0=\{(c, 0, c)\colon 0<c\leqq 1\} \tag{1.8}$$

in (x, y, c) space. If we let each of the points of L_0 evolve following trajectories of (1.5), up to time $t=1$, we get the curve

$$L_1=\left\{\left(c\cos\left(\frac{1}{\sqrt{c}}\right), -\sqrt{c}\sin\left(\frac{1}{\sqrt{c}}\right), c\right)\colon 0<c\leqq 1\right\}. \tag{1.9}$$

The set L_1 is not subanalytic because, if we let P denote the plane $y=0$, then $L_1\cap P$ is the set of points $((-1)^k/\pi^2k^2, 0, 1/\pi^2k^2)$, i.e. an infinite sequence that has a finite limit point.

The preceding considerations do not yet provide an example of the kind we want, for two reasons:

(a) the equations (1.7) are not a system of first-order differential equations of the usual kind, because the equation $c\dot{y}=-x$ is not of the form $\dot{y}=\phi(c, x, y)$ with ϕ real analytic in c, x, y,

(b) the initial condition $(x, y, c)\in L_0$ does not specify a single initial point, but a whole segment.

However, we can modify our construction so as to obtain an example of the desired kind. We do this in two steps, by first taking care of (a), and then of (b).

To take care of (a) we must find a way of forcing $c\dot{y}$ to be equal to $-x$ without actually writing "$c\dot{y}=-x$." We can do this by introducing two new variables p, q with initial conditions

$$p(0)=q(0)=0, \tag{1.10}$$

and equations

$$\dot{p}=cy+q,\qquad \dot{q}=x. \tag{1.11a, b}$$

If we succeed in forcing p to equal 0, then we will get $\ddot{p}\equiv 0$, i.e. $c\dot{y}=-x$ (given that $\dot{c}=0$). To force p to equal 0, we introduce another variable s, and impose the boundary conditions

$$s(0)=s(1)=0 \tag{1.12}$$

and the equation

(1.11c) $$\dot{s} = p^2.$$

We now have a total of six variables (x, y, c, p, q and s), the three equations (1.11a, b, c), as well as the two equations

(1.11d, e) $$\dot{x} = y, \qquad \dot{c} = 0,$$

giving a total of five equations which express the time-derivatives of five of our fix variables as functions of these variables. On the missing variable y, we impose no explicit restriction. We allow y to vary freely, by writing

(1.11f) $$\dot{y} = u_y$$

where u_y is a control. The terminal condition $s(1) = 0$, together with the initial conditions $s(0) = p(0) = q(0) = 0$, will have the effect of forcing $c\dot{y}$ to equal $-x$.

The six equations (1.11a, b, c, d, e, f) define a control system with a single scalar input u_y, in the space $\mathbb{R}^6$ of the six variables x, y, c, p, q and s. If we let $\tilde{L}_0$ denote the segment defined by

(1.13) $$y = p = q = s = 0, \qquad x = c, \quad 0 < c \leqq 1,$$

and we let $\tilde{L}_1$ denote the set of all points that are reachable in time 1 by some trajectory of (1.11) from some point of $\tilde{L}_0$, then the set $\tilde{L}_1$ can be proved not to be subanalytic. Indeed, if we let H denote the hyperplane $s = 0$, then $\tilde{L}_1 \cap H$ is precisely the set $L_1 \times \{(0, 0, 0)\}$ because, if a trajectory $t \to x(t), \cdots, s(t)$ of (1.11) satisfies $s(0) = s(1) = 0$, then $p(t)$ must equal 0 for all $t \in [0, 1]$ and then, as explained before, the equation $c\dot{y}(t) = -x(t)$ will hold, so that $t \to (x(t), y(t), c(t))$ is actually a trajectory of our first system.

So we have succeeded in exhibiting a control system Σ_1 with the property that the time 1 reachable set from an initial segment $\tilde{L}_0$ is not subanalytic. Our next task is to take care of difficulty (b), i.e. to modify the system so that the initial condition that gives rise to a nonsubanalytic reachable set is a point, rather than a segment. It is quite easy to produce a system Σ_2 such that the time one reachable set from some point is a segment. If we then follow the trajectories of system Σ_1 from that segment, during one unit of time, we will get our nonsubanalytic reachable set. Our problem is to produce this "system switching" within one single system, i.e. to find one system whose trajectories will be forced to be those of Σ_2 for $t \leqq 1$, and then those of Σ_1 for $t \geqq 1$. So we must find a way to write equations for x, y, c, p, q, s and other variables as well, so that the equations of Σ_1 will have to hold for x, y, c, p, q, s for $1 \leqq t \leqq 2$, while enough freedom is preserved so that, for instance, points with $x = 0, y = c > 0$ will be reachable from the origin in time 1.

Recall that the equations $\dot{s} = p^2, \dot{p} = cy + q$ force $cy + q$ to vanish identically along trajectories that satisfy the terminal conditions $s(0) = s(1) = 0$. Our first task is to modify these equations so as to force $cy + q$ to vanish identically for $1 \leqq t \leqq 2$, without forcing it to vanish identically for all t. We will do this while not modifying the equation $\dot{p} = cy + q$. So we must force p to be constant on $[1, 2]$. To do this, we introduce new variables $\xi, \alpha_\xi, \beta_\xi, \gamma_\xi$ and controls u_ξ, v_ξ, w_ξ, subject to the equations

(1.14a) $$\dot{\alpha}_\xi = (1 - t)\xi + 1 - u_\xi,$$

(1.14b) $$\dot{\beta}_\xi = (2 - t)\xi + 1 - v_\xi,$$

(1.14c) $$\dot{\xi} = w_\xi,$$

(1.14d) $$\dot{\gamma}_\xi = \xi,$$

and then we replace $\dot{s} = p^2$ by

(1.15) $$\dot{s} = (p - \gamma_\xi)^2 + \alpha_\xi^2 + \beta_\xi^2.$$

If the terminal conditions $s(0) = s(2) = 0$ hold, then (1.15) forces α_ξ and β_ξ to vanish identically, and p to equal γ_ξ. If the controls u_ξ, v_ξ are $\leqq 1$, then (1.14a, b) force $(1-t)\xi$ to be $\leqq 0$, and $(2-t)\xi$ to be $\leqq 0$ as well. Since $1-t$ and $2-t$ have opposite signs on [1, 2] this forces ξ to vanish on [1, 2], and so γ_ξ is constant on [1, 2], and therefore p is constant on [1, 2].

To get c to be constant on [1, 2] we could use a similar trick. We could introduce variables $\eta, \alpha, \beta_\eta, \gamma_\eta$ and controls u_η, v_η, w_η, subject to equations exactly like (1.14a, b, c, d), with ξ replaced by η. We would then add the sum $(c-\gamma_n)^2 + \alpha_\eta^2 + \beta_\eta^2$ to the right side of (1.15), and then we will have forced c to be constant on [1, 2].

There is, however, a slight drawback to using this approach to make c constant on [1, 2]. We want to force c, x and y to satisfy $c(1) = x(1) > 0$, $y(1) = 0$. The equation $\dot{\alpha}_\eta = (1-t)\eta + 1 - u_\eta$ will force η to be $\leqq 0$ on [0, 1], and so γ_η will be $\leqq 0$ on [0, 1] (if $\gamma_\eta(0) = 0$). So c will be $\leqq 0$ on [0, 1] (because $c \equiv \gamma_\eta$) and this will contradict $c(1) > 0$. This can easily be taken care of, by forcing c to equal $-\gamma_\eta$ rather than γ_η, i.e. by inserting the term $(c+\gamma_\eta)^2$, rather than $(c-\gamma_\eta)^2$, in the right side of (1.15). However, if we do this, we will get $\dot{c} = -\dot{\gamma}_\eta = -\eta$, and so $\dot{c}$ will be $\geqq 0$ on [0, 1]. Therefore $c \geqq 0$ on [0, 1]. As we will see below, the equality $x(1) = c(1)$ will be enforced by making x equal to c on [0, 1]. So, if $c \geqq 0$ on [0, 1], then $x \geqq 0$ on [0, 1]. Since $q(0) = 0$, we find that $q(1)$ can only vanish if $x \equiv 0$ on [0, 1], in which case $x(1) = 0$ and $c(1) = 0$. On the other hand, since we will be forcing $cy + q$ to vanish on [1, 2], and $y(1)$ to be equal to zero, $q(1)$ will have to vanish. So, if we use the approach outlined above, we will not be able to get the condition $c(1) > 0$ satisfied.

So we must use a slightly different approach to force c to be constant on [1, 2]. We introduce $\eta, \alpha_\eta, \beta_\eta, \gamma_\eta, u_\eta, v_\eta, w_\eta$ as above, subject to the four equations

(1.16a) $$\dot{\alpha}_\eta = (1-t)\eta + 1 - u_\eta,$$

(1.16b) $$\dot{\beta}_\eta = (2-t)\eta + 1 - v_\eta,$$

(1.16c) $$\dot{\eta} = w_\eta,$$

(1.16d) $$\dot{\gamma}_\eta = \eta,$$

but we do not attempt to force c to equal γ_η (or $-\gamma_\eta$). Instead, we introduce new variables λ, b, subject to the equations

(1.17a) $$\dot{c} = b,$$

(1.17b*) $$\dot{\lambda}(t) = b\phi(t),$$

(1.17c) $$\dot{b} = u_b,$$

where u_b is a new control, and ϕ is a function that changes sign on [0, 1], but has isolated zeros. We then force λ to coincide with γ_η by inserting the term $(\lambda - \gamma_\eta)^2$ in the right side of (1.15). In this way, we guarantee that b will vanish, and c will be constant, on [1, 2], without forcing c to have constant sign on [0, 1].

Since (1.17b*) is not autonomous, we render it autonomous by writing instead, for instance,

(1.17b) $$\dot{\lambda} = b\mu,$$

(1.17d) $$\dot{\mu} = 2 - 10t.$$

(Notice that t can be considered as a state variable, with equation $\dot{t} = 1$.)

The solution of (1.17d) with initial condition $\mu(0) = 0$ is the polynomial $2t - 5t^2$, which satisfies the conditions that were required of $\phi(t)$.

Finally, we must force $x(1)$ to equal $c(1)$, and $y(1)$ to equal 0. We achieve this by introducing yet another variable σ, together with variables $\alpha_\sigma, \beta_\sigma$, subject to the equations

(1.18a) $$\dot{\alpha}_\sigma = -t\sigma + 1 - u_\sigma,$$

(1.18b) $$\dot{\beta}_\sigma = (1-t)\sigma + 1 - v_\sigma,$$

where u_σ, v_σ are new controls with values in $[-1, 1]$. As before, we insert $\alpha_\sigma^2 + \beta_\sigma^2$ in the right side of (1.15). This forces σ to vanish on $[0, 1]$. We then demand that σ itself satisfy the equation

(1.18c) $$\dot{\sigma} = (x - c)^2.$$

This will force x to equal c on $[0, 1]$. Since we are already forcing b to vanish on $[1, 2]$, we have $\dot{c}(1) = 0$, i.e. $\dot{x}(1) = 0$, i.e. $y(1) = 0$.

Summarizing, we have a set of 21 variables, namely, t, y, x, q, b, c, σ, α_σ, β_σ, p, ξ, γ_ξ, α_ξ, β_ξ, μ, λ, η, γ_η, α_η, β_η and s, which we will now relabel as $x_0, x_1, \cdots, x_{20}$, and ten controls u_x, u_b, u_ξ, v_ξ, w_ξ, u_η, v_η, w_η, u_σ and v_σ, which we will relabel as $u_1, \cdots, u_{10}$. The equations listed in § 2 are precisely the ones we have derived, namely:

(i) the fifteen equations (1.14a, b, c, d), (1.16a, b, c, d), (1.17a, b, c, d), (1.18a, b, c),

(ii) the equations $\dot{t} = 1$, $\dot{x} = y$, $\dot{y} = u_y$, $\dot{p} = cy + q$, $\dot{q} = x$, and

(iii) the final version of (1.15), obtained after all the squares of functions that will be forced to vanish are added to the right side of (1.15), which gives

(1.19) $$\dot{s} = (p - \gamma_\xi)^2 + \alpha_\xi^2 + \beta_\xi^2 + \alpha_\eta^2 + \beta_\eta^2 + (\lambda - \gamma_\eta)^2 + \alpha_\sigma^2 + \beta_\sigma^2.$$

2. The main example. We let $n = 21$, $m = 10$. We use $x_0, \cdots, x_{20}$ for the coordinates in $\mathbb{R}^{21}$. The control set is the cube

(2.1) $$K = \{(u_1, \cdots, u_{10}) : |u_i| \leqq 1,\ i = 1, \cdots, 10\}.$$

Equation (1.3) will now be written in full, as a system of 21 scalar equations. For convenience, we divide the 21 equations into five groups.

(I) a) $\dot{x}_0 = 1$, b) $\dot{x}_1 = u_1$,
c) $\dot{x}_2 = x_1$, d) $\dot{x}_3 = x_2$, e) $\dot{x}_4 = u_2$,
f) $\dot{x}_5 = x_4$.

(II) a) $\dot{x}_6 = (x_2 - x_5)^2$,
b) $\dot{x}_7 = -x_0 x_6 + 1 - u_3$,
c) $\dot{x}_8 = (1 - x_0)x_6 + 1 - u_4$.

(III) a) $\dot{x}_9 = x_1 x_5 + x_3$,
b) $\dot{x}_{10} = u_5$,
c) $\dot{x}_{11} = x_{10}$,
d) $\dot{x}_{12} = (1 - x_0)x_{10} + 1 - u_6$,
e) $\dot{x}_{13} = (2 - x_0)x_{10} + 1 - u_7$.

(IV) a) $\dot{x}_{14} = 2 - 10x_0$,
b) $\dot{x}_{15} = x_4 x_{14}$,
c) $\dot{x}_{16} = u_8$,
d) $\dot{x}_{17} = x_{16}$,
e) $\dot{x}_{18} = (1 - x_0)x_{16} + 1 - u_9$,
f) $\dot{x}_{19} = (2 - x_0)x_{16} + 1 - u_{10}$.

(V) $\dot{x}_{20} = x_7^2 + x_8^2 + (x_9 - x_{11})^2 + x_{12}^2 + x_{13}^2 + (x_{15} - x_{17})^2 + x_{18}^2 + x_{19}^2$.

The system Σ defined by the above equations satisfies all the hypotheses of § 1. (To verify the nonexplosion condition (e), observe that each of the 21 equations is of the form $\dot{x}_g i = \psi_i(x_0, \cdots, x_{i-1}, u_1, \cdots, u_{10})$, and therefore the solution can be found by successive integrations.)

From now on, we will only be dealing with this particular system, and with the attainable sets from $p = 0$, which we denote by $A(T)$, $A(\leqq T)$. Our goal is to prove that $A(2)$ and $A(\leqq 2)$ are not subanalytic.

First, we list some properties of subanalytic sets:

(i) Every finite set is subanalytic.

(ii) Every set of the form $\{x: \phi(x) \in A\}$, where $\phi: \mathbb{R}^n \to \mathbb{R}^m$ is a real analytic function, and A is subanalytic in $\mathbb{R}^m$, is subanalytic in $\mathbb{R}^n$.

(iii) The intersection of two subanalytic sets is subanalytic.

(iv) Every bounded subanalytic subset of $\mathbb{R}^n$ has a finite number of connected components.

In view of (iv), every hyperplane is subanalytic. Clearly, $A(2)$ is the intersection of $A(\leqq 2)$ with the hyperplane $x_0 = 2$ (because of (Ia)). So, if $A(\leqq 2)$ were subanalytic, it would follow that $A(2)$ is subanalytic. Therefore we will only study $A(2)$. We let S be the linear subspace of $\mathbb{R}^{21}$ defined by $x_1 = x_{20} = 0$. In view of (i) and (ii), S is subanalytic. Let $B = A(2) \cap S$. If we prove that B is not subanalytic, then it clearly follows that $A(2)$ is not subanalytic. On the other hand, the nonexplosion condition (e) implies that $A(2)$ is compact. Therefore B is compact. We will prove that B has infinitely many connected components. In view of (iv), this will imply that B is not subanalytic.

To prove that B has infinitely many components, we construct an infinite sequence $H_0, H_1, H_2, \cdots$ of pairwise disjoint hyperplanes, and we prove that

$$B \subseteq \bigcup_{k=0}^{\infty} H_k, \tag{2.2}$$

and

$$B \cap H_k \neq \varnothing \quad \text{for } k > 10. \tag{2.3}$$

The desired conclusion clearly follows from (2.2) and (2.3).

The H_k are constructed as follows: H_k is the hyperplane given by $x_5 = c_k$, where the constants c_k are defined by $c_0 = 0$ and

$$c_k = \frac{1}{\pi^2 k^2} \quad \text{for } k = 1, 2, \cdots. \tag{2.4}$$

First we prove (2.2). Let $I_1 = [0, 1]$, $I_2 = [1, 2]$, $I = [0, 2]$. Let $y \in B$, and let $u(\cdot): [0, 2] \to K$ be an admissible control whose corresponding trajectory $x(\cdot): [0, 2] \to \mathbb{R}^{21}$, with initial condition $x(0) = 0$, satisfies $x(2) = y$. Let $x_0(\cdot), \cdots, x_{20}(\cdot)$ be the components of the vector-valued function $x(\cdot)$. We will now derive a series of condi-

tions that these functions must satisfy. To simplify the formulas, we will often write x_i rather than $x_i(t)$.

Since $y \in B$, we have $x_{20}(2) = 0$ and so (V) implies that the following equalities hold identically on I:

$$x_7 \equiv x_8 \equiv x_{12} \equiv x_{13} \equiv x_{18} \equiv x_{19} \equiv 0, \tag{2.5}$$

$$x_9 \equiv x_{11}, \qquad x_{15} \equiv x_{17}. \tag{2.6}$$

Then (IIb) gives $x_0 x_6 \equiv 1 - u_3$, and (IIc) implies $(x_0 - 1)x_6 \equiv 1 - u_4$. Since $1 - u_3 \geqq 0$, $1 - u_4 \geqq 0$, we get $x_6 \geqq 0$ and $x_6 \leqq 0$ on I_1.

Therefore

$$x_6 \equiv 0 \quad \text{on } I_1. \tag{2.7}$$

Then (IIa) gives

$$x_2 \equiv x_5 \quad \text{on } I_1. \tag{2.8}$$

From $x_{12} \equiv x_{13} \equiv 0$ and (IIId), (IIIe), we get $(1 - x_0)x_{10} + 1 - u_6 \equiv 0$, which implies that $(1 - x_0)x_{10} \leqq 0$, and $(2 - x_0)x_{10} + 1 - u_7 \equiv 0$, which implies that $(2 - x_0)x_{10} \leqq 0$. Since $1 - x_0 < 0$ and $2 - x_0 > 0$ on $]1, 2[$, we conclude that

$$x_{10} \equiv 0 \quad \text{on } I_2. \tag{2.9}$$

Since $x_9 \equiv x_{11}$, we get from (IIIa) that

$$x_1 x_5 + x_3 \equiv x_{10} \quad \text{on } I. \tag{2.10}$$

Therefore

$$x_1 x_5 + x_3 \equiv 0 \quad \text{on } I_2. \tag{2.11}$$

Since $x_{18} \equiv x_{19} \equiv 0$, a reasoning similar to that used for x_{10} shows that

$$x_{16} \equiv 0 \quad \text{on } I_2. \tag{2.12}$$

On the other hand, we know that $x_{15} \equiv x_{17}$. So (IVb) and (IVd) yield

$$x_4 x_{14} \equiv x_{16}. \tag{2.13}$$

Therefore

$$x_4 x_{14} \equiv 0 \quad \text{on } I_2. \tag{2.14}$$

The function $x_{14}(\cdot)$ can be computed from (IVa). We get

$$x_{14}(t) = t(2 - 5t). \tag{2.15}$$

In particular, $x_{14}(t) \neq 0$ for $t \in I_2$, and so

$$x_4 \equiv 0 \quad \text{on } I_2. \tag{2.16}$$

By (If), x_5 is constant on I_2. Let c denote the value of $x_5(t)$ for $t \in I_2$. Then (2.11) says that

$$cx_1 + x_3 \equiv 0 \quad \text{on } I_2. \tag{2.17}$$

Therefore

$$c\dot{x}_1 + \dot{x}_3 \equiv 0 \quad \text{on } I_2. \tag{2.18}$$

Using $\dot{x}_3 = x_2$, $\dot{x}_2 = x_1$, we conclude that

$$c\ddot{x}_2 + x_2 \equiv 0 \quad \text{on } I_2. \tag{2.19}$$

Clearly, if we compute $x_2(1)$ and $\dot{x}_2(1)$, the function $x_2(\cdot)$ will be completely determined on I_2 by these data, together with c.

Since $x_2 \equiv x_5$ on I_1, we have

$$x_2(1) = x_5(1) = c. \tag{2.20}$$

Moreover, $x_2 \equiv x_5$ on I_1 implies $\dot{x}_2 \equiv \dot{x}_5$ on I_1, i.e. $x_1 \equiv x_4$ on I_1. Since $x_4 \equiv 0$ on I_2, we get

$$\dot{x}_2(1) = x_1(1) = 0. \tag{2.21}$$

If $c < 0$, let $c = -\mu^{-2}$. Then $x_2(t)$ must be of the form

$$x_2(t) = \alpha\, e^{\mu(t-1)} + \beta\, e^{-\mu(t-1)} \tag{2.22}$$

on I_2, for some $\alpha, \beta \in \mathbb{R}$. The initial conditions $x_2(1) = c$, $\dot{x}_2(1) = 0$, give $\alpha + \beta = c$, $\mu(\alpha - \beta) = 0$. Since $\mu \neq 0$, we have $\alpha = \beta$ and so $\alpha = \beta = c/2$. Therefore

$$x_2(t) = \frac{c}{2}(e^{\mu(t-1)} + e^{-\mu(t-1)}) \tag{2.23}$$

for $1 \leqq t \leqq 2$. Therefore

$$x_1(t) = \dot{x}_2(t) = \frac{c}{2}[\mu\, e^{\mu(t-1)} - \mu\, e^{-\mu(t-1)}] \tag{2.24}$$

for $1 \leqq t \leqq 2$. In particular, we get

$$x_1(2) = \frac{c\mu}{2}(e^{\mu} - e^{-\mu}) \neq 0. \tag{2.25}$$

But this is a contradiction, since $x(2) = y$, and $y \in B$, so that $x_1(2) = 0$. Therefore the possibility that $c < 0$ is excluded.

Now suppose that $c > 0$. Write $c = \mu^{-2}$, $\mu > 0$. We then have

$$x_2(t) = \alpha \sin[\mu(t-1)] + \beta \cos[\mu(t-1)] \tag{2.26}$$

for $1 \leqq t \leqq 2$.

Since $\dot{x}_2(1) = 0$, we have $\alpha = 0$. So $\beta = c$, and

$$x_2(t) = c \cos[\mu(t-1)] \quad \text{for } 1 \leqq t \leqq 2. \tag{2.27}$$

Therefore

$$x_1(t) = \dot{x}_2(t) = -\mu c \sin[\mu(t-1)] \tag{2.28}$$

for $1 \leqq t \leqq 2$. Since $x_1(2) = 0$ (because $y \in B$), we find that $\sin \mu = 0$. Therefore $\mu = \pi k$ for some integer $k \neq 0$, and so $c = c_k$.

So we have shown that $c = c_k$ for some k. Since $x_5 \equiv c$ on I_2, we conclude that $x_5(2) = c_k$. Therefore $y \in H_k$. The proof of (2.2) is complete.

We now prove (2.3). We fix a number c such that

$$0 < c < 10^{-3}. \tag{2.29}$$

For this c, we construct an admissible control $u^c(\cdot)$ and a corresponding trajectory $x^c(\cdot)$ with initial condition $x^c(0) = 0$. This trajectory will have the property that

$$x^c(2) \in S \cap H_k \quad \text{if } c = c_k, \quad k > 10. \tag{2.30}$$

This will clearly imply that $B \cap H_k \neq \varnothing$ for $k > 10$, as desired.

From now on, c is fixed, so we will write $u(\cdot), x(\cdot)$ for $u^c(\cdot), x^c(\cdot)$. We first define $u(\cdot)$ on I_1.

For $t \in I_1$, we let

$$u_1 = u_2 = -12c(15t^2 - 14t + 2), \tag{2.31}$$

$$u_3 = u_4 = 1, \tag{2.32}$$

$$u_5 = x_2 + x_1^2 + u_1 x_2, \tag{2.33}$$

$$u_6 = 1 + (1-t)x_{10}, \tag{2.34}$$

$$u_7 = 1 + (2-t)x_{10}, \tag{2.35}$$

$$u_8 = t(2-5t)u_1 + (2-10t)x_1, \tag{2.36}$$

$$u_9 = 1 + (1-t)x_{16}, \tag{2.37}$$

$$u_{10} = 1 + (2-t)x_{16}. \tag{2.38}$$

Notice that formulas (2.31) and (2.32) define u_1, u_2, u_3, and u_4, and therefore make it possible, using (Ia) and (Ib), to compute the functions x_1 and x_2. Then u_5 is well defined by (2.33). Using u_5, (IIIb) determines x_{10}, and so u_7 is well defined by (2.35). Then u_8 is well defined by (2.36), since this equation only involves u_1 and x_1, and then (IVc) determines x_{16}. Therefore (2.37) and (2.38) can be used to determine u_9 and u_{10}.

The proof that the control defined by (2.31), $\cdots$, (2.38) is admissible is quite easy. Since $|15t^2 - 14t + 2| \leqq 3$ for $0 \leqq t \leqq 1$, we get $|u_1| \leqq 36c < \frac{1}{2}$, $|x_1| \leqq t/2$, $|x_2| \leqq t^2/4$, $|x_2 + x_1^2 + u_1 x_2| \leqq 3t^2/4 < 1$, $|u_5| \leqq 1$, $|x_{10}| \leqq t$. On the other hand, we can compute x_1, x_2, x_3 explicitly, and get

$$x_1 = -12ct(1-t)(2-5t), \tag{2.39}$$

$$x_2 = -ct^2(15t^2 - 28t + 12), \tag{2.40}$$

$$x_3 = -ct^3(1-t)(4-3t). \tag{2.41}$$

So

$$x_1 x_2 + x_3 = -ct^3(1-t)[4 - 3t - 12c(2-5t)(15t^2 - 28t + 12)]. \tag{2.42}$$

Since $4 - 3t \geqq 1$ for $t \in [0, 1]$, it is clear that the expression in the square brackets is positive for all $t \in [0, 1]$, if c is sufficiently small. An elementary calculation shows that the expression is positive for $0 < c < \frac{1}{72}$, so in particular it is positive for c in the range defined by (2.29). So $x_1 x_2 + x_3 < 0$ for $0 < t < 1$. By construction

$$\dot{x}_{10} = u_5 = x_2 + x_1^2 + u_1 x_2 = \widehat{x_3 + x_1 x_2}.$$

Therefore $x_{10} \equiv x_3 + x_1 x_2$ on I_1, and so $x_{10} < 0$ for $0 < t < 1$. Since $|x_{10}| \leqq t$, we have $1 \geqq u_6 \geqq 0$ and $1 \geqq u_7 \geqq -1$ on I_1. This shows in particular, that u_6 and u_7 are admissible.

Also, we get

$$u_8 = -12ct(2-5t)(25t^2 - 26t + 4) \tag{2.43}$$

which implies $|u_8| \leqq 144c < 1$. Therefore $|x_{16}| \leqq 1$. Since $u_8 = x_{14} u_1 + \dot{x}_{14} x_1$, we get $u_8 = \widehat{x_1 x_{14}}$. But $\dot{x}_{16} = u_8$ and so $x_{16} = x_1 x_{14}$. Since $x_1 x_{14} = -12ct^2(1-t)(2-5t)^2 \leqq 0$, we get $x_{16} \leqq 0$ on I_1. The admissibility of u_9 and u_{10} now follows exactly as that of u_6 and u_7 followed from the bounds for x_{10}.

We then define controls $u_1, \ldots, u_{10}$ on I_2, by letting

$$u_1 = -\cos\left(\frac{t-1}{\sqrt{c}}\right), \tag{2.44}$$

$$u_2 = u_5 = u_8 = 0, \tag{2.45}$$

$$u_3 = 1 - tx_6, \qquad u_4 = 1 + (1-t)x_6, \tag{2.46}$$

$$u_6 = u_7 = u_9 = u_{10} = 1. \tag{2.47}$$

The admissibility is immediate, with the possible exception of u_3 and u_4. But $\dot{x}_6 = (x_2 - x_5)^2$, and $x_2 \equiv x_5$ on I_1, so that $x_6(1) = 0$. Moreover $\dot{x}_2 = x_1$ and $\dot{x}_5 = x_4$, so $x_1(1) = x_4(1)$. And $\dot{x}_1 - \dot{x}_4 = u_1 - u_2$ so that, for $t \in I_2$, we have $|x_1(t) - x_4(t)| \leqq t - 1$. Therefore $|x_2(t) - x_5(t)| \leqq (t-1)^2/2$ for $t \in I_2$, and so $|x_6(t)| \leqq 1$ for $t \in I_2$. Also, $x_6(t) \geqq 0$, so $|u_3| \leqq 1$ and $|u_4| \leqq 1$, proving that the controls are admissible.

An elementary calculation (which we leave to the reader) now shows that

$$x_1(t) = -\sqrt{c}\sin\left(\frac{t-1}{\sqrt{c}}\right) \quad \text{for } t \in I_2,$$

so that, in particular,

$$x_1(2) = -\sqrt{c}\sin\left(\frac{1}{\sqrt{c}}\right).$$

Moreover, one sees easily that $x_{20}(t) = 0$ for $t \in I$, and in particular $x_{20}(2) = 0$. If $c = c_k$, it follows that $x_1(2) = x_{20}(2) = 0$, and so $x(2) \in S$. Finally, it is easily verified that $x_5(t) = c$ on I_2. So, if $c = c_k$, the point $x(2)$ is in H_k. This shows that, if $c = c_k$ and (2.29) holds, the set $B \cap H_k$ is nonempty. If $k > 10$, then $c_k < 10^{-3}$, and so the preceding result applies. The proof of (2.3) is now complete.

3. A nonsubanalytic optimal time function. We now modify the example of § 2 and construct a completely controllable system for which the optimal time function is not subanalytic.

The state space for our new system is $\mathbb{R}^{22}$. The variables are $x_0, \cdots, x_{20}$ and y. We write $\mathbf{x}$ for $(x_1, \cdots, x_{20})$. The controls are $u_1, \cdots, u_{10}$, and a new control v. For $i = 1, \cdots, 20$, the system of § 2 gives an expression for $\dot{x}_i$ which is of the form

$$\dot{x}_i = \phi_i(x_0, \mathbf{x}) + \psi_i(\mathbf{u}), \tag{3.1}$$

where $\mathbf{u}$ stands for $(u_1, \cdots, u_{10})$, the function ϕ_i is a polynomial of degree $\leqq 2$, and ψ_i is linear homogeneous. (For instance, $\phi_7(x_0, \mathbf{x}) = -x_0x_6 + 1$, $\psi_7(\mathbf{u}) = -u_3$.)

The equations for the new system are

$$\dot{x}_0 = 1 - y^2, \tag{3.2a}$$

$$\dot{x}_1 = 1 - y + u_1, \tag{3.2b}$$

$$\dot{x}_i = (1-y)\phi_i(x_0, \mathbf{x}) + yx_{i-1} + \psi_i(\mathbf{u}) \quad \text{for } i = 2, \cdots, 20, \tag{3.2c}$$

$$\dot{y} = v, \tag{3.2d}$$

and the control constraints are $|u_i| \leqq 1$, $|v| \leqq 1$.

We show that the system (3.2) is completely controllable. Let $p^i = (x_0^i, \mathbf{x}^i, y^i)$, $i = 1, 2$, be points of $\mathbb{R}^{22}$. We will show that p^1 can be steered to p^2. It is clear that p^1 can be steered to some point $(x_0, \mathbf{x}, y)$ for which $y = 1$. (Just use $v \equiv -1$ if $y > 1$, $v \equiv 1$ if

$y<1$, during a suitably chosen time.) Similarly, p^2 can be reached from some point for which $y=1$. This means that we can assume, without loss of generality, that $y^1=y^2=1$. Next we show that, for some $\mathbf{x}$, the point $q=(x_0^2, \mathbf{x}, 1)$ can be reached from p^1. Suppose first that $x_0^2>x_0^1$. For $T>0$, $T\leqq 2$, let $v_T:[0, T]\to\mathbb{R}$ be the control which is equal to -1 on $[0, T/2]$ and to 1 on $]T/2, T]$. If $T>2$, let v_T equal -1 on $[0, 1]$, 0 on $]1, T-1[$, and 1 on $[T-1, T]$. If we solve (3.2d) with $v=v_T$ and with initial condition $y(0)=1$, then the solution y_T is a function on $[0, T]$ that satisfies $y_T(T)=1$. Let

$$\xi_T=\int_0^T [1-y_T(t)^2]\,dt. \tag{3.3}$$

Then it is clear that ξ_T depends continuously on T, $\xi_0=0$, and $\lim_{T\to+\infty}\xi_T=+\infty$. So we can choose $T>0$ such that $\xi_T=x_0^2-x_0^1$. Choose $\mathbf{u}(\cdot)$ to be an arbitrary admissible control for the system of § 2, defined on $[0, T]$, and let $v(\cdot)=v_T$. Then $(u(\cdot), v(\cdot))$ steers p^1 to a point $q=(x_0^2, \mathbf{x}, 1)$. If $x_0^2<x_0^1$, then the proof that p can be steered to a q of the desired form is similar. (Use $v_T=1$ on $[0, T/2]$, $v_T=-1$ on $]T/2, T]$, for arbitrary $T\geqq 0$.)

So the complete controllability of the system (3.2) will follow if we prove that p^1 can be steered to p^2 whenever $p^i=(x_0, \mathbf{x}^i, 1)$, $i=1, 2$. To go from p^1 to p^2, we use $v\equiv 0$, so that $y\equiv 1$ and x_0 remains constant. Also, we set $u_2\equiv\cdots\equiv u_{10}\equiv 0$. The equations for $\dot{x}_i$, $i=1, \cdots, 20$, become

$$\dot{x}_1=u_1, \qquad |u_1|\leqq 1, \tag{3.4a}$$

$$\dot{x}_i=x_{i-1} \quad \text{for } i=2, \cdots, 20. \tag{3.4b}$$

Now, it is quite easy to see (cf. [5]) that the system (3.4) is completely controllable. Therefore, we can choose a control $u_1(\cdot)$ that will steer $\mathbf{x}^1$ to $\mathbf{x}^2$. This completes the proof of complete controllability.

The Bellman function $V:\mathbb{R}^{22}\to\mathbb{R}$ for the optimal time problem associated with our system (3.2) is defined by letting $V(p)$ be the infimum of the times T, taken over all trajectories of (3.2) that go from 0 to p in time T. It is easy to see that the nonexplosion condition (e) of § 1 holds for the system (3.2). This implies that, for each p, there exists an *optimal trajectory* from 0 to p, i.e. a trajectory that goes from 0 to p in time $V(p)$. (This follows from the fact that the sets

$$\tilde{A}_p(T)=\{(t, q): 0\leqq t\leqq T, q\in A_p(t)\}$$

are compact.)

We now prove that the level set

$$V^{-1}(2)=\{p: V(p)=2\} \tag{3.5}$$

is not subanalytic. To see this, we let H denote the hyperplane $\{(x_0, \mathbf{x}, y): x_0=2\}$, and we study the intersection

$$D=V^{-1}(2)\cap H. \tag{3.6}$$

We will show that D is not subanalytic. Let $A(2)$ be the time 2 reachable set from $0\in\mathbb{R}^{21}$ for the system of § 2. Then $A(2)\times\{0\}$ is a subset of H, because every $(x_0, \mathbf{x})\in A(2)$ satisfies $x_0=2$. Since $A(2)$ is not subanalytic in $\mathbb{R}^{21}$, the set $A(2)\times\{0\}$ is not subanalytic in $\mathbb{R}^{22}$ (by condition (i) of § 2).

To establish that D is not subanalytic, we prove that

$$D=A(2)\times\{0\}. \tag{3.7}$$

Let $p=(x_0^*, \mathbf{x}^*, y^*) \in D$. Then p is reachable from 0 in time 2 by a trajectory of (3.2), given by functions $x_0(t)$, $\mathbf{x}(t)$, $y(t)$, $0 \leqq t \leqq 2$. Moreover, $p \in H$ and so $x_0^* = x_0(2) = 2$. Since $\dot{x}_0(t) = 1 - y(t)^2$ on $[0, 2]$, the only way that $x_0(2)$ can be equal to 2 is if $y(t) = 0$ for $0 \leqq t \leqq 2$. So $y^* = 0$.

Moreover, if we plug in $y \equiv 0$ in the equations of (3.2), we find that $t \to (x_0(t), \mathbf{x}(t))$ is a trajectory of the system of § 2, which goes from 0 to $(x_0^*, \mathbf{x}^*)$ in time 2. Therefore $(x_0^*, \mathbf{x}^*) \in A(2)$. Since $y^* = 0$, we have shown that $p \in A(2) \times \{0\}$. Conversely, suppose that $p \in A(2) \times \{0\}$. Let $p = (x_0^*, \mathbf{x}^*, 0)$, with $(x_0^*, \mathbf{x}^*) \in A(2)$. If $t \to (x_0(t), \mathbf{x}(t))$ is a trajectory of the system of § 2 that goes from 0 to $(x_0^*, \mathbf{x}^*)$ in time 2, then $t \to (x_0(t), \mathbf{x}(t), 0)$ is a trajectory of (3.2) that goes from 0 to p in time 2. On the other hand, we have $x_0(t) = t$ and so $x_0^* = 2$.

If $t \to (\tilde{x}_0(t), \tilde{\mathbf{x}}(t), y(t))$, $0 \leqq t \leqq T$, is any trajectory of (3.2) that goes from 0 to p in time T, then (3.2a) implies that $\dot{\tilde{x}}(t) \leqq 1$. Since $\tilde{x}(0) = 0$, $\tilde{x}(T) = 2$, we see that $T \geqq 2$.

Hence 2 is the optimal time for steering 0 to p. So $V(p) = 2$. Since $x_0^* = 2$, we have $p \in H$. Therefore $p \in D$. This completes the proof of (3.7). As explained before, it follows that $V^{-1}(2)$ is not subanalytic.

Finally, if we let $G \subseteq \mathbb{R}^{22} \times \mathbb{R}$ denote the graph of V, the set $V^{-1}(2)$ is the inverse image of G under the map $z \to (z, 2)$ from $\mathbb{R}^{22}$ to $\mathbb{R}^{22} \times \mathbb{R}$. If G were subanalytic, it would follow from (i) of § 2 that $V^{-1}(2)$ is subanalytic. Therefore G is not subanalytic, i.e. V is not a subanalytic function.

4. A simple example for linear systems with nonpolyhedral control constraints. We consider the system

$$\dot{x} = u, \quad \dot{y} = v, \quad \dot{z} = -y \tag{4.1}$$

with control constraint

$$(u, v) \in K \tag{4.2}$$

where K is the set defined by

$$u^2 - 1 \leqq 2v \leqq 1 - u^2. \tag{4.3}$$

We pick an arbitrary $T > 0$, and we show that the attainable set $A_0(T)$ is not subanalytic.

Let X be a column vector with components x, y, z. Then we can write (4.1) as

$$\dot{X} = AX + w, \qquad w \in \Omega, \tag{4.4}$$

where A is the matrix

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \end{bmatrix}$$

and $\Omega = K \times \{0\}$. If $w(\cdot) : [0, T] \to \Omega$ is measurable, then the corresponding solution of (4.4), with initial condition $X(0) = 0$, is given by

$$X(t) = e^{tA} Y(t) \tag{4.5}$$

where

$$Y(t) = \int_0^t e^{-sA} w(s)\, ds. \tag{4.6}$$

In particular,

$$A_0(T) = e^{TA}\tilde{A}_0(T), \tag{4.7}$$

where $\tilde{A}_0(T)$ is the set of all $Y(T)$ given by (4.6), as $w(\cdot)$ varies over all measurable functions from $[0, T]$ to Ω.

So $A_0(T)$ is subanalytic iff $\tilde{A}_0(T)$ is subanalytic.

For $s \in \mathbb{R}$, let $\lambda(s): \mathbb{R}^3 \to \mathbb{R}$ be the linear functional given by

$$\langle \lambda(s), (x, y, z) \rangle = sx + sy + z.$$

Then $\lambda(\cdot)$ is a real analytic function from $\mathbb{R}$ to (the dual of) $\mathbb{R}^3$. For $s \in]0, 1[$, let $p(s)$ be the point of $\tilde{A}_0(T)$ on which $\lambda(s)$ has its maximum value. (The existence of a maximum follows from the compactness of $\tilde{A}_0(T)$. The uniqueness would follow from general considerations about strict convexity, but in our case, it will follow directly from the explicit calculation of $p(s)$ carried out below.)

We compute $p(s)$. Clearly,

$$p(s) = \int_0^T e^{-tA} w(t)\, dt \tag{4.8}$$

where, for each $t \in [0, T]$, $w(t)$ is the value of $w \in \Omega$ that maximizes

$$\langle \lambda(s), e^{-tA} w \rangle. \tag{4.9}$$

Now, it is easy to see that

$$e^{-tA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & t & 1 \end{bmatrix}. \tag{4.10}$$

If we write $w = (u, v, 0)$, then (4.9) becomes $su + (s+t)v$. The value of (u, v) that maximizes this quantity, subject to $(u, v) \in K$, is

$$u(t) = \frac{s}{s+t}, \tag{4.11}$$

$$v(t) = \tfrac{1}{2}(1 - u(t)^2). \tag{4.12}$$

If we write

$$p(s) = (p_1(s), p_2(s), p_3(s)),$$

we see that

$$p_1(s) = \int_0^T \frac{s}{s+t}\, dt, \tag{4.13}$$

so that

$$p_1(s) = s[\log(T+s) - \log s], \tag{4.14}$$

where "log" denotes natural logarithm.

We now use some well-known properties of subanalytic sets. First, if a real valued function ϕ of one variable, on some interval $]0, a[$, is bounded, analytic and subanalytic, then it is actually semianalytic (i.e. the graph of ϕ is a semianalytic subset of $\mathbb{R}^2$), and then ϕ has a Puiseaux series near $s = 0$. This implies that, as $s \to 0$, $\phi(s)$ is asymptotic to cs^α for some rational $\alpha \geqq 0$, and some constant $c \neq 0$. The function $p_1(\cdot)$ given by

(4.14) clearly does not have this type of asymptotic behavior. Therefore $p_1(\cdot)$ is not subanalytic. So $p(\cdot)$ is not subanalytic.

On the other hand, $p(s) = q$ iff $q \in \tilde{A}_0(T)$ and

$$\langle \lambda(s), q \rangle = \max \{\langle \lambda(s), r \rangle : r \in \tilde{A}_0(T)\}.$$

If $\tilde{A}_0(T)$ were subanalytic, then it would follow from general properties of subanalytic sets that $\{(s, q): p(s) = q\}$ is subanalytic in $\mathbb{R}^4$, i.e. that $p(\cdot)$ is a subanalytic function. Therefore $\tilde{A}_0(T)$ is not subanalytic, and then $A_0(T)$ is not subanalytic either.

Remark. The system (4.1) is completely controllable and, moreover, for arbitrarily small $\varepsilon > 0$ the reachable set $A_0(\varepsilon)$ contains 0 as an interior point. This implies that the optimal time function for (4.1) (with initial point 0) is continuous, and, for each $T > 0$, the level set $V^{-1}(T)$ is precisely the boundary of $A_0(T)$. Clearly, the fact that $A_0(T)$ is not subanalytic implies that its boundary is not subanalytic (using the fact that $A_0(T)$ is compact and convex). So $V^{-1}(T)$ is not subanalytic, and V is not a subanalytic function.

REFERENCES

[1] P. BRUNOVSKÝ, *Every normal linear system has a regular time-optimal synthesis*, Math. Slovaca, 28 (1978), pp. 81-100.

[2] ———, *On the structure of optimal feedback systems*, Proc. Int. Congress of Mathematicians, Helsinki, 1978, pp. 841-846.

[3] ———, *Existence of regular synthesis for general problems*, J. Differential Equations, 38 (1980), pp. 317-343.

[4] ———, *Regular synthesis for the linear-quadratic optimal control problem with linear control constraints*, J. Differential Equations, 38 (1980), pp. 344-360.

[5] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1976.

[6] C. MARCHAL, *Chattering arcs and chattering controls*, J. Optim. Theory Appl., 11 (1973), pp. 411-468.

[7] H. J. SUSSMANN, *Analytic stratifications and control theory*, Proc. Int. Congress of Mathematicians, Helsinki, 1978, pp. 865-871.

[8] ———, *A bang-bang theorem with bounds on the number of switchings*, this Journal, 17 (1979), pp. 371-393.

[9] ———, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31-52.

[10] ———, *Lie brackets, real analyticity and geometric control*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman and H. J. Sussmann, eds., Birkhäuser, Boston, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

OPTIMAL CONTROL OF PARTIALLY OBSERVABLE STOCHASTIC SYSTEMS WITH AN EXPONENTIAL-OF-INTEGRAL PERFORMANCE INDEX*

A. BENSOUSSAN† AND J. H. VAN SCHUPPEN‡

Abstract. The stochastic control problem with linear stochastic differential equations driven by Brownian motion processes and as cost functional the exponential of a quadratic form is considered. The solution consists of a linear control law and of a linear stochastic differential equation. The latter has the same structure as the Kalman filter but depends explicitly on the cost functional. The separation property does not hold in general for the solution to this problem.

Key words. stochastic control, linear systems, exponential-of-integral cost functional, linear-exponential-Gaussian problem

1. Introduction. The class of so-called Linear-Exponential-Gaussian (LEG) stochastic control problems has been introduced by Jacobson [4] and Speyer et al. [7] Since then several papers have presented solutions to special cases of this problem. The general case of the discrete-time problem has been solved by Whittle [11]. Below the solution to the general case of a continuous-time partially observed stochastic system is presented.

The simplest special case of the LEG stochastic control problem is that of the completely observable system

$$dx = (Fx + Bv)\,dt + G\,dw, \qquad x_0 = \mu_0, \tag{1.1}$$

with the cost functional

$$J(v(\cdot)) = E\left[\mu \exp\left(\frac{\mu}{2}\left[Mx_{t_1}^2 + \int_0^{t_1} (Qx^2 + Nv^2)\,dt\right]\right)\right]. \tag{1.2}$$

Under certain definite conditions there exists a linear optimal control which is implementable by a finite dimensional system, see [4].

Subsequently Speyer et al. [7], [8] have considered the case of a partially observed system

$$\begin{aligned} dx &= (Fx + Bv)\,dt + G\,dw, \qquad x_0 = \mu_0, \\ dy &= Hx\,dt + R^{1/2}\,db, \qquad y_0 = 0, \end{aligned} \tag{1.3}$$

with the cost functional (1.2) with $Q = 0$. Again there exists a finite dimensional implementable optimal control.

Yet another case of a partially observed stochastic control problem is considered by Kumar, van Schuppen [6]. There the general cost functional (1.2) is combined with the stochastic system (1.3), but with $G = 0$. It is proven that the optimal control is given by

$$u_t = -N^{-1}(t)B^*(t)[L(t)\hat{x}_t + M(t)\eta(t)] \tag{1.4}$$

where $\hat{x}_t$ is produced by the Kalman filter and

$$\eta(t) = \int_0^t K(t, s)u_s\,ds.$$

* Received by the editors December 29, 1983, and in revised form June 25, 1984.
† I.N.R.I.A., Domaine de Voluceau, Rocquencourt, 78150 Le Chesnay, France.
‡ Centre for Mathematics and Computer Science, P.O. Box 4079, 1009 AB Amsterdam, The Netherlands.

The control is thus implementable by a finite dimensional system. It has long been thought that the general case does not have a finite dimensional implementable optimal controller.

In this paper the solution to the general stochastic control problem will be presented, consisting of the stochastic system (1.3) and the cost functional (1.2). It will be proven that the optimal control is given by

$$u_t = -N^{-1}(t)B^*(t)S(t)r_t \tag{1.5}$$

where

$$dr = [F - PH^*R^{-1}H + \mu PQ]r_t\,dt + Bu\,dt + PH^*R^{-1}\,dy, \qquad r_0 = \mu_0, \tag{1.6}$$

P is the solution of a filter type Riccati differential equation

$$\dot{P} - FP - PF^* + P(H^*R^{-1}H - \mu Q)P - GG^* = 0, \qquad P(0) = P_0 \tag{1.7}$$

and S is the solution of a control type Riccati differential equation, see (4.1), that however depends explicitly on the function P. The recursions for the sufficient statistics reduce to the Kalman filter if $Q = 0$. The presentation of the solution (1.5) and (1.6) is much more convenient than that given by (1.4).

The motivation for considering LEG stochastic control problems is that for certain applications the exponential-of-integral cost functional may be better suitable than the usual quadratic cost functional. The reason for this is that the exponential form introduces a nonlinear relation between small and large deviations from the equilibrium state. The economic interpretation of the solution of the LEG problem is discussed by van der Ploeg [10]. One may interpret the solution as an attitude of either risk-preference, for $\mu < 0$, or of risk-aversion, for $\mu > 0$, see [10], [11], [12].

There is also a system theoretic motivation to consider LEG stochastic control problems. A major question in stochastic control theory is to classify those stochastic control systems and cost functionals that lead to finite dimensional control algorithms. An attempt to define a finite dimensional control algorithm will not be given here, but the solution presented by (1.5) and (1.6) illustrates what the authors have in mind. It is expected that the availability of the solution to the LEG problem may provide insight into the above stated question. Apparently the fact that the conditional cost function has for all $t \in T$ the same analytic form, plays a key role. In addition the solution provides an example of a sufficient statistic for a stochastic control problem which does not have the separation property.

A brief summary of the paper follows. In the following section a problem formulation is given. In § 3 an equivalent expression for the cost functional is derived. The solution is presented in § 4.

2. The problem formulation.

Notation. Let $(\Omega, \mathcal{A}, P)$ be a complete probability space and $T = [0, t_1]$, on which are defined

$$\begin{aligned} &w: \Omega \times T \to R^k && \text{a standard Wiener process;} \\ &\tilde{b}: \Omega \times T \to R^d && \text{a standard Wiener process;} \\ &R: T \to R^{d \times d} && \text{a symmetric positive definite matrix for which there exists a } r_0 \in (0, \infty) \text{ such that for all } t \in T\ R(t) \geqq r_0 I; \\ &y: \Omega \times T \to R^d \end{aligned} \tag{2.1}$$

$$dy_t = R^{1/2}(t)\,d\tilde{b}_t, \qquad y_0 = 0; \tag{2.2}$$

$$x_0: \Omega \to R^n \qquad \text{a Gaussian random variable with mean } \mu_0 \text{ and variance } P_0, \text{ with } P_0 \text{ nonsingular.} \tag{2.3}$$

Assume that x_0, y, w are independent objects. The process y will be termed the *observation process.*

Consider processes $v: \Omega \times T \to R^m$ which belong to $L_y(0, t_1; R^m)$, meaning that they are adapted to the σ-algebra family $Y^t = \sigma(\{y_s, \forall s \leqq t\})$ generated by the observation process. For uninitiated readers of stochastic control the well-known fact is pointed out that v in $L_y(0, t_1; R^m)$ is equivalent to specifying a nonanticipating function $f: (R^d)^T \to R^m$ such that $v_t = f(y(\cdot))$. Here nonanticipating means that for any $t \in T$ $v_t = f(y(\cdot))$ depends only on the path of y before time t. One may consider f to be a control law.

For $v \in L_y(0, t_1; R^m)$ define the *state process* as the solution of the stochastic differential equation

$$dx_t = [F(t)x_t + B(t)v_t]\, dt + G(t)\, dw_t, \qquad x_0, \tag{2.4}$$

where

$$F: T \to R^{n \times n}, \quad B: T \to R^{n \times m}, \quad G: T \to R^{n \times k}.$$

Define the process $b: \Omega \times T \to R^d$

$$b_t = \tilde{b}_t - \int_0^t R^{-1/2}(s) H(s) x_s\, ds \tag{2.5}$$

for $H: T \to R^{d \times n}$. Define the change of probability

$$\begin{aligned} \frac{d\tilde{P}}{dP} &= \exp\left(\int_0^{t_1} R^{-1/2} Hx \cdot d\tilde{b} - \frac{1}{2} \int_0^{t_1} H^* R^{-1} Hx \cdot x\, ds \right) \\ &= \exp\left(\int_0^{t_1} R^{-1} Hx \cdot dy - \frac{1}{2} \int_0^{t_1} H^* R^{-1} Hx \cdot x\, ds \right). \end{aligned} \tag{2.6}$$

Since the integrand entering into the stochastic integral at the right-hand side of (2.6) is unbounded, an assumption is necessary to ensure that $\tilde{P}$ is indeed a probability measure. The following criterion will be used, see [1], [3, p. 83]: there exist $\mu, c \in (0, \infty)$ such that

$$E[\exp(\mu H^* R^{-1} Hx_t \cdot x_t)] \leqq c \tag{2.7}$$

for all $t \in T$. Define $x_1: \Omega \times T \to R^n$, $x_2: \Omega \times T \to R^n$

$$\dot{x}_{1t} = Fx_{1t} + Bv_t, \qquad x_{10} = 0, \tag{2.8}$$

$$dx_{2t} = Fx_{2t}\, dt + G\, dw_t, \qquad x_{20} = x_0. \tag{2.9}$$

Then

$$x_t = x_{1t} + x_{2t}, \tag{2.10}$$

and x_1, x_2 are independent with respect to the probability measure P. Therefore (2.7) can be majorized as follows

$$\begin{aligned} E[\exp(\mu H^* R^{-1} H(x_1 + x_2)^2)] &\leqq E[\exp(2\mu H^* R^{-1} H(x_1^2 + x_2^2))] \\ &\leqq E[\exp(2\mu H^* R^{-1} Hx_{1t}^2)] E[\exp(2\mu H^* R^{-1} Hx_{2t}^2)]. \end{aligned}$$

Since x_{2t} is a Gaussian random variable, it is possible to find a $\mu \in (0, \infty)$ such that

the second expectation is finite. Because

$$\|x_{1t}\|^2 \leqq c \int_0^{t_1} \|v_s\|^2 \, ds,$$

(2.7) will be satisfied if there exists a $\nu \in (0, \infty)$ such that

$$E\left[\exp\left(\nu \int_0^{t_1} \|v_s\|^2 \, ds\right)\right] < \infty. \tag{2.11}$$

The parameter ν may depend on the control, but not on Ω. The preliminary set of admissible controls is therefore defined as

$$\mathbf{U}_1 = \{v \in L_y(0, t_1; R^m) | \exists \nu \in (0, \infty) \text{ such that (2.11) holds}\}.$$

Other restrictions will be stated later.

With respect to the probability measure $\tilde{P}$ the process b is a standard Wiener process, x_0, b, w are independent objects, and

$$dy_t = H(t)x_t \, dt + R^{1/2}(t) \, db_t, \qquad y_0 = 0. \tag{2.12}$$

Furthermore the measures P and $\tilde{P}$ are identical when restricted to the σ-algebra's generated by x_0 and w.

In the rest of the paper the time parameter will often be suppressed.

The stochastic control problem. Because of the dependence of $\tilde{P}$ on the control v the notation $\tilde{P}^v$ will be used. Consider the cost function

$$J(v)(\cdot)) = \tilde{E}^v\left[\mu \exp\left(\frac{\mu}{2}\left[Mx_{t_1}^2 + \int_0^{t_1} (Qx^2 + Nv^2) \, ds\right]\right)\right] \tag{2.13}$$

where $\mu \in R$, $\mu \neq 0$, is given,

$$\begin{aligned} &M \in R^{n \times n} \text{ is symmetric and nonnegative definite,} \\ &Q: T \to R^{n \times n} \text{ is also symmetric and nonnegative definite, and} \\ &N: T \to R^{m \times m} \text{ is symmetric positive definite for which there} \\ &\qquad \text{exists a } n_0 \in (0, \infty) \text{ such that for all } t \in T \; N(t) \geqq n_0 I. \end{aligned} \tag{2.14}$$

Note that for both $\mu > 0$ and $\mu < 0$, $J(v(\cdot))$ should be minimized.

In order that $J(v(\cdot))$ is finite for at least some $v \in \mathbf{U}_1$, an assumption is necessary. Let

$$\mathbf{U}_2 = \{u \in \mathbf{U}_1 | J(u(\cdot)) < \infty\} \tag{2.15}$$

be called the *class of admissible controls.* It will be assumed that $\mathbf{U}_2$ is nonempty. A condition guaranteeing that $\mathbf{U}_2$ is nonempty is the following. Let $v \equiv 0$. Then $x = x_2$, and in this case the probability laws of x_2 with respect to P and $\tilde{P}^v$ are the same. Therefore

$$J(0(\cdot)) = E\left[\mu \exp\left(\frac{\mu}{2}\left[Mx_{t_1}^2 + \int_0^{t_1} Qx_{2s}^2 \, ds\right]\right)\right]. \tag{2.16}$$

If $J(0(\cdot)) < \infty$, then $\mathbf{U}_2$ is nonempty, and it is likely that it will contain more than one

element. One can also reformulate (2.13) as

$$J(v(\cdot)) = E\left[\mu \exp\left(\frac{\mu}{2}\left[Mx_{t_1}^2 + \int_0^{t_1} Nv^2\, ds\right]\right)\right.$$

(2.17)

$$\left. + \int_0^{t_1} \left(\frac{\mu Q}{2} - \frac{H^* R^{-1} H}{2}\right) x^2\, ds + \int_0^{t_1} R^{-1} H x_s \cdot dy_s\right].$$

DEFINITION 2.1. a) An admissible control u^* will be called *optimal* for the cost functional (2.13) if

$$J(u^*(\cdot)) \leqq J(v(\cdot)) \tag{2.18}$$

for all $v \in \mathbf{U}_2$. Here the state process and the observation process are given respectively by (2.4) and (2.12).

b) An admissible control u^* will be called *conditionally optimal* for the cost functional (2.13) if for all $t \in T$

$$\tilde{E}^{u^*}[c^{u^*} | Y^t] \leqq \tilde{E}^{v}[c^{v} | Y^t] \quad \text{a.s. } \tilde{P}^{u^*} \tag{2.19}$$

for all $v \in \mathbf{U}_2$ such that for all $s \leqq t$, $u_s^* = v_s$. Here

$$c^v = \mu \exp\left(\frac{\mu}{2}\left[Mx_{t_1}^2 + \int_0^{t_1} (Qx^2 + Nv^2)\, ds\right]\right). \tag{2.20}$$

Note that because of the condition that for $s \leqq t$, $u_s^* = v_s$ and causality, $\tilde{P}^{u^*}$ and $\tilde{P}^v$ agree on Y^t. Hence (2.19) holds almost surely with respect to $\tilde{P}^{u^*}$.

The definition of conditional optimality is due to C. Striebel [9, Ch. 4]. If $u^* \in \mathbf{U}_2$ is conditionally optimal, then it is also optimal; take $t = 0$ in (2.19) and use $y_0 = 0$. However the converse is not true, see [6, p. 315] for a counterexample.

Problem 2.2. The *Linear-Exponential-Gaussian stochastic control problem* is to determine an admissible control u_1^* that is optimal, and an admissible control u_2^*, possibly different from u_1^*, that is conditionally optimal. The state process and the observation process are given respectively by (2.4) and (2.12).

3. Calculation of the cost function. The solution to the stochastic control problem 2.2 that will be given in § 4 is based on an alternate expression for the cost functional. This result will be derived below.

Definitions. The following variables are introduced:

$P: T \to R^{n \times n}$

$$\dot{P} - FP - PF^* + P(H^* R^{-1} H - \mu Q)P - GG^* = 0, \qquad P(0) = P_0; \tag{3.1}$$

$r: \Omega \times T \to R^n$

$$dr = [F - PH^* R^{-1} H + \mu PQ]r\, dt + Bv\, dt + PH^* R^{-1}\, dy, \qquad r_0 = \mu_0; \tag{3.2}$$

for any $v \in \mathbf{U}_2$ $\pi^v: \Omega \times R^n \times T \to R$

$$\pi^v(x, t) = \exp\left(-\frac{1}{2} P(t)^{-1}(x - r) \cdot (x - r)\right.$$

$$+ \int_0^t R^{-1} Hr \cdot dy - \frac{1}{2}\int_0^t H^* R^{-1} Hr \cdot r\, ds \tag{3.3}$$

$$\left. + \frac{\mu}{2}\int_0^t (Qr^2 + Nv^2)\, ds + \frac{\mu}{2}\int^t \operatorname{tr}(PQ)\, ds\right)(2\pi)^{-n/2} |P(t)|^{-1/2};$$

$$K(v(\cdot)) = E\left[\mu \int \exp\left(\frac{\mu}{2} Mx^2\right)\pi^v(x, t_1)\, dx\right]; \tag{3.4}$$

$\Sigma: T \to R^{n\times n}$

$$\dot{\Sigma} - \Sigma GG^*\Sigma + F\Sigma + \Sigma F^* - \mu Q + H^*R^{-1} = 0, \qquad \Sigma(t_1) = -\mu M. \tag{3.5}$$

Assumptions 3.1.

$$H^*R^{-1}H - \mu Q \geqq 0 \quad \text{(then a solution to (3.1) exists)}; \tag{3.6}$$

$$P(t) \geqq c_1 I \quad \text{for some } c_1 \in (0, \infty) \text{ and for all } t \in T; \tag{3.7}$$

$$P^{-1}(t) + \Sigma(t) > 0 \quad \text{for all } t \in T. \tag{3.8}$$

THEOREM 3.2. *Assume that* (3.6), (3.7) *and* (3.8) *hold. Assume further that the Riccati equation* (3.5) *has a symmetric bounded solution. For any control v in the class of admissible controls* $\mathbf{U}_2$ *one has the equality*

$$J(v(\cdot)) = K(v(\cdot)) \tag{3.9}$$

where $J(v(\cdot))$ *is defined by* (2.13) *and* $K(v(\cdot))$ *by* (3.4).

The proof of Theorem 3.2 is based on several lemmas.

Preliminary calculations. It will be convenient to introduce the processes

$$\begin{aligned} z_t &= \exp\left(\int_0^t R^{-1}Hx \cdot dy - \frac{1}{2}\int_0^t H^*R^{-1}Hx \cdot x\, ds\right), \\ \lambda_t &= \exp\left(\frac{\mu}{2}\int_0^t (Qx^2 + Nv^2)\, ds\right). \end{aligned} \tag{3.10}$$

Thus

$$J(v(\cdot)) = E\left[\mu \exp\left(\frac{\mu}{2} Mx_{t_1}^2\right)\lambda_{t_1} z_{t_1}\right]. \tag{3.11}$$

Define

$$\begin{aligned} s_t = {} & P^{-1}(t)r_t^2 - 2\int_0^t R^{-1}Hr \cdot dy + \int_0^t H^*R^{-1}Hr \cdot r\, ds \\ & - \mu\int_0^t (Qr^2 + Nv^2)\, ds - \mu\int_0^t \operatorname{tr}(PQ)\, ds + \ln((2\pi)^n |P(t)|). \end{aligned} \tag{3.12}$$

Then

$$\pi^v(x, t) = \exp\left(-\tfrac{1}{2}[P^{-1}(t)x \cdot x - 2P^{-1}(t)r_t \cdot x + s_t]\right). \tag{3.13}$$

The following result is then obtained.

LEMMA 3.3. *The process s is a solution of the differential equation*

$$\frac{ds_t}{dt} = \operatorname{tr}(G^*P^{-1}G + 2F) - |G^*P^{-1}r|^2 + 2P^{-1}Bv \cdot r - \mu Nv^2. \tag{3.14}$$

Proof. One uses

$$dP^{-1}r^2 = (P^{-1})'r \cdot r\, dt + 2P^{-1}r \cdot dr + \operatorname{tr}(PH^*R^{-1}H)\, dt,$$

$$\frac{dP^{-1}(t)}{dt} = -P^{-1}\dot{P}P^{-1} = -P^{-1}F - F^*P^{-1} + H^*R^{-1}H - \mu Q - P^{-1}GG^*P^{-1};$$

hence,

$$
\begin{aligned}
dP^{-1}r^2 = &[-2Fr\cdot P^{-1}r + R^{-1}Hr\cdot Hr - \mu Qr^2 - |G^*P^{-1}r|^2]\,dt \\
&+2P^{-1}r\cdot[F - PH^*R^{-1}H + \mu PQ]r\,dt \\
&+2P^{-1}r\cdot Bv\,dt + 2r\cdot H^*R^{-1}\,dy + \operatorname{tr}(PH^*R^{-1}H)\,dt \\
= &[-R^{-1}Hr\cdot Hr + \mu Qr^2 - |G^*P^{-1}r|^2 + 2P^{-1}r\cdot Bv + \operatorname{tr}(PH^*R^{-1}H)]\,dt \\
&+2r\cdot H^*R^{-1}\,dy.
\end{aligned} \tag{3.15}
$$

Moreover writing

$$
\dot{P} = [F + PF^*P^{-1} - PH^*R^{-1}H + GG^*P^{-1} + \mu PQ]P,
$$

one deduces

$$
\begin{aligned}
d\ln|P(t)|/dt &= \operatorname{tr}(F + PF^*P^{-1} - PH^*R^{-1}H + GG^*P^{-1} + \mu PQ) \\
&= \operatorname{tr}(2F - PH^*R^{-1}H + GG^*P^{-1} + \mu PQ).
\end{aligned} \tag{3.16}
$$

From (3.12), (3.15), and (3.16), one easily deduces (3.14). □

Assumption (3.8) for $t = t_1$ implies that

$$
P^{-1}(t_1) - \mu M > 0
$$

is positive definite. Hence one can calculate the integral

$$
\begin{aligned}
&\int \exp\left(-\frac{1}{2}P^{-1}(t_1)(x - r_{t_1})^2 + \frac{\mu}{2}Mx^2\right) dx \\
&\quad = \exp\left(\frac{\mu}{2}r_{t_1}\cdot[I - \mu MP(t_1)]^{-1}Mr_{t_1}\right) \\
&\qquad \cdot (2\pi)^{n/2}|P(t_1)|^{1/2}|[I - \mu MP(t_1)]|^{-1/2}.
\end{aligned}
$$

Therefore one can write

$$
\begin{aligned}
K(v(\cdot)) &= E\left[\mu\int \exp((\mu/2)Mx^2)\pi(x, t_1)\,dx\right] \\
&= E\left[\mu.\exp\left[\int_0^{t_1} R^{-1}Hr\cdot dy - \frac{1}{2}\int_0^{t_1} H^*R^{-1}Hr\cdot r\,dt\right.\right. \\
&\qquad \left.\left. + \frac{\mu}{2}\int_0^{t_1}(Qr^2 + Nv^2)\,dt + \frac{\mu}{2}[I - \mu MP(t_1)]^{-1}Mr_{t_1}\cdot r_{t_1}\right]\right] \\
&\qquad \cdot \exp\left(\frac{\mu}{2}\int_0^{t_1}\operatorname{tr}(PQ)\,dt\right)|[I - \mu MP(t_1)]|^{-1/2}
\end{aligned} \tag{3.17}
$$

assuming that the expectation is finite.

Equation for π^v and its adjoint. To show that $K(v(\cdot))$ is an alternative expression for the cost, an equation for π^v and its adjoint are needed.

LEMMA 3.4. *The process $\pi^v(x, t)$ has the Ito differential*

$$
\begin{aligned}
d\pi = &\left[\frac{1}{2}\operatorname{tr}(GG^* D_x^2 \pi) - (Fx + Bv)\cdot D_x\pi \right.\\
&\left. + \frac{\mu}{2}\pi(Qx^2 + Nv^2) - \pi \operatorname{tr}(F)\right] dt + \pi R^{-1}Hx \cdot dy\\
= \pi&\left[\left(P^{-1}Fx\cdot x + \frac{\mu}{2}Qx\cdot x + \frac{1}{2}|G^*P^{-1}x|^2\right.\right.\\
&+ x\cdot(P^{-1}Bv - F^*P^{-1}r - P^{-1}GG^*P^{-1}r)\\
&-\frac{1}{2}\operatorname{tr}(G^*P^{-1}G + 2F) + \frac{1}{2}|G^*P^{-1}r|^2\\
&\left.\left. + \frac{\mu}{2}Nv^2 - P^{-1}Bv\cdot r\right) dt + R^{-1}Hx\cdot dy\right].
\end{aligned}
\tag{3.18}
$$

Proof. This follows by simple calculations from (3.13). □

At this stage it is convenient to use the pathwise form of (3.18) [2]. In order to derive it, it is however necessary to assume that

$$
R^{-1}(t)H(t) \text{ is differentiable.} \tag{3.19}
$$

Let us consider $q^v : R^n \times \Omega \times T \to R$

$$
q^v(x, t) = \pi^v(x, t)\exp(-y_t \cdot R^{-1}(t)H(t)x).
$$

LEMMA 3.5. *The process $q^v(\cdot, \cdot)$ satisfies the equation*

$$
\begin{aligned}
\frac{\partial q(x, t)}{\partial t} = &\frac{1}{2}\operatorname{tr}(GG^* D_x^2 q) + D_x q\cdot(GG^*H^*R^{-1}y - Fx - Bv)\\
&+\frac{1}{2}q[|G^*H^*R^{-1}y|^2 + \mu(Qx^2 + Nv^2) - H^*R^{-1}Hx^2\\
&-2(Fx + Bv)\cdot H^*R^{-1}y - 2(R^{-1}H)'x\cdot y - 2\operatorname{tr}(F)].
\end{aligned}
\tag{3.20}
$$

Proof. One has

$$
\begin{aligned}
dq(x, t) = &\, d\pi \exp(-y_t\cdot R^{-1}Hx)\\
&+ \pi\exp(-y_t\cdot R^{-1}Hx)[-dy\cdot R^{-1}Hx - y_t\cdot(R^{-1}H)'x\,dt]\\
&-\tfrac{1}{2}\pi\exp(-y_t\cdot R^{-1}Hx)H^*R^{-1}Hx\cdot x\,dt
\end{aligned}
$$

from which one easily derives (3.20). □

Next one derives the adjoint equation of (3.20) with respect to (3.4), which reads

$$
\begin{aligned}
-\frac{\partial p(x, t)}{\partial t} = &\frac{1}{2}\operatorname{tr}(GG^* D_x^2 p) - D_x p\cdot(GG^*H^*R^{-1}y - Fx - Bv)\\
&+\frac{1}{2}p[|G^*H^*R^{-1}y|^2 + \mu(Qx^2 + Nv^2) - H^*R^{-1}Hx^2\\
&-2(Fx + Bv)\cdot H^*R^{-1}y - 2(R^{-1}H)'x\cdot y],\\
p(x, t_1) = &\,\mu\exp\left(\frac{\mu}{2}Mx^2 + y_{t_1}\cdot R^{-1}(t_1)H(t_1)x\right).
\end{aligned}
\tag{3.21}
$$

In fact it is possible to solve (3.21) exactly.

LEMMA 3.6. *Assume that the Riccati equation* (3.5) *has a symmetric bounded solution. Define*

$\sigma: \Omega \times T \to R^n$

$$\begin{aligned} &\dot{\sigma} + (F^* - \Sigma GG^*)(\sigma - H^* R^{-1} y) - \Sigma Bv - (H^* R^{-1})' y = 0, \\ &\sigma(t_1) = H^*(t_1) R^{-1}(t_1) y_{t_1}; \end{aligned} \tag{3.22}$$

$\rho: \Omega \times T \to R$

$$\begin{aligned} \dot{\rho} = &-\operatorname{tr}(GG^*\Sigma) + |G^*\sigma|^2 - 2\sigma \cdot (GG^* H^* R^{-1} y - Bv) \\ &+ [|G^* H^* R^{-1} y|^2 + \mu N v^2 - 2Bv \cdot H^* R^{-1} y], \qquad \rho(t_1) = 0. \end{aligned} \tag{3.23}$$

Then

$$p(x, t) = \mu \exp\left(-\tfrac{1}{2}[\Sigma(t)x \cdot x - 2\sigma(t) \cdot x + \rho(t)]\right)$$

is a solution of (3.21).

Proof. One has

$$\frac{\partial p(x, t)}{\partial t} = p(x, t)\left[-\frac{1}{2}\dot{\Sigma}x \cdot x + \dot{\sigma} \cdot x - \frac{1}{2}\dot{\rho}\right],$$

$$D_x p = p[-\Sigma x + \sigma],$$

$$D_x^2 p = p[-\Sigma x + \sigma] \otimes [-\Sigma x + \sigma] - p\Sigma.$$

Substitution in (3.21) yields

$$\begin{aligned} \tfrac{1}{2}\dot{\Sigma}x \cdot x - \dot{\sigma} \cdot x + \tfrac{1}{2}\dot{\rho} = &-\tfrac{1}{2}\operatorname{tr}(GG^*\Sigma) + \tfrac{1}{2}|G^*(-\Sigma x + \sigma)|^2 \\ &- (-\Sigma x + \sigma) \cdot (GG^* H^* R^{-1} y - Fx - Bv) \\ &+ \tfrac{1}{2}[|G^* H^* R^{-1} y|^2 + \mu(Qx^2 + Nv^2) - H^* R^{-1} Hx^2 \\ &\qquad - 2(Fx + Bv) \cdot H^* R^{-1} y - 2(R^{-1}H)' x \cdot y] \end{aligned}$$

and the result follows with (3.5), (3.22), and (3.23). □

LEMMA 3.7. *The functional* $K(v(\cdot))$ *defined by* (3.4) *can be calculated by*

$$K(v(\cdot)) = E\left[\int p(x, 0)\pi(x, 0)\, dx\right]. \tag{3.24}$$

Proof. By the expression for $p(x, t_1)$ in (3.21) one has

$$\mu \int \exp\left(\frac{\mu}{2} M x^2\right) \pi(x, t_1)\, dx = \int p(x, t_1) q(x, t_1)\, dx.$$

This integral makes sense by assumption (3.8). An integration by parts yields

$$\mu \int \exp\left(\frac{\mu}{2} M x^2\right) \pi(x, t_1)\, dx = \int p(x, 0) q(x, 0)\, dx = \int p(x, 0)\pi(x, 0)\, dx.$$

Taking the expectation, one deduces (3.24). □

Equality of the two costs.

3.8. *Proof of Theorem* 3.2. From (3.11) and (3.21) follows that

$$J(v(\cdot)) = E[p(x_{t_1}, t_1)\lambda_{t_1} z_{t_1} \exp(-y_{t_1} \cdot R^{-1}(t_1) H(t_1) x_{t_1})].$$

But

$$y_{t_1} \cdot R^{-1}(t_1)H(t_1)x_{t_1} = \int_0^{t_1} R^{-1}Hx \cdot dy + \int_{t_1}^{t_1} y \cdot [(R^{-1}H)'x + R^{-1}H(Fx+Bv)]\,dt + \int_0^{t_1} y \cdot R^{-1}HG\,dw;$$

hence

$$J(v(\cdot)) = E\left[p(x_{t_1}, t_1)\lambda_{t_1} \exp\left(-\int_0^{t_1} y \cdot ((R^{-1}H)'x + R^{-1}H(Fx+Bv))\,dt - \frac{1}{2}\int_0^{t_1} H^*R^{-1}Hx \cdot x\,dt - \int_0^{t_1} y \cdot R^{-1}HG\,dw \right)\right].$$

Attention will be concentrated on

$$X = E\left[p(x_{t_1}, t_1)\lambda_{t_1} \exp\left(-\left[\int_0^{t_1} y \cdot ((R^{-1}H)'x + R^{-1}H(Fx+Bv))\,dt + \frac{1}{2}\int_0^{t_1} H^*R^{-1}Hx \cdot x\,dt + \int_0^{t_1} y \cdot R^{-1}HG\,dw \right]\right) \Big| Y^{t_1}\right].$$

Recall that y, w, x_0 are independent objects. Since v is adapted to Y, one can calulate X by freezing the values of y and v, and taking the expectation with respect to the remaining source of noise, namely w.

Note that for y and v frozen, $p(\cdot, \cdot)$ is a $C^{2,1}$ deterministic function. Therefore

$$\begin{aligned} dp(x_t, t) &= \left[\frac{\partial p(x,t)}{\partial t} + D_x p \cdot (Fx_t + Bv_t) + \frac{1}{2}\operatorname{tr}(D_x^2 pGG^*)\right] dt + D_x p \cdot Gw \\ &= \Big[D_x p \cdot GG^*H^*R^{-1}y - \frac{1}{2}p(|G^*H^*R^{-1}y|^2) \\ &\quad + \mu(Qx_t^2 + Nv_t^2) - H^*R^{-1}Hx_t^2 \\ &\quad - 2(Fx_t + Bvt) \cdot H^*R^{-1}y - 2(R^{-1}H)'x_t \cdot y_t)\Big]\,dt + D_x p \cdot G\,dw. \end{aligned} \tag{3.25}$$

Next

$$\begin{aligned} d(p(x_t, t)\lambda_t) &= \lambda_t[D_x p \cdot GG^*H^*R^{-1}y - \tfrac{1}{2}p(|G^*H^*R^{-1}y|^2 \\ &\quad - H^*R^{-1}Hx_t^2 - 2(Fx_t + Bv_t) \cdot H^*R^{-1}y_t - 2(R^{-1}H)'x_t \cdot y_t)]\,dt \\ &\quad + \lambda_t D_x p \cdot G\,dw_t. \end{aligned} \tag{3.26}$$

Let us denote

$$\theta_t = \exp\left[-\int_0^t y \cdot ((R^{-1}H)'x_t + R^{-1}H(Fx+Bv))\,dt + \frac{1}{2}\int_0^t R^{-1}H^*Hx \cdot x\,dt + \int_0^t y \cdot R^{-1}HG\,dw_t \right];$$

hence

$$\begin{aligned} d\theta_t &= \theta_t[-y \cdot ((R^{-1}H)'x + (R^{-1}H)(Fx+Bv))\,dt \\ &\quad + \tfrac{1}{2}H^*R^{-1}Hx_t^2\,dt - y_t \cdot R^{-1}HG\,dw_t + \tfrac{1}{2}|G^*H^*R^{-1}y|^2\,dt]. \end{aligned} \tag{3.27}$$

Combining (3.26) and (3.27), one obtain

$$dp(x_t, t)\lambda_t\theta_t = [\lambda_t\theta_t D_x p + \lambda_t\theta_t H^* R^{-1} y_r] \cdot G\, dw_t.$$

Recalling that y is frozen, one can take expectation with respect to w; hence

$$X = E[p(x_0, 0)|\, Y^{t_1}] = \int p(x, 0)\pi(x, 0)\, dx$$

by x_0 independent of y with a Gaussian distribution and by definition of $\pi(x, 0)$. Thus

$$J(v(\cdot)) = E[X] = E\left[\int p(x, 0)\pi(x, 0)\, dx\right] = K(v(\cdot))$$

by (3.24). □

Remark 3.9. Assumption (3.19) is not necessary to prove Theorem 3.2. because of the following argument. One first approximates $R^{-1}H$ by a differentiable function. In this case the preceding proof shows that Theorem 3.2 holds. Secondly one observes that the final result does not depend on the derivative and that hence one can pass to the limit. Condition (3.19) is thus only an intermediary technical assumption.

4. Solution of the stochastic control problem. The result of Theorem 3.2 implies that minimization of the cost functional $J(v(\cdot))$ is equivalent to the minimization of the cost functional $K(v(\cdot))$. The importance of this lies in the fact that the minimization of $K(v(\cdot))$ appears as a stochastic control problem with full information specified by the state equation (3.2) and cost functional (3.4). This problem is now easily solved.

It should however be pointed out that the function π defined in (3.3) is in general not the conditional density of x given y because it depends on the parameters Q and N of the cost functional. A key point in the proof of Theorem 3.2 is that one can push the cost functional into the expression for the conditional density. The fact that this is possible is based on the exponential form of the Gaussian density and the cost functional.

Consider the Riccati equation for $S: T \to R^{n\times n}$

$$\begin{aligned} &\dot S + S(F + \mu PQ) + (F^* + \mu QP)S + Q - S(BN^{-1}B^* - \mu PH^*R^{-1}HP)S = 0,\\ &S(t_1) = \tfrac{1}{2}[(I - \mu MP(t_1))^{-1}M + M(I - \mu P(t_1)M)^{-1}]. \end{aligned} \tag{4.1}$$

It will be assumed that (4.1) has a symmetric solution. Consider the equation

$$d\hat r = [F - PH^*R^{-1}H + \mu PQ - BN^{-1}B^*S]\hat r\, dt + PH^*R^{-1}\, dy, \qquad \hat r_0 = \mu_0, \tag{4.2}$$

which corresponds to the state equation (3.2) with the control

$$u_t^* = -N^{-1}(t)B^*(t)S(t)\hat r_t. \tag{4.3}$$

For some r given by (3.2) define

$$\begin{aligned} h_t = &\int_0^t R^{-1}Hr \cdot dy - \frac{1}{2}\int_0^t H^*R^{-1}Hr \cdot r\, ds\\ &+ \frac{\mu}{2}\int_0^t (Qr^2 + Nv^2)\, ds + \frac{\mu}{2}S(t)r_t \cdot r_t\\ &- \frac{\mu}{2}\int_0^t \operatorname{tr}(SPH^*R^{-1}HP)\, ds. \end{aligned} \tag{4.4}$$

Attention is now restricted from the class of admissible controls $\mathbf{U}_2$ to

$$\mathbf{U}_3 = \left\{ u \in \mathbf{U}_2 \,\middle|\, E\left[\int_0^{t_1} \exp(2h_s) |R^{-1} H(I + \mu PS) r|^2 \, ds \right] < \infty \right\}.$$

THEOREM 4.1. *Assume the conditions of Theorem* 3.2, *and that a symmetric solution to* (4.1) *exists. Assume that* u^*, *defined by* (4.3), *belongs to the class of admissible controls, and moreover to* $\mathbf{U}_3$.

a. *Then* u^* *is optimal and*

$$\min_{v \in \mathbf{U}_3} J(v(\cdot)) = J(u^*(\cdot))$$

$$= \mu \exp\left(\frac{\mu}{2} \left[S_0 \mu_0 \cdot \mu_0 + \int_0^{t_1} \operatorname{tr}(PQ + SPH^* R^{-1} HP) \, ds \right] \right) \cdot |[I - \mu MP(t_1)]|^{-1/2}.$$

b. *Then also* u^* *is conditionally optimal in* $\mathbf{U}_3$.

Remarks 4.2. 1. In the representation of the solution as given by (3.1), (4.2), and (4.3), the separation property does not hold in general. Note that in the sufficient statistic π for the cost functional both r and P depend on the state cost matrix Q, while the control Riccati differential equation for S depends on the matrix function P.

2. Observe that the function $\pi^v(x, t)$ is in general not the conditional density of x_t given Y^t since its parameters r and P depend on the cost functional through Q.

3. The concept of a sufficient statistic for a stochastic control problem has been defined by C. Striebel [9, 3.2]. For the stochastic control problem under consideration it follows from the proof of Theorem 4.1 that $\hat{r}$, as defined by (4.2), is a sufficient statistic. Note that because P is a deterministic function, it is therefore considered not to be a sufficient statistic.

4. An attempt to define minimality of a sufficient statistic for a stochastic control problem will not be made here. Because $\hat{r}$ takes values in R^n, the state space of the given stochastic system, it seems likely that this sufficient statistic is minimal in any reasonable sense. In the special case of $G = 0$ a sufficient statistic of much higher dimension has been found in Speyer et al. [7] for discrete-time systems, and in Kumar, van Schuppen [6] for continuous-time systems.

5. Theorem 4.1 contains the special cases discussed by Speyer et al. [7], with $Q = 0$, and by Kumar, van Schuppen [6], with $G = 0$. The discrete-time version of the problem considered here has been solved by Whittle [11, Thm. 5], see also [12, Part 1, Chap. 19, Thm. 0.1].

6. When μ is small, one has that

$$[J_\mu(v(\cdot)) - \mu]/\mu^2 \to \frac{1}{2} \tilde{E}^v \left[\int_0^{t_1} (Qx^2 + Nv^2) \, ds + Mx_{t_1}^2 \right].$$

Hence for μ small the LEG stochastic control problem becomes close to the standard linear-quadratic-Gaussian stochastic control problem for which the separation principle holds. This can also be seen from the explicit expressions for the optimal control. For μ small, S becomes close to the solution of the Riccati differential equation of the deterministic linear quadratic control problem

$$\dot{\Pi} + \Pi F + F^* \Pi + Q - \Pi B N^{-1} B^* \Pi = 0, \qquad \Pi(t_1) = M.$$

Moreover, then (3.1) reduces to the Riccati equation of the Kalman filter, and (3.2) reduces to the Kalman filter itself.

A sufficient condition for conditional optimality that will be used in the proof of Theorem 4.1, will be stated.

THEOREM 4.3 (C. Striebel). *If there exists a* $u^* \in \mathbf{U}_3$, *and for any admissible control* $v \in \mathbf{U}_3$ *an* Y^t *adapted process* $h^v : \Omega \times T \to R$ *such that*

1. *for any* $v \in \mathbf{U}_2$ $\tilde{E}^v[c|Y^{t_1}] = h^v_{t_1}$, *and* h^v *is a submartingale on* Y^t *with respect to* $\tilde{P}^v$;
2. h^{v^*} *is a* Y^t *martingale with respect to* $\tilde{P}^{v^*}$;

then v^* *is conditionally optimal. One calls* h^v *the conditional cost functional associated with* v.

Proof of Theorem 4.1. a) By Theorem 3.2 $J(v(\cdot)) = K(v(\cdot))$. Recall that

$$h_t = \int_0^t R^{-1}Hr \cdot dy - \frac{1}{2}\int_0^t H^*R^{-1}Hr \cdot r\, ds + \frac{\mu}{2}\int_0^t (Qr^2 + Nv^2)\, ds + \frac{\mu}{2} S(t)r_t \cdot r_t - \frac{\mu}{2}\int_0^t \operatorname{tr}(SPH^*R^{-1}HP)\, ds.$$

Then, from (3.17) and (4.1), one obtains

$$K(v(\cdot)) = \mu \exp\left(\frac{\mu}{2}\int_0^{t_1} \operatorname{tr}(PQ + SPH^*R^{-1}HP)\, ds\right) \cdot |[I - \mu MP(t_1)]|^{-1/2} E[\exp(h_{t_1})]. \tag{4.5}$$

Calculations show that

$$dh_t = R^{-1}Hr \cdot dy - \frac{1}{2}H^*R^{-1}Hr \cdot r\, dt + \frac{\mu}{2}(Qr^2 + Nv^2)\, dt + \frac{\mu}{2}\dot{S}r \cdot r\, dt + \mu Sr \cdot [(F - PH^*R^{-1}H + \mu PQ)r + Bv]\, dt + \mu Sr \cdot PH^*R^{-1}\, dy,$$

$$d\exp(h_t) = \exp(h_t)\Big[R^{-1}H(I + \mu PS)r \cdot dy - \frac{1}{2}H^*R^{-1}Hr \cdot r\, dt + \frac{\mu}{2}(Qr^2 + Nv^2)\, dt + \frac{\mu}{2}\dot{S}r \cdot r\, dt + \mu Sr \cdot [(F - Ph^*R^{-1}H + \mu PQ)r + Bv]\, dt + \frac{1}{2}|R^{-1/2}H(I + \mu PS)r|^2\, dt\Big]$$

$$= \exp(h_t)\Big[R^{-1}H(I + \mu PS)r \cdot dy + \frac{\mu}{2}(\dot{S} + S(F + \mu PQ) + (F^* + \mu QP)S - S(BN^{-1}B^* - \mu PH^*R^{-1}HP)S + Q)r \cdot r\, dt + \frac{\mu}{2}N(v + N^{-1}B^*Sr)^2\, dt\Big]. \tag{4.6}$$

If S satisfies (4.1), then

$$\exp(h_{t_1}) \geqq \exp\left(\frac{\mu}{2}S_0r_0r_0 \cdot r_0\right) + \int_0^{t_1} \exp(h_s)R^{-1}H(I + \mu PS)r \cdot dy \tag{4.7}$$

and from (4.5) and the definition of $\mathbf{U}_3$ then follows that

$$K(v(\cdot)) \geqq \mu \exp\left(\frac{\mu}{2}\left[S_0\mu_0 \cdot \mu_0 + \int_0^{t_1} \operatorname{tr}(PQ + SPH^*R^{-1}HP)\, ds\right]\right) \cdot |[I - \mu MP(t_1)]|^{-1/2}. \tag{4.8}$$

For the control u^* defined by (4.3) one gets equality in (4.7) and (4.8); hence it is optimal and

$$\min_{u \in \mathbf{U}_2} J(v(\cdot)) = \min_{u \in \mathbf{U}_2} K(v(\cdot)) = \mu \exp\left(\frac{\mu}{2}\left[S_0\mu_0 \cdot \mu_0 + \int_0^{t_1} \operatorname{tr}(PQ + SPH^*R^{-1}HP)\, ds\right]\right) \cdot |[I - \mu MP(t_1)]|^{-1/2}.$$

b) Let

$$c = \mu \exp\left(\frac{\mu}{2}\left[Mx_{t_1}^2 + \int_0^{t_1} (Qx^2 + Nv^2)\, ds\right]\right),$$

$$\rho_t = \exp\left(\int_0^t R^{-1} Hx \cdot dy - \frac{1}{2}\int_0^t H^*R^{-1}Hx \cdot x\, ds\right),$$

$$\bar{\rho}_t = \exp\left(\int_0^t R^{-1}Hr \cdot dy - \frac{1}{2}\int_0^t H^*R^{-1}Hr \cdot r\, ds\right),$$

$$h_t = \mu a_t \exp\left(\frac{\mu}{2}\left[S_t r_t \cdot r_t + \int_0^t (Qr^2 + Nv^2)\, ds\right]\right),$$

$$a_t = \exp\left(\frac{\mu}{2}\left[\int_0^{t_1} \operatorname{tr}(PQ)\, ds + \int_t^{t_1} (PH^*R^{-1}HPS)\, ds\right]\right)|[I - \mu MP(t_1)]|^{-1/2},$$

$$k_t = h_t\bar{\rho}_t.$$

By the proof of Theorem 3.2

$$E[c\rho_{t_1} | Y^{t_1}] = X = \int p(x, 0)\pi(x, 0)\, dx,$$

which by the proof of Lemma 3.7 equals

$$= \int \mu \exp\left(\frac{\mu}{2}Mx^2\right)\pi(x, t_1)\, dx.$$

With the calculations above (3.17) one obtains

$$E[c\rho_{t_1} | Y^{t_1}] = k_{t_1} = h_{t_1}\bar{\rho}_{t_1}.$$

Setting in this expression $M = 0$, $Q = 0$, $N = 0$, it materializes that

$$E[\mu\rho_{t_1} | F^{y_1}] = \mu\bar{\rho}_{t_1},$$
$$\tilde{E}^v[c | Y^{t_1}] = E[c\rho_{t_1} | Y^{t_1}] / E[\rho_{t_1} | Y^{t_1}] = h_{t_1}.$$

It is claimed that if k is a submartingale with respect to P, that then h is a

submartingale with respect to $\tilde{P}^v$. For if $s, t \in T$, $s < t$, then

$$\begin{aligned}\tilde{E}^v[h_t|Y^s] &= E[h_t\rho_{t_1}|Y^s]/E[\rho_{t_1}|Y^s] \\ &= E[h_tE[\bar{\rho}_{t_1}|Y^{t_1}]|Y^s]/E[E[\bar{\rho}_{t_1}|Y^{t_1}]|Y^s] \\ &= E[h_tE[\bar{\rho}_{t_1}|Y^t]|Y^s]/E[\bar{\rho}_{t_1}|Y^s] \\ &= E[h_t\bar{\rho}_t|Y^s]/\bar{\rho}_s \geqq h_s\bar{\rho}_s/\bar{\rho}_s = h_s.\end{aligned}$$

It is then clear that if k is martingale with respect to P, that then also h is a martingale with respect to $\tilde{P}^v$.

Using the fact that S satisfies (4.1) and that u^* is given by (4.3), a lengthy calculation shows that

$$\begin{aligned}dk = d(h\bar{\rho}) = \left(\frac{\mu^2}{2}\right)a_t \exp\left(\frac{\mu}{2}\left[S_tr_t \cdot r_t + \int_0^t (Qr^2 + Nv^2)\,ds\right]\right) \\ \cdot [N(v_t - u_t^*)^2\,dt + 2(R^{-1}Hr/\mu + R^{-1}HPSr) \cdot dy].\end{aligned}$$

Thus for any $v \in \mathbf{U}_3$, $k = h\bar{\rho}$ is a submartingale with respect to P, and for $v = u^*$ a martingale. By the above claim h is then for any $v \in \mathbf{U}_2$ a $\tilde{P}^v$ submartingale and for $v = u^*$ a martingale. From Theorem 4.3 it then follows that u^* is conditionally optimal. □

Note that in the proof of Theorem 4.1 a key element is the invariance of the conditional cost function h.

REFERENCES

[1] A. BENSOUSSAN AND J. L. LIONS, *Applications of Variational Inequalities in Stochastic Control*, North-Holland, Amsterdam, 1982.

[2] M. H. A. DAVIS, *Pathwise non-linear filtering*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., D. Reidel, Dordrecht, 1981, pp. 505–528.

[3] I. I. GIKHMAN AND I. V. SKOROKHOD, *Stochastic Differential Equations*, Springer-Verlag, Berlin, 1972.

[4] D. H. JACOBSON, *Optimal stochastic linear systems with exponential criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, AC-18 (1973), pp. 124–131.

[5] J. C. KRAINAK, F. W. MACHELL, S. I. MARCUS AND J. L. SPREYER, *The dynamic linear exponential Gaussian team problem*, IEEE Trans. Automat. Control, 27 (1982), pp. 860–869.

[6] P. R. KUMAR AND J. H. VAN SCHUPPEN, *On the optimal control of stochastic systems with an exponential-of-integral performance index*, J. Math. Anal. Appl., 80 (1981), pp. 312–332.

[7] J. L. SPEYER, J. DEYST AND D. H. JACOBSON, *Optimization of stochastic linear systems with additive measurement and process noise using exponential performance criteria*, IEEE Trans. Automat. Control, AC-19 (1974), pp. 358–366.

[8] J. L. SPEYER, *An adaptive terminal guidance scheme based on an exponential cost criterion with application to homing missile guidance*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 371–375.

[9] C. STRIEBEL, *Optimal control of discrete time stochastic systems*, Lecture Notes in Economics and Mathematical Systems 110, Springer-Verlag, Berlin, 1975.

[10] F. VAN DER PLOEG, *Risk-sensitive stabilization policy: complacency and neurotic breakdown*, preprint, 1983.

[11] P. WHITTLE, *Risk-sensitive linear/quadratic/Gaussian control*, Adv. Appl. Prob., 13 (1974), pp. 764–777.

[12] ———, *Optimization over Time: Dynamic Programming and Stochastic Control*, 2 volumes, John Wiley, New York, 1982–1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

008

ASYMPTOTIC EVOLUTION OF A STOCHASTIC CONTROL PROBLEM*

R. TARRES†

Abstract. Here we study the minimization problem of an integral discounted functional, on a set of nonexplosive and nonconstrained diffusions. The integrand is "weakly coercive," which leads us, using the dynamic programming method, to characterize the optimal cost among the solutions of the solving equation, with radiative conditions expressing the centripetal aspect of the optimal control. The evolution of the problem when the discount vanishes is then considered: the integrand being "strongly coercive" (i.e. its gradient being outward), a limit problem is defined and similarly solved; an inward optimal control exists, which is the limit of the ones of the initial problems. The existence properties are obtained by means of a priori estimates concerning suitable solutions of the solving equations in the whole space.

Key words. ergodic stochastic control, dynamic programming, a priori asymptotic estimates

1. Statement of the problems. Let us consider the stochastic differential equation of Ito type:

$$\xi(0)=x, \qquad d\xi(t)=p(\xi(t))\,dt+\rho\,dw(t), \tag{1.1}$$

where w is a normalized Brownian motion on $\mathbb{R}^m$ $(m\geqq 1)$, $\rho>0$ is a constant, $p\in\Lambda$.

Λ is the control set of our problem and is defined as follows: Λ is the set of polynomially growing[1] and locally Lipschitzian functions $p:\mathbb{R}^m\mapsto\mathbb{R}^m$ such that

$$\begin{aligned}&\text{there exists a positive constant } c_p \text{ such that}\\ &\xi p(\xi)\leqq c_p(1+|\xi|^2) \quad \text{for each } \xi\in\mathbb{R}^m.\end{aligned} \tag{1.2}$$

It is well known (see, for instance [3], [10], [11]) that for each initial state $x\in\mathbb{R}^m$ and for each control $p\in\Lambda$, the equation (1.1) has a solution $\xi_{x,p}$ defined on $[0,+\infty[$, unique in the sense of pathwise uniqueness on each interval $[0,T]$; $\xi_{x,p}$ is a nonexplosive diffusion process whose diffusion matrix is $\rho^2 I$ and whose drift is the closed-loop deterministic control p.

For each constant s $(s>0)$, $x\in\mathbb{R}^m$ and $p\in\Lambda$, the relation

$$J_{s,x}(p)=E\int_0^{+\infty} e^{-st}(g(\xi_{x,p}(t))+f(p(\xi_{x,p}(t))))\,dt \tag{1.3}$$

defines a *discounted* cost; the functions $f\in C^2(\mathbb{R}^m;\mathbb{R}_+)$ and $g\in C^2(\mathbb{R}^m;\mathbb{R}_+)$ are given. We are interested in the following two problems:

the stationary discounted problem $(P_{x,s})$: minimize $J_{s,x}(p)$, $p\in\Lambda$;

what is the behaviour of $(P_{x,s})$ when the discount s vanishes?

Constraint is imposed neither to the trajectories nor to the controls p of the diffusion: the processes $\xi_{x,p}$ evolve on the *whole* space $\mathbb{R}^m$ and the only restriction (1.2) above on admissible controls is that they have to be nonexplosive ($\xi_{x,p}$ is well defined for all $t>0$).

On the other hand, we shall impose some *coercivity* and *growth* hypotheses about the functions g and f in the running cost; particularly, the solution of the *asymptotic*

* Received by the editors April 5, 1983, and in revised form January 10, 1984.

† Département de Mathématiques, Faculté des Sciences et Techniques, 29283 Brest Cedex, France. Present address, Ecole Normale Supérieure d'Enseignement Polytechnique, B.P. 1523, Oran St. Charles, Algérie.

[1] $p:\mathbb{R}^m\mapsto\mathbb{R}^m$ is a polynomially growing function if and only if there exist positive constants b_p and m_p such that for all $\xi\in\mathbb{R}^m$, $|p(\xi)|\leqq b_p(1+|\xi|^{m_p})$. ($|\cdot|$ will denote the Euclidean norm.)

problem (i.e. the behaviour of $(P_{x,s})$ when s vanishes) will be interpreted as a minimization problem of the cost

$$\mu_x(p) = \liminf_{\tau \to +\infty} \left(\frac{1}{\tau} E \int_0^\tau (g(\xi_{x,p}(t)) + f(p(\xi_{x,p}(t)))) \, dt \right),$$

mainly by means of the hypothesis that the *gradient of g is outward (or centrifugal) enough*: in these conditions, the moments of $\xi_{x,p}(t)$ are bounded for any *inward (or centripetal) controls*, and particularly for the optimal one, which will be proved to be inward; this phenomenon (namely that the coercivity of g implies the centripetal aspect of the optimal control), which explains that the asymptotic problem can be solved, seems to us the main feature of our paper.

These problems have been studied in the case of a periodical running cost g and in the case of a reflected diffusion evolving in a bounded domain, and analogous results have been obtained (see the works of J. M. Lasry and P. L. Lions: [19], [20], [21], [22]). Other authors (see for example [6], [13], [25]) have studied the asymptotic evolution of some Markovian control problems depending on a parameter, and have obtained the existence of a limit problem with suitable convergence properties of the optimal policies.

For other works concerning discounted control problems having some connection with our paper (asymptotic problem, infinite horizon, processes evolving on the whole space, coercivity hypotheses), see for example [1], [2], [4], [5], [7], [14], [16], [23].

First we shall solve the problem $(P_{x,s})$ with a weak coercivity hypothesis about g; secondly the asymptotic problem will be considered by means of a stronger coercivity hypothesis (the gradient of g will be "outward"). For similar and other results about the one-dimensional case, see [27], [28].

2. The problem $(P_{x,s})$. It will be solved on the control set Λ_1 defined by

$$\begin{aligned} &p \in \Lambda_1 \text{ if and only if } p \in \Lambda \text{ and there exist } c'_p \geqq 0 \text{ and } \alpha_p \in [1, 2[\text{ such that} \\ &\xi p(\xi) \leqq c'_p(1 + |\xi|^{\alpha_p}) \quad \text{for each } \xi \in \mathbb{R}^m. \end{aligned} \tag{2.1}$$

The method of solution is the *dynamic programming* one (see for instance [3], [4], [5], [8], [9], [15]). The solving equation of $(P_{x,s})$ is

$$-\tfrac{1}{2}\rho^2 \Delta u(x) + su(x) + h(-\nabla u(x)) = g(x) \quad \text{for each } x \in \mathbb{R}^m \tag{2.2}$$

where $h = f^*$ is the conjugate function of f (according to convex analysis), defined on $\mathbb{R}^m$ by

$$h(z) = \sup_{y \in \mathbb{R}^m} (zy - f(y)) \quad \text{for each } z \in \mathbb{R}^m; \tag{2.3}$$

we assume that

$$\begin{aligned} &\text{for some constants } c_0 \text{ and } c'_0 \text{ such that } 0 < c_0 \leqq c'_0, \\ &c_0 \eta^2 \leqq \sum_{1 \leqq i,j \leqq m} \frac{\partial^2 f}{\partial y_i \, \partial y_j}(y) \eta_i \eta_j \leqq c'_0 \eta^2 \end{aligned} \tag{2.4}$$

for each $y \in \mathbb{R}^m$ and $\eta = (\eta_1, \cdots, \eta_m) \in \mathbb{R}^m$.

Therefore $h \in C^2(\mathbb{R}^m;]-f(0), +\infty[)$, f and h are strictly convex functions, ∇f and ∇h

are inverse C^1-diffeomorphisms of $\mathbb{R}^m$,

$$f(y)=\sup_{z\in\mathbb{R}^m}(yz-h(z)),$$
$$h(z)+f(y)=yz \text{ if and only if } z=(\nabla f)(y), \tag{2.5}$$

and

$$\frac{1}{c_0'}\zeta^2\leqq\sum_{1\leqq i,j\leqq m}\frac{\partial^2 h}{\partial z_i\,\partial z_j}(z)\zeta_i\zeta_j\leqq\frac{1}{c_0}\zeta^2$$

for each $z\in\mathbb{R}^m$ and $\zeta=(\zeta_1,\cdots,\zeta_m)\in\mathbb{R}^m$, and hence

$$\frac{\lambda}{c_0'}\leqq\frac{\partial h}{\partial\nu}(z)-\frac{\partial h}{\partial\nu}(z')\leqq\frac{\lambda}{c_0} \tag{2.6}$$

for each $\lambda\geqq 0$, $z\in\mathbb{R}^m$, $z'\in\mathbb{R}^m$ and $\nu\in\mathbb{R}^m$ satisfying $z-z'=\lambda\nu$ and $\nu^2=1$.

Let us make the additional following assumption:

For some positive constant μ_0,

$$\left|\frac{\partial h}{\partial\nu}(z)-\frac{\partial h}{\partial\nu}(z')\right|\leqq\mu_0 \tag{2.7}$$

for each $\nu\in\mathbb{R}^m$ with $\nu^2=1$ and for each z and $z'\in\mathbb{R}^m$ satisfying $(z-z')\nu=0$.

This last hypothesis does not exist if $m=1$; it restricts the lateral variation of $(\partial h/\partial\nu)(z)$. For example, if $f(y)=c_0y^2$, $h(z)=z^2/4c_0$ and (2.7) is satisfied with $\mu_0=0$.

In the classical stochastic control problem $(P_{x,s})$ with reflection on the boundary of a bounded set, we have a limit condition of Neumann type on the boundary of this set to characterize the optimal cost among all the solutions of the solving equation (2.2); in our problem, we do not have such conditions, and this characterization is obtained by means of a *radiative condition.* This question is naturally connected with those of the *coercivity and growth hypotheses* concerning the function g. The first one, which we call the weak coercivity hypothesis, is sufficient to the solution of $(P_{x,s})$; it consists in a limitation of the centripetal aspect of the gradient of g: the radial component of this gradient is more than a negative constant; on the other hand, the gradient of g may be centrifugal (but with a polynomial growth). Precisely, we shall assume that

For some positive constants K and K_1 and for some integer $q\geqq 3$,

$$\frac{\partial g}{\partial\nu}(\xi)\geqq -K-K_1(|\xi|-\xi\nu)^q|\xi|^q \tag{2.8}$$

for each $\xi\in\mathbb{R}^m$ and $\nu\in\mathbb{R}^m$ with $\nu^2=1$.

Let us note that this assumption implies the polynomial growth of ∇g, and that, in radial restriction, it can be written as follows:

$$\xi(\nabla g(\xi))\geqq -K|\xi| \quad \text{for each } \xi\in\mathbb{R}^m,$$
$$\text{or equivalently,}$$
$$\frac{\partial g}{\partial\nu}(\lambda\nu)\geqq -K \quad \text{for each } \lambda\geqq 0 \text{ and } \nu\in\mathbb{R}^m \text{ with } \nu^2=1. \tag{2.9}$$

If $m=1$, (2.8) is obviously equivalent to (2.9) with the polynomial growth of g'.

As a consequence of (2.8), we shall see that a *suitable radiative condition* serving to characterize the *optimal cost* u_s among all the solutions of the solving equation (2.2) expresses the suitable centripetal aspect of the corresponding optimal control: $(\nabla h)(-\nabla u_s(\cdot)) \in \Lambda_1$.

The solution of $(P_{x,s})$ on Λ_1 is summarized in Theorem 1:

THEOREM 1. *We make all the above assumptions. Then, for each fixed $s>0$, there exists one and only one solution $u \in C^3(\mathbb{R}^m;\mathbb{R})$ of the solving equation* (2.2) *satisfying the "weak radiative condition"*

$$(\nabla h)(-\nabla u(\cdot)) \in \Lambda_1. \tag{2.10}$$

Let u_s denote this solution, and p_s denote the control defined by

$$p_s(\xi) = (\nabla h)(-\nabla u_s(\xi)) \quad \textit{for each } \xi \in \mathbb{R}^m. \tag{2.11}$$

Then

$$p_s \in \Lambda_1$$

and for each $p \in \Lambda_1$ and $x \in \mathbb{R}^m$, (2.12)

$$0 \leqq u_s(x) = J_{s,x}(p_s) \leqq J_{s,x}(p);$$

in other words, for each $x \in \mathbb{R}^m$, $u_s(x)$ is the optimal cost for $(P_{x,s})$ on Λ_1 and p_s is an optimal control (independent of the initial state x) for this problem.

Remark. Let us denote by Λ_2 the subset of Λ_1 defined by

$p \in \Lambda_2$ if and only if $p \in \Lambda$ and there exists $c'_p \in \mathbb{R}_+$ such that (2.13)

$$\xi p(\xi) \leqq c'_p(1+|\xi|) \quad \text{for each } \xi \in \mathbb{R}^m.$$

Then we have $p_s \in \Lambda_2 \subset \Lambda_1$.

For the proof, see § 4.

3. Asymptotic behaviour of $(P_{x,s})$ when s vanishes. This study leads us to introduce the asymptotic stationary problem

$$(Q_x): \qquad \text{minimize } \mu_x(p), \qquad p \in \Lambda_3,$$

where $\mu_x(p)$ is defined by

$$\mu_x(p) = \liminf_{\tau \to +\infty} \frac{1}{\tau} E \int_0^\tau (g(\xi_{x,p}(t)) + f(p(\xi_{x,p}(t)))) \, dt \tag{3.1}$$

for each $x \in \mathbb{R}^m$ and $p \in \Lambda_3$,

and where the control set Λ_3 is the subset of Λ_2 defined by

$p \in \Lambda_3$ if and only if $p \in \Lambda$ and there exist two constants $c'_p \geqq 0$ and $d'_p > 0$ such that $\xi p(\xi) \leqq c'_p - d'_p|\xi|^2$ for each $\xi \in \mathbb{R}^m$. (3.2)

We shall solve this problem by means of a coercivity hypothesis concerning g stronger than (2.8): the gradient of g is quite centrifugal; its radial component has at least a linear growth. Precisely, our strong coercivity hypothesis is the following:

For some constants $K \geqq 0$, $K_1 \geqq 0$, $K_2 > 0$ and for some integer $q \geqq 3$,

$$\frac{\partial g}{\partial \nu}(\xi) \geqq -K - K_1(|\xi| - \xi\nu)^q |\xi|^q + K_2 \xi\nu \tag{3.3}$$

for each $\xi \in \mathbb{R}^m$ and $\nu \in \mathbb{R}^m$ with $\nu^2 = 1$.

It is clear that (3.3) implies (2.8) and therefore the polynomial growth of ∇g; its radial restriction can be written as follows:

$$\frac{\partial g}{\partial \nu}(\lambda \nu) \geqq -K + K_2 \lambda \tag{3.4}$$

for each $\lambda \geqq 0$ and $\nu \in \mathbb{R}^m$, with $\nu^2 = 1$, or equivalently, for some positive constants K' and $K_2' > 0$, $\xi \nabla g(\xi) \geqq -K' + K_2' |\xi|^2$ for each $\xi \in \mathbb{R}^m$.

If $m = 1$, (3.3) is obviously equivalent to (3.4) with the polynomial growth of g'.

With such a coercivity hypothesis, the optimal control p_s of $(P_{x,s})$ is naturally quite centripetal (we shall see that $p_s \in \Lambda_3$, i.e. $p_s(\xi)$ is at least linearly centripetal for $|\xi|$ large enough; it tends to bring back the evolution of the process in the region where g takes small values; so that the situation is similar to the one corresponding to the case where the diffusion is reflecting at the boundary of a bounded set, the centripetal aspect of the optimal control and of reasonable controls ($p \in \Lambda_3$) replacing the reflection phenomenon. Therefore, it is not surprising that the solution of (Q_x) described in the Theorem 2 given below is analogous to the one obtained by J. M. Lasry (see [20]) in the case of a diffusion $\xi_{x,p}$ evolving on a bounded set of $\mathbb{R}^m$, with reflection at the boundary. In fact, this analogy is a conjecture of J. M. Lasry and is the starting point of the present work.

The solution of (Q_x) and the convergence properties are summarized in Theorem 2:

THEOREM 2. *We make all the above hypotheses (the assumption* (2.8) *of Theorem* 1 *is replaced by* (3.3)). *The conclusions of Theorem* 1 *are still valid; moreover,* $p_s \in \Lambda_3$ *for each* $s > 0$.

Let us consider the solving equation of (Q_x):

$$-\tfrac{1}{2}\rho^2 \Delta v(x) + \lambda + h(-\nabla v(x)) = g(x) \quad \text{for each } x \in \mathbb{R}^m \tag{3.5}$$

(*the unknown of* (3.5) *is the pair* $(\lambda, v) \in \mathbb{R} \times C^2(\mathbb{R}^m; \mathbb{R})$).

There exist a pair $(\lambda_0, v_0) \in \mathbb{R}_+ x C^3(\mathbb{R}^m; \mathbb{R})$ *and a sequence* $(s_n)_{n \in \mathbb{N}}$ *on* $]0, +\infty[$ *converging to* 0 *and such that*:

1. (λ_0, v_0) *is a solution of* (3.5) *satisfying the "strong radiative condition"*

$$p_0 = (\nabla h)(-\nabla v_0(\cdot)) \in \Lambda_3. \tag{3.6}$$

2. $\left.\begin{aligned} \lambda_0 &= \lim_{n \uparrow \infty} s_n u_{s_n} \\ p_0 &= \lim_{n \uparrow \infty} p_{s_n} \end{aligned}\right\}$ *C^1 convergence on all compact subsets of* $\mathbb{R}^m$

$v_0 = \lim_{n \uparrow \infty} (u_{s_n} - u_{s_n}(0))$ *C^2 convergence on all compact subsets of* $\mathbb{R}^m$.

3. *For each* $p \in \Lambda_3$ *and* $x \in \mathbb{R}^m$,

$$\lambda_0 = \mu_x(p_0) \leqq \mu_x(p); \tag{3.7}$$

in other words, for each $x \in \mathbb{R}^m$, *the constant* λ_0 *is the optimal cost* (*independent of* x) *for* (Q_x) *and* p_0 *is an optimal control* (*independent of* x) *for this problem.*

Finally, the "strong radiative condition" (3.6), *connected with* (3.5), *implies the uniqueness of* λ_0, *first term of the pair* $(\lambda_0, v_0) \in \mathbb{R} \times C^2(\mathbb{R}^m; \mathbb{R})$; *if* $m = 1$ *the same conditions, with* $v_0(0) = 0$, *implies the uniqueness of* $(\lambda_0, v_0) \in \mathbb{R} \times C^2(\mathbb{R}^m; \mathbb{R})$.

Remark.

$$\lambda_0 = \lim_{\tau \uparrow +\infty} \left(\frac{1}{\tau} E \int_0^\tau (g(\xi_{x,p}(t)) + f(p(\xi_{x,p}(t)))) \, dt \right).$$

For the proof, see § 5.

4. Proof of Theorem 1. (For details, see [29]. If $m = 1$, see [27], [28].)

This proof, which is given in § 4d, is a consequence of the following three lemmas; Lemma 3 contains the a priori estimates leading to the optimality of p_s by means of the stochastic calculus Lemmas 1 and 2.

4a.

LEMMA 1. *Let $u_s \in C^2(\mathbb{R}^m; \mathbb{R})$ be a solution of* (2.2); *and suppose that ∇u_s is a polynomially growing function. Then*

for each $p \in \Lambda$, $\tau \geqq 0$ and $x \in \mathbb{R}^m$,

$$u_s(x) \leqq E \int_0^\tau e^{-st} (g(\xi_{x,p}(t)) + f(p(\xi_{x,p}(t)))) \, dt + e^{-s\tau} E(u_s(\xi_{x,p}(\tau))) \tag{4.1}$$

and this relation becomes an equality if

$$p = p_s = (\nabla h)(-\nabla u_s(\cdot)) \in \Lambda. \tag{4.2}$$

Proof. This property is a consequence of the Ito formula applied to the process $\alpha_{s,x,p}$ defined by $\alpha_{s,x,p}(t) = e^{-st} u_s(\xi_{x,p}(t))$; then it is sufficient to write the mathematical expectations using (2.2) and the definition of h.

4b.

LEMMA 2. *If $\varphi : \mathbb{R}^m \mapsto \mathbb{R}$ is a polynomially growing and measurable function, and if $p \in \Lambda_1$, then $E(\varphi(\xi_{x,p}(\cdot)))$ is a polynomially growing function for all $x \in \mathbb{R}^m$.*

Proof. It is sufficient to verify the result when $\varphi(\xi) = (\xi^2)^n$, $n \in \mathbb{N}^*$. Let L_p be the differential operator associated with (2.2):

$$L_p u = \tfrac{1}{2} \rho^2 \Delta u + p \nabla u. \tag{4.3}$$

Since $p \in \Lambda_1$, there exist $c_p'' \geqq 0$, $d_p'' \geqq 0$ and $\alpha_p' \in]0, 1[$ such that,

$$(L_p \varphi)(\xi) \leqq G(\varphi(\xi)) \quad \text{for each } \xi \in \mathbb{R}^m, \tag{4.4}$$

where $G(u) = c_p'' u^{\alpha_p'} + d_p''$ for all $u \in \mathbb{R}_+$.

It is well known that, if $m(t) = E(\varphi(\xi_{x,p}(t)))$, then the right derivative m_+' of m exists on $\mathbb{R}_+$, and, for each $t \in \mathbb{R}_+$

$$m_+'(t) = E(L_p \varphi(\xi_{x,p}(t))) \tag{4.5}$$

(see [9], [10]). Therefore

$$m_+'(t) \leqq E(G(\varphi(\xi_{x,p}(t)))) \leqq G(m(t)) \quad \text{for each } t \in \mathbb{R}_+ \tag{4.6}$$

because of Jensen's inequality and concavity of the function G; consequently, $m(t) \leqq \zeta(t)$, where ζ is the maximal solution, defined on $\mathbb{R}_+$, of the differential equation $du/dt = G(u)$, with initial condition $u(0) = (x^2)^n$ (for such results concerning differential inequalities, see [17]); the verification that ζ is a polynomially growing function is easy and completes the proof.

4c.

LEMMA 3. *Under the hypotheses of Theorem* 1, (2.2) *has at least one solution*

$u_s \in C^3(\mathbb{R}^m;\mathbb{R})$ *such that*

$$p_s = (\nabla h)(-\nabla u_s(\cdot)) \in \Lambda_2. \tag{4.7}$$

Proof. $m \geqq 2$ (if $m=1$, see [28]). First we shall prove a priori estimates about solutions $u_{s,r}$ of the solving equation (2.2) in the ball of radius r; secondly, the asymptotic evolution of $u_{s,r}$ when $r \to +\infty$ leads to the announced existence property.

A. *A priori estimates in the balls.* For each $s>0$ and $r>0$, let $u_{s,r} \in C^2(\bar{B}_r)$ (where $B_r = \mathscr{B}(0;r)$) be the solution of the solving equation (2.2) satisfying the Neumann condition on the boundary S_r of B_r

$$\frac{\partial u_{s,r}}{\partial \nu}(r\nu) = qr^{2q} \quad \text{for each } \nu \in S_1. \tag{4.8}$$

This existence and uniqueness property is classical; more precisely, for some $\alpha \in]0,1[$, $u_{s,r} \in C^{3,\alpha}(B_r) \cap C^{2,\alpha}(\bar{B}_r)$; see § 6 for a suitable proof, or [18, Chap. 10, Thm. 2.2] for a more general analysis of the problem.

We have for $u_{s,r}$ and $\partial u_{s,r}/\partial \nu$ the following estimates: for some $r_0>0$, $D>0$ and $q_1 > 2q+1$ (independent of r and s) and $\mathscr{C}_s > 0$ (independent of r),

$$\inf g - h(0) \leqq s u_{s,r}(\xi) \leqq s|\xi|^{q_1} + D, \tag{4.9}$$

$$\frac{\partial u_{s,r}}{\partial \nu}(\xi) \geqq -\mathscr{C}_s - (|\xi| - \xi\nu)^q |\xi|^q \tag{4.10}$$

for each $r \geqq r_0$, $s>0$, $\xi \in \bar{B}_r$ and $\nu \in S_1$.

Proof of (4.9). The function $U_{1,s}$ defined by $U_{1,s}(\xi) = (1/s)(\inf g - h(0))$ (respectively the function $U_{2,s}$ defined by $U_{2,s}(\xi) = |\xi|^{q_1} + D/s$, for sufficiently large $D>0$ and $q_1 > 2q+1$) for each $\xi \in \mathbb{R}^m$, is a subsolution (respectively a supersolution) of (2.2) in $\mathbb{R}^m$. The application of extremality conditions to $u_{s,r} - U_{1,s}$ (respectively to $u_{s,r} - U_{2,s}$) at a point where the minimum (respectively the maximum) of this function is reached implies (4.9): the minimum of $u_{s,r} - U_{1,s}$ cannot be reached on S_r since $\nu \nabla u_{s,r}(r\nu) > 0$ (respectively the maximum of $u_{s,r} - U_{2,s}$ cannot be reached on S_r since $\nu \nabla((u_{s,r} - U_{2,s})(r\nu)) = qr^{2q} - q_1 r^{q_1 - 1} < 0$ for sufficiently large r ($r \geqq r_0$) because of $q_1 - 1 > 2q$).

Proof of (4.10). 1) Let $V_{\nu,s}$ be the C^2 functions defined by

$$V_{\nu,s}(\xi) = -\mathscr{C}_s - (|\xi| - \xi\nu)^q |\xi|^q \quad \text{for each } \xi \text{ and } \nu \in \mathbb{R}^m. \tag{4.11}$$

We have, for $\mathscr{C}_s$ large enough, for each $\nu \in \mathbb{R}^m$, $\nu^2 = 1$:

$$\begin{aligned} &-\tfrac{1}{2}\rho^2 \Delta V_{\nu,s}(\xi) + sV_{\nu,s}(\xi) - (\nabla h(-V_{\nu,s}(\xi)\nu - q(|\xi| - \xi\nu)^{q-1}|\xi|^q \xi'_\nu))\nabla V_{\nu,s}(\xi) \\ &+ q\mu_0 (|\xi| - \xi\nu)^{q-1} |\xi|^{q-2} (2|\xi| - \xi\nu)|\xi'_\nu| \leqq \frac{\partial g}{\partial \nu}(\xi) \quad \text{for each } \xi \in \mathbb{R}^m, \end{aligned} \tag{4.12}$$

where ξ'_ν is the vector of $\mathbb{R}^m$ defined by $\xi'_\nu = \xi - (\xi\nu)\nu$. Therefore the functions $V_{\nu,s}$ are, in a certain sense, subsolutions of the derived equations of (2.2) in the direction ν.

The proof of (4.12) is quite technical; nevertheless we give it hereafter, but the reader may directly go to step 2.

The gradient and the Laplacian of $V_{\nu,s}$ are respectively

$$\nabla V_{\nu,s}(\xi) = -q|\xi|^{q-2}(-(|\xi| - \xi\nu)^{q+1}\nu + (|\xi| - \xi\nu)^{q-1}(2|\xi| - \xi\nu)\xi'_\nu), \tag{4.13}$$

$$\Delta V_{\nu,s}(\xi) = -q|\xi|^{q-2}(|\xi| - \xi\nu)^{q-1}((2m+5q-5)|\xi| - (m+3q-2)\xi\nu). \tag{4.14}$$

Because of the hypotheses concerning h, we have

$$(4.15)\quad \frac{\partial h}{\partial \nu}(-V_{\nu,s}(\xi)\nu - q(|\xi|-\xi\nu)^{q-1}|\xi|^q\xi'_\nu) \geqq -(\mu_0+\mu_1)+\frac{1}{c'_0}(\mathscr{C}_s+(|\xi|-\xi\nu)^q|\xi|^q)$$

and, if $\xi'_\nu \neq 0$, denoting $\xi''_\nu = (1/|\xi'_\nu|)\xi'_\nu$,

$$(4.16)\quad \frac{\partial h}{\partial \xi''_\nu}(-V_{\nu,s}(\xi)\nu - q(|\xi|-\xi\nu)^{q-1}|\xi|^q\xi'_\nu) \leqq \mu_0+\mu_1-\frac{q}{c'_0}(|\xi|-\xi\nu)^{q-1}|\xi|^q|\xi'_\nu|$$

where

$$\mu_1 = \sup_{\nu\in S_1}\left|\frac{\partial h}{\partial \nu}(0)\right| = |\nabla h(0)|.$$

Therefore the following relation implies (4.12):

$$\begin{aligned}
&\tfrac{1}{2}\rho^2 q|\xi|^{q-2}(|\xi|-\xi\nu)^{q-1}((2m+5q-5)|\xi|-(m+3q-2)\xi\nu)\\
&-s\mathscr{C}_s - s(|\xi|-\xi\nu)^q|\xi|^q\\
&-q|\xi|^{q-2}(|\xi|-\xi\nu)^{q+1}\left(-(\mu_0+\mu_1)+\frac{1}{c'_0}(\mathscr{C}_s+(|\xi|-\xi\nu)^q|\xi|^q)\right)\\
&+q|\xi|^{q-2}(|\xi|-\xi\nu)^{q-1}(2|\xi|-\xi\nu)|\xi'_\nu|\left((\mu_0+\mu_1)-\frac{q}{c'_0}(|\xi|-\xi\nu)^{q-1}|\xi|^q|\xi'_\nu|\right)\\
&+q\mu_0(|\xi|-\xi\nu)^{q-1}|\xi|^{q-2}(2|\xi|-\xi\nu)|\xi'_\nu|\\
&+K+K_1(|\xi|-\xi\nu)^q|\xi|^q \leqq 0 \quad \text{for each } \xi\in\mathbb{R}^m \text{ and } \nu\in S_1.
\end{aligned} \tag{4.17}$$

If $\mathscr{C}_s > c'_0(\mu_0+\mu_1)$, and because of

$$|\xi'_\nu| = (|\xi|-\xi\nu)^{1/2}(|\xi|+\xi\nu)^{1/2},$$

it is sufficient to have

$$\begin{aligned}
&\tfrac{1}{2}\rho^2 q(3m+8q-7)|\xi|^{q-1}(|\xi|-\xi\nu)^{q-1}-s\mathscr{C}_s+K\\
&-\frac{q}{c'_0}|\xi|^{2q-2}(|\xi|-\xi\nu)^{2q+1}\\
&+6q(2\mu_0+\mu_1)|\xi|^{q-1/2}(|\xi|-\xi\nu)^{q-1/2}\\
&-\frac{q^2}{c'_0}|\xi|^{2q-1}(|\xi|-\xi\nu)^{2q-1}(|\xi|+\xi\nu)+K_1(|\xi|-\xi\nu)^q|\xi|^q \leqq 0.
\end{aligned} \tag{4.18}$$

When $\xi\nu \geqq 0$, $|\xi|-\xi\nu \leqq |\xi| \leqq |\xi|+\xi\nu$; therefore we can replace (4.18) by

$$\begin{aligned}
&\tfrac{1}{2}\rho^2 q(3m+8q-7)(|\xi|(|\xi|-\xi\nu))^{q-1}-s\mathscr{C}_s+K\\
&+6q(2\mu_0+\mu_1)(|\xi|(|\xi|-\xi\nu))^{q-1/2}\\
&-\frac{q^2}{c'_0}(|\xi|(|\xi|-\xi\nu))^{2q-1/2}+K_1((|\xi|-\xi\nu)|\xi|)^q \leqq 0,
\end{aligned} \tag{4.19}$$

which is satisfied if $s\mathscr{C}_s \geqq K + C_1$, C_1 being the supremum of the function

$$t \mapsto -\frac{q^2}{c'_0}t^{2q-1/2}+K_1t^q+6q(2\mu_0+\mu_1)t^{q-1/2}+\tfrac{1}{2}\rho^2 q(3m+8q-7)t^{q-1}.$$

When $\xi\nu \leqq 0$, $|\xi| \leqq |\xi| - \xi\nu \leqq 2|\xi|$; in this case we can replace (4.18) by

$$\begin{aligned}&\frac{1}{2}\rho^2 q(3m+8q-7)2^{q-1}|\xi|^{2q-2} - \frac{q}{c_0'}|\xi|^{4q-1}\\ &+6q(2\mu_0+\mu_1)2^{q-1/2}|\xi|^{2q-1} + K_1 2^q|\xi|^{2q} \leqq s\mathscr{C}_s - K\end{aligned} \tag{4.20}$$

which is satisfied if $s\mathscr{C}_s \geqq K + C_2$, C_2 being the supremum of the left-hand side of (4.20).

In short, (4.12) is a consequence of

$$\mathscr{C}_s > c_0'(\mu_0+\mu_1), \quad s\mathscr{C}_s \geqq K + C_1, \quad s\mathscr{C}_s > K + C_2. \tag{4.21}$$

2) Now we shall prove (4.10), which can be written as follows:

$$v_{\nu,s,r}(\xi) = \nu\nabla u_{s,r}(\xi) \geqq V_{\nu,s}(\xi) \quad \text{for each } \xi \in \overline{B_r} \text{ and } \nu \in S_1. \tag{4.10}$$

Let us consider the application $(\nu, \xi) \mapsto w_\nu(\xi) = (v_{\nu,s,r} - V_{\nu,s})(\xi)$ which is continuous on the compact $S_1 \times \overline{B_r}$ (s and r are fixed); let (ν_0, ξ_0) be a point where the infimum of this function is reached. If $w_{\nu_0}(\xi_0) \geqq 0$, then $w_\nu(\xi) \geqq 0$ for each $(\nu, \xi) \in S_1 \times \overline{B_r}$. We shall see that the contrary ($w_{\nu_0}(\xi_0) < 0$) is impossible: Let us suppose that $w_{\nu_0}(\xi_0) < 0$.

Let us write the extremality condition of first order concerning the variations of $\nu \in S_1$, for $\nu \mapsto w_\nu(\xi_0)$; the gradient of this application, at point $\nu_0 \in S_1$, is

$$\nabla u_{s,r}(\xi_0) - q(|\xi_0| - \xi_0\nu_0)^{q-1}|\xi_0|^q\xi_0. \tag{4.22}$$

The gradient of $\nu \mapsto \nu^2 - 1$ is $2\nu_0$. Therefore, we have

$$\lambda\nu_0 = \nabla u_{s,r}(\xi_0) - q(|\xi_0| - \xi_0\nu_0)^{q-1}|\xi_0|^q\xi_0 \quad \text{for some } \lambda \in \mathbb{R}, \tag{4.23}$$

$$\lambda = \lambda\nu_0^2 = v_{\nu_0,s,r}(\xi_0) - q(|\xi_0| - \xi_0\nu_0)^{q-1}|\xi_0|^q\xi_0\nu_0, \tag{4.24}$$

$$\nabla u_{s,r}(\xi_0) = v_{\nu_0,s,r}(\xi_0)\nu_0 + q(|\xi_0| - \xi_0\nu_0)^{q-1}|\xi_0|^q\xi_{0\nu_0}'. \tag{4.25}$$

a) If $\xi_0 \in B_r$, we have, from the derived equation of (2.2) satisfied by $v_{\nu_0,s,r}$, from (4.12), (4.25) and from the next extremality conditions (since $\xi_0 \in B_r$)

$$\nabla v_{\nu_0,s,r}(\xi_0) = \nabla V_{\nu_0,s}(\xi_0) \quad \text{and} \quad \Delta w_{\nu_0}(\xi_0) \geqq 0, \tag{4.26}$$

after calculation and denoting (ν, ξ) in place of (ν_0, ξ_0):

$$\begin{aligned}sw(\xi) \geqq\ & q\mu_0(|\xi| - \xi\nu)^{q-1}|\xi|^{q-2}(2|\xi| - \xi\nu)|\xi_\nu'|\\ &+ q|\xi|^{q-2}(|\xi| - \xi\nu)^{q+1}\left(\frac{\partial h}{\partial \nu}(-v_{\nu,s,r}(\xi)\nu - q(|\xi| - \xi\nu)^{q-1}|\xi|^q\xi_\nu')\right.\\ &\qquad\left. -\frac{\partial h}{\partial \nu}(-V_{\nu,s}(\xi)\nu - q(|\xi| - \xi\nu)^{q-1}|\xi|^q\xi_\nu')\right)\\ &- q|\xi|^{q-2}(|\xi| - \xi\nu)^{q-1}(2|\xi| - \xi\nu)|\xi_\nu'|\\ &\cdot\left(\frac{\partial h}{\partial \xi_\nu''}(-v_{\nu,s,r}(\xi)\nu - q(|\xi| - \xi\nu)^{q-1}|\xi|^q\xi_\nu')\right.\\ &\qquad\left. -\frac{\partial h}{\partial \xi_\nu''}(-V_{\nu,s}(\xi)\nu - q(|\xi| - \xi\nu)^{q-1}|\xi|^q\xi_\nu')\right).\end{aligned} \tag{4.27}$$

Therefore, because of $-v_{\nu_0,s,r}(\xi_0) > -V_{\nu_0,s}(\xi_0)$ and (2.7), $sw_{\nu_0}(\xi_0) \geqq 0$, which is impossible.

b) If $\xi_0 \in S_r$, it follows from (4.25) and (4.8) that (denoting $(\nu, \xi) = (\nu_0, \xi_0)$):

$$w_\nu(\xi)\xi\nu = qr^{2q+1} - q(|\xi| - \xi\nu)^{q-1}|\xi|^q\xi_\nu'^2 + (\mathscr{C}_s + (|\xi| - \xi\nu)^q|\xi|^q)\xi\nu. \tag{4.28}$$

α) If $\xi_0\nu_0 \neq 0$, we have, denoting $|\xi_0| - \xi_0\nu_0 = ry$,

$$w_{\nu_0}(\xi_0)\xi_0\nu_0 = \mathscr{C}_s\xi_0\nu_0 + r^{2q+1}g_q(y), \tag{4.29}$$

where

$$g_q(y) = (q-1)y^{q+1} - (2q-1)y^q + q. \tag{4.30}$$

It is easy to prove that $g_q(y) \geqq 0$ if $0 \leqq y \leqq 1$ and that $g_0(y) \leqq 0$ if $1 \leqq y \leqq 2$; hence $w_{\nu_0}(\xi_0) \geqq 0$, which is impossible.

β) If $\xi_0\nu_0 = 0$: we draw one's inspiration from methods used by J. M. Lasry ([20]). Let $t \mapsto \eta_t$ be a curve of class C^1 in S_1 such that

$$r\eta_0 = \xi_0 \quad \text{and} \quad \eta_0' = \nu_0. \tag{4.31}$$

We have successively

$$\eta_t \nabla u_{s,r}(r\eta_t) = qr^{2q}, \tag{4.32}$$

$$\nu_0 \nabla u_{s,r}(r\eta_0) + r\eta_0 u''_{s,r}(r\eta_0)\nu_0 = 0. \tag{4.33}$$

Moreover, ξ_0 being a minimum of $w_{\nu_0}(\xi)$ for $\xi \in \overline{B_r}$, and η_0 being an outward normal to S_r at point ξ_0,

$$\frac{\partial}{\partial \eta_0}(v_{\nu_0,s,r}(\xi_0)) \leqq \frac{\partial}{\partial \eta_0}(V_{\nu_0,s}(\xi_0)) = -2qr^{2q-1} \leqq 0. \tag{4.34}$$

On the other hand,

$$\begin{aligned} \frac{\partial}{\partial \eta_0}(v_{\nu_0,s,r}(\xi_0)) &= \nu_0 u''_{s,r}(\xi_0)\eta_0 = \eta_0 u''_{s,r}(\xi_0)\nu_0 \\ &= -\frac{1}{r}\nu_0 \nabla u_{s,r}(r\eta_0) = -\frac{1}{r} v_{\nu_0,s,r}(\xi_0). \end{aligned} \tag{4.35}$$

Hence $v_{\nu_0,s,r}(\xi_0) \geqq 0$, which contradicts $v_{\nu_0,s,r}(\xi_0) < V_{\nu_0,s}(\xi_0) < 0$.

B. *A priori estimates in the whole space.* Now we shall study the asymptotic evolution of $u_{s,r}$ when $r \to +\infty$:

The above estimates (4.9) and (4.10) are uniform with respect to r; for each $\beta \in]0, 1[$ it is then possible to construct by recurrence $u_s \in C^{2,\beta}(\mathbb{R}^m; \mathbb{R})$, solution of (2.2) on $\mathbb{R}^m$ and a sequence $(r_n)_{n\in\mathbb{N}}$ in $[r_0, +\infty[$ such that

$$\lim_{n\uparrow+\infty} r_n = +\infty, \tag{4.36}$$

$$\lim_{n\uparrow+\infty} u_{s,r_n}/\overline{B_R} = u_s/\overline{B_R} \quad \text{in } C^{2,\beta}(\overline{B_R}) \quad \text{for each } R \geqq r_0, \tag{4.37}$$

$$\inf g - h(0) \leqq s u_s(\xi) \leqq s|\xi|^{q_1} + D \quad \text{for each } \xi \in \mathbb{R}^m, \tag{4.38}$$

$$\frac{\partial u_s}{\partial \nu}(\xi) \geqq -\mathscr{C}_s - (|\xi| - \xi\nu)^q |\xi|^q \quad \text{for each } (\nu, \xi) \in S_1 \times \mathbb{R}^m. \tag{4.39}$$

Indeed, for each $R \geqq r_0$, there exists a positive constant C, independent of $r \geqq 2R$ (depending on β, ρ, s, R, $|g|_{0,\beta}$,

$$\sup_{r\geqq 2R}\Big(\sup_{\xi\in B_{2R}} |u_{s,r}(\xi)|\Big) \quad \text{and} \quad \sup_{r\geqq 2R}\Big(\sup_{\xi\in B_{2R}} |\nabla u_{s,r}(\xi)|\Big);$$

these last two numbers are $< +\infty$ because of (4.9) and (4.10)) such that

$$|u_{s,r}/\overline{B_R}|_{2,\beta,B_R} \leqq C \quad \text{for each } r \geqq 2R. \tag{4.40}$$

This estimate is a consequence of (4.9), (4.10) and of the Hölder interior estimates of the gradient and the second derivatives concerning quasilinear elliptic equations of divergence form (see [18, Chap. 4, § 6]).

Therefore, the above existence property follows from the Ascoli-Arzela theorem. For the regularity property $u_s \in C^3(\mathbb{R}^m; \mathbb{R})$, see [11].

C. *The radiative condition.* We have

$$\nu \nabla h(-\nabla u_{s,r}(\lambda \nu)) = \nu \nabla h(-v_{\nu,s,r}(\lambda \nu)\nu + V(\lambda, \nu, s, r)) \tag{4.41}$$

for each $\nu \in S_1$ and $\lambda \geqq 0$, $V(\lambda, \nu, s, r)$ being orthogonal to ν. Hence, applying (4.10) and (2.7) gives

$$\nu \nabla h(-\nabla u_{s,r}(\lambda \nu)) \leqq \mu_0 + \mu_1 + \frac{\mathscr{C}_s}{c_0} \tag{4.42}$$

and

$$\nu \nabla h(-\nabla u_s(\lambda \nu)) \leqq \mu_0 + \mu_1 + \frac{\mathscr{C}_s}{c_0} \tag{4.43}$$

for each $\nu \in S_1$ and $\lambda \geqq 0$, which completes the proof of Lemma 3.

4d. We are now in a position to deduce Theorem 1 from Lemmas 1, 2, 3. Let u_s be a solution of (2.2) satisfying (4.7). ∇u_s has polynomial growth, and hence so does u_s; therefore

$$\lim_{t \uparrow +\infty} e^{-st} E(u_s(\xi_{x,p}(t))) = 0 \quad \text{for each } p \in \Lambda_1 \text{ and } x \in \mathbb{R}^m. \tag{4.44}$$

Lemma 1 completes the proof.

5. Proof of Theorem 2. For details, see [29]. If $m = 1$, see [28]. This proof follows the same methods as those of Theorem 1.

5a.

LEMMA 4. *Let* $(\lambda_0, v_0) \in \mathbb{R} \times C^2(\mathbb{R}^m; \mathbb{R})$ *be a solution of* (3.5) *such that* ∇v_0 *has polynomial growth. Then*

For each $p \in \Lambda$, $\tau > 0$ *and* $x \in \mathbb{R}^m$,

$$\lambda_0 \leqq \frac{1}{\tau} E \int_0^\tau (g(\xi_{x,p}(t)) + f(p(\xi_{x,p}(t)))) \, dt - \frac{1}{\tau} v_0(x) + \frac{1}{\tau} E(v_0(\xi_{x,p}(\tau))), \tag{5.1}$$

and this relation becomes an equality if

$$p = p_0 = (\nabla h)(-\nabla v_0(\cdot)) \in \Lambda. \tag{5.2}$$

Proof. We apply the Ito formula to the process $\alpha_{x,p}$ defined by $\alpha_{x,p}(t) = -\lambda_0 t + v_0(\xi_{x,p}(t))$. Then it is sufficient to write the mathematical expectations using (3.5) and the definition of h.

5b.

LEMMA 5. *If* $\varphi: \mathbb{R}^m \mapsto \mathbb{R}$ *is a polynomially growing and measurable function, and if* $p \in \Lambda_3$, *then for each* $x \in \mathbb{R}^m$, $E(\varphi(\xi_{x,p}(\cdot)))$ *is a bounded function.*

Proof. The proof is similar to that of Lemma 2; using the same notation, since $p \in \Lambda_3$, there exist $c_p'' > 0$ and $d_p'' > 0$ such that

$$\begin{aligned} &(L_p \varphi)(\xi) \leqq G(\varphi(\xi)) \quad \text{for each } x \in \mathbb{R}^m (\varphi(\xi) = |\xi|^{2n}) \\ &\text{where } G(u) = c_p'' - d_p'' u \text{ for all } u \in \mathbb{R}_+; \end{aligned} \tag{5.3}$$

hence, $m(t) = E(\varphi(\xi_{x,p}(t))) \leqq \zeta(t)$, where ζ is the maximal solution, defined on $\mathbb{R}_+$, of the differential equation $du/dt = G(u)$ with initial condition $u(0) = (x^2)^n$. The verification that ζ is bounded is easy and completes the proof.

5c.

LEMMA 6. *Under the hypotheses of Theorem* 2, (2.2) *has at least one solution* $u_s \in C^3(\mathbb{R}^m; \mathbb{R}_+)$ *such that*

$$p_s = (\nabla h)(-\nabla u_s(\cdot)) \in \Lambda_3. \tag{5.4}$$

Furthermore, there exist a pair $(\lambda_0, v_0) \in \mathbb{R}_+ \times C^3(\mathbb{R}^m; \mathbb{R})$ *and a sequence* $(s_n)_{n\in\mathbb{N}}$ *on* $]0, +\infty[$ *converging to* 0 *and such that*:

1. (λ_0, v_0) *is a solution of* (3.5) *satisfying the "strong radiative condition"*

$$p_0 = (\nabla h)(-\nabla v_0(\cdot)) \in \Lambda_3; \tag{3.6}$$

2.

$$\begin{aligned} &\lambda_0 = \lim_{n\uparrow+\infty} s_n u_{s_n} \quad (C^1 \text{ convergence on all compact subsets of } \mathbb{R}^m) \\ &v_0 = \lim_{n\uparrow+\infty} (u_{s_n} - u_{s_n}(0)) \quad (C^2 \text{ convergence on all compact subsets of } \mathbb{R}^m). \end{aligned} \tag{5.5}$$

Proof. $m \geqq 2$ (if $m = 1$, see [28]). The proof is similar to that of Lemma 3; the main difference between Lemmas 3 and 5 lies in the a priori estimates of $su_{s,r}$ and $\nabla u_{s,r}$, which are now independent of $s < s_0$ (and of r, as in Lemma 3). Let s_0 be a positive constant.

A. *A priori estimates in the balls.* For each $s \in]0, s_0[$ and $r > 0$, let $u_{s,r} \in C^2(\overline{B_r})$ be the solution of (2.2) satisfying the Neumann condition on S_r

$$\frac{\partial u_{s,r}}{\partial \nu}(r\nu) = qr^{2q} + \mathscr{C}' r \quad \text{for each } \nu \in S_1, \tag{5.6}$$

where $\mathscr{C}'$ is a constant such that

$$s_0 \mathscr{C}' + \frac{\mathscr{C}'^2}{c_0} < K_2 \quad \text{and} \quad \mathscr{C}' > 0. \tag{5.7}$$

We have for $u_{s,r}$ and $\partial u_{s,r}/\partial \nu$ the following estimates:

For some $r_0 > 0$, $D > 0$, $q_1 > 2q + 1$, $\mathscr{C} > 0$ and $\mathscr{C}' > 0$, we have

$$\inf g - h(0) \leqq s u_{s,r}(\xi) \leqq s_0 |\xi|^{q_1} + D \tag{5.8}$$

and

$$\frac{\partial u_{s,r}}{\partial \nu}(\xi) \geqq -\mathscr{C} - (|\xi| - \xi\nu)^q |\xi|^q + \mathscr{C}' \xi\nu \tag{5.9}$$

for each $r \geqq r_0$, $s \in]0, s_0]$, $\xi \in \overline{B_r}$ and $\nu \in S_1$.

Proof of (5.8). This proof is analogous to that of (4.9); the limit conditions are modified, but this fact has no consequences on the reasoning.

Proof of (5.9). 1) The functions V_ν defined for some constants $\mathscr{C} > 0$ and $\mathscr{C}' > 0$ by

$$V_\nu(\xi) = -\mathscr{C} - (|\xi| - \xi\nu)^q |\xi|^q + \mathscr{C}' \xi\nu \quad \text{for each } \xi \in \mathbb{R}^m \text{ and } \nu \in S_1 \tag{5.10}$$

are subsolutions, in the following sense, of the derived equation of (2.2) in the direction ν:

for each $\nu \in S_1$, $s \in]0, s_0]$, $\xi \in \mathbb{R}^m$,

$$
\begin{aligned}
&-\tfrac{1}{2}\rho^2 \Delta V_\nu(\xi) + sV_\nu(\xi) \\
&-\nabla h(-V_\nu(\xi)\nu - (q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')\xi'_\nu)\nabla V_\nu(\xi) \\
&+ q\mu_0(|\xi| - \xi\nu)^{q-1}|\xi|^{q-2}(2|\xi| - \xi\nu)|\xi'_\nu| \leqq \frac{\partial g}{\partial \nu}(\xi).
\end{aligned} \tag{5.11}
$$

The proof of (5.11) is quite technical; we give it below, but the reader may directly go to step 2. See Lemma 3 for the definition of ξ'_ν, ξ''_ν and μ_1. Because of the hypotheses concerning h we have:

$$
\begin{aligned}
&\frac{\partial h}{\partial \nu}(-V_\nu(\xi)\nu - (q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')\xi'_\nu) \\
&\qquad \geqq -(\mu_0 + \mu_1) + \frac{1}{c'_0}(\mathscr{C} + (|\xi| - \xi\nu)^q|\xi|^q) - \frac{\mathscr{C}'}{c''_0}\xi\nu
\end{aligned} \tag{5.12}
$$

where $c''_0 = c_0$ if $\xi\nu \geqq 0$ and $c''_0 = c'_0$ if $\xi\nu \leqq 0$, and, if $\xi'_\nu \neq 0$,

$$
\begin{aligned}
&\frac{\partial h}{\partial \xi''_\nu}(-V_\nu(\xi)\nu - (q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')\xi'_\nu) \\
&\qquad \leqq (\mu_0 + \mu_1) - \frac{1}{c'_0}(q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')|\xi'_\nu|.
\end{aligned} \tag{5.13}
$$

Therefore the following relation implies (5.11):

$$
\begin{aligned}
&\tfrac{1}{2}\rho^2 q|\xi|^{q-2}(|\xi| - \xi\nu)^{q-1}((2m + 5q - 5)|\xi| - (m + 3q - 2)\xi\nu) \\
&- s\mathscr{C} - s(|\xi| - \xi\nu)^q|\xi|^q + s\mathscr{C}'\xi\nu \\
&- (q|\xi|^{q-2}(|\xi| - \xi\nu)^{q+1} + \mathscr{C}')\left(-(\mu_0 + \mu_1) + \frac{1}{c'_0}(\mathscr{C} + (|\xi| - \xi\nu)^q|\xi|^q) - \frac{\mathscr{C}'}{c''_0}\xi\nu\right) \\
&+ q|\xi|^{q-2}(|\xi| - \xi\nu)^{q-1}(2|\xi| - \xi\nu)|\xi'_\nu| \\
&\quad \cdot \left((\mu_0 + \mu_1) - \frac{1}{c'_0}(q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')|\xi'_\nu|\right) \\
&+ q\mu_0(|\xi| - \xi\nu)^{q-1}|\xi|^{q-2}(2|\xi| - \xi\nu)|\xi'_\nu| \\
&+ K + K_1(|\xi| - \xi\nu)^q|\xi|^q - K_2\xi\nu \leqq 0.
\end{aligned} \tag{5.14}
$$

If $\mathscr{C} > c'_0(\mu_0 + \mu_1)$, it is sufficient to have

$$
\begin{aligned}
&\frac{1}{2}\rho^2 q(3m + 8q - 7)|\xi|^{q-1}(|\xi| - \xi\nu)^{q-1} + \left(s_0\mathscr{C}' + \frac{\mathscr{C}'^2}{c''_0} - K_2\right)\xi\nu \\
&- \mathscr{C}'\left(\frac{\mathscr{C}}{c'_0} - \mu_0 - \mu_1\right) + K - \frac{q}{c'_0}|\xi|^{2q-2}(|\xi| - \xi\nu)^{2q+1} \\
&+ q\frac{\mathscr{C}'}{c''_0}\xi\nu|\xi|^{q-2}(|\xi| - \xi\nu)^{q+1} \\
&+ 6q(2\mu_0 + \mu_1)|\xi|^{q-1/2}(|\xi| - \xi\nu)^{q-1/2} \\
&- \frac{q^2}{c'_0}|\xi|^{2q-1}(|\xi| - \xi\nu)^{2q-1}(|\xi| + \xi\nu) + K_1|\xi|^q(|\xi| - \xi\nu)^q \leqq 0.
\end{aligned} \tag{5.15}
$$

When $\xi\nu \geqq 0$, $|\xi| - \xi\nu \leqq |\xi| \leqq |\xi| + \xi\nu$; therefore we can replace (5.15) by

$$\frac{1}{2}\rho^2 q(3m+8q-7)(|\xi|(|\xi|-\xi\nu))^{q-1} + \left(q\frac{\mathscr{C}'}{c_0} + K_1\right)(|\xi|(|\xi|-\xi\nu))^q$$

$$+6q(2\mu_0+\mu_1)(|\xi|(|\xi|-\xi\nu))^{q-1/2} - \frac{q^2}{c_0'}(|\xi|(|\xi|-\xi\nu))^{2q-1/2} \tag{5.16}$$

$$+\left(s_0\mathscr{C}' + \frac{\mathscr{C}'^2}{c_0} - K_2\right)\xi\nu - \mathscr{C}'\left(\frac{\mathscr{C}}{c_0'} - \mu_0 - \mu_1\right) + K \leqq 0,$$

which is satisfied if (5.7) and

$$C_1 + \mathscr{C}'(\mu_0+\mu_1) + K \leqq \frac{\mathscr{C}'}{c_0'}\mathscr{C}, \tag{5.17}$$

C_1 being the supremum of the function

$$t \mapsto -\frac{q^2}{c_0'}t^{2q-1/2} + \left(q\frac{\mathscr{C}'}{c_0} + K_1\right)t^q + 6q(2\mu_0+\mu_1)t^{q-1/2} + \tfrac{1}{2}\rho^2 q(3m+8q-7)t^{q-1}.$$

This condition (5.17) implies $\mathscr{C} > c_0'(\mu_0+\mu_1)$ prescribed above. When $\xi\nu \leqq 0$, $|\xi| \leqq |\xi| - \xi\nu \leqq 2|\xi|$; in this case, we can replace (5.15) by

$$\frac{1}{2}\rho^2 q(3m+8q-7)2^{q-1}|\xi|^{2q-2} + \left|s_0\mathscr{C}' + \frac{\mathscr{C}'^2}{c_0'} - K_2\right||\xi|$$

$$-\frac{q}{c_0'}|\xi|^{4q-1} + 6q(2\mu_0+\mu_1)2^{q-1/2}|\xi|^{2q-1} + K_1 2^q|\xi|^{2q} \tag{5.18}$$

$$\leqq \mathscr{C}'\left(\frac{\mathscr{C}}{c_0'} - \mu_0 - \mu_1\right) - K,$$

which is satisfied if

$$C_2 + \mathscr{C}'(\mu_0+\mu_1) + K \leqq \frac{\mathscr{C}'}{c_0'}\mathscr{C}, \tag{5.19}$$

C_2 being the supremum of the left-hand side of (5.18).

In short, (5.11) is a consequence of (5.7), (5.17) and (5.19).

2) Now we shall prove (5.9), which can be written as follows:

$$v_{\nu,s,r}(\xi) = \nu u_{s,r}(\xi) \geqq V_\nu(\xi) \quad \text{for each } \xi \in \overline{B_r} \text{ and } \nu \in S_1. \tag{5.9}$$

Let us consider, as in the proof of Lemma 3, the application $(\nu, \xi) \mapsto w_\nu(\xi) = v_{\nu,s,r}(\xi) - V_{\nu,s}(\xi)$, which is continuous on the compact $S_1 \times \overline{B_r}$ (s and r are fixed); let (ν_0, ξ_0) be a point where the infimum of this function is reached. If $w_{\nu_0}(\xi_0) \geqq 0$, then $w_\nu(\xi) \geqq 0$ for each $(\nu, \xi) \in S_1 \times \overline{B_r}$. If not, $v_{\nu_0,s,r}(\xi_0) < V_{\nu_0}(\xi_0)$ but we shall see that it is impossible:

The extremality condition of first order concerning the variations of $\nu \in S_1$ can be written

$$\nabla u_{s,r}(\xi_0) = v_{\nu_0,s,r}(\xi_0)\nu_0 + (q(|\xi_0| - \xi_0\nu_0)^{q-1}|\xi_0|^q + \mathscr{C}')\xi'_{0_{\nu_0}}. \tag{5.20}$$

a) If $\xi_0 \in B_r$, we obtain, as in Lemma 3, from the derived equation of (2.2) satisfied by $v_{\nu_0,s,r}$, from (5.11), (5.20), and from the next extremality conditions (since $\xi_0 \in B_r$)

$$\nabla v_{\nu_0,s,r}(\xi_0) = \nabla V_{\nu_0}(\xi_0) \quad \text{and} \quad \Delta w_{\nu_0}(\xi_0) \geqq 0, \tag{5.21}$$

after calculation and denoting (ν, ξ) in place of (ν_0, ξ_0):

$$
\begin{aligned}
sw_\nu(\xi) \geqq\ & q\mu_0(|\xi| - \xi\nu)^{q-1}|\xi|^{q-2}(2|\xi| - \xi\nu)|\xi'_\nu| \\
& + (q|\xi|^{q-2}(|\xi| - \xi\nu)^{q+1} + \mathscr{C}') \\
& \cdot \left(\frac{\partial h}{\partial \nu}(-v_{\nu,s,r}(\xi)\nu - (q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')\xi'_\nu) \right. \\
& \left. - \frac{\partial h}{\partial \nu}(-V_\nu(\xi)\nu - (q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')\xi'_\nu) \right) \\
& - q|\xi|^{q-2}(|\xi| - \xi\nu)^{q-1}(2|\xi| - \xi\nu)|\xi'_\nu| \\
& \cdot \left(\frac{\partial h}{\partial \xi''_\nu}(-v_{\nu,s,r}(\xi)\nu - (q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')\xi'_\nu) \right. \\
& \left. - \frac{\partial h}{\partial \xi''_\nu}(-V_\nu(\xi)\nu - (q(|\xi| - \xi\nu)^{q-1}|\xi|^q + \mathscr{C}')\xi'_\nu) \right).
\end{aligned} \tag{5.22}
$$

Therefore, because of $-v_{\nu_0,s,r}(\xi_0) \geqq -V_{\nu_0}(\xi_0)$ and (2.7), $sw_{\nu_0}(\xi_0) \geqq 0$, which is impossible.

b) If $\xi_0 \in S_r$, it follows from (5.6) and (5.20) that the expression of $w_{\nu_0}(\xi_0)\xi_0\nu_0$ is the same as that of Lemma 3, given by the formula (4.28) (replacing $\mathscr{C}_s$ by $\mathscr{C}$, and (ν, ξ) by (ν_0, ξ_0)). Therefore,

α) If $\xi_0\nu_0 \neq 0$, we obtain, as in Lemma 3, that $w_{\nu_0}(\xi_0) \geqq 0$, which is impossible.

β) The hypothesis $\xi_0\nu_0 = 0$ leads as in Lemma 3 (it suffices to replace the right-hand side of (4.32) by $qr^{2q} + \mathscr{C}'r$) to $v_{\nu_0}(\xi_0) \geqq 0$, which is impossible.

B. *A priori estimates in the whole space.* As in part B of the proof of Lemma 3, for each $\beta \in]0, 1[$ and $s \in]0, s_0]$, there exists a solution of (2.2) in $C^{2,\beta}(\mathbb{R}^m; \mathbb{R}_+)$ satisfying the same inequalities as $u_{s,r}$. It is clear, because of the uniqueness property of Theorem 1, that this solution is the optimal cost u_s of $(P_{x,s})$ in Λ_1. Let us write the estimates of u_s and ∇u_s:

$$
\inf g - h(0) \leqq su_s(\xi) \leqq s_0|\xi|^{q_1} + D, \tag{5.23}
$$

$$
\frac{\partial u_s}{\partial \nu}(\xi) \geqq -\mathscr{C} - (|\xi| - \xi\nu)^q|\xi|^q + \mathscr{C}'\xi\nu \tag{5.24}
$$

for each $s \in]0, s_0]$, $\xi \in \mathbb{R}^m$ and $\nu \in S_1$.

C. *Asymptotic estimates.* These last estimates are independent of $s \in]0, s_0]$. We will now show that, for each $R > 0$, there exists a positive constant C, independent of $s \in]0, s_0]$, such that

$$
|(u_s - u_s(0))/\overline{B_R}|_{2,\beta,B_R} \leqq C. \tag{5.25}
$$

Because of (5.24), we have

$$
\sup_{s \in]0,s_0]} \left(\sup_{\xi \in B_{2R}} |\nabla u_s(\xi)| \right) \leqq M_1
$$

for some positive constant M_1; hence, we have, successively

$$
\sup_{s \in]0,s_0]} \left(\sup_{\xi \in B_{2R}} |u_s(\xi) - u_s(0)| \right) \leqq M
$$

for some positive constant M, and $|g + u_s - u_s(0) - su_s|_{0,\beta,B_{2R}} \leqq M_2$ for some positive

constant M_2. Therefore, rewriting (2.2) in the form

$$\begin{aligned}(5.26)\qquad &-\tfrac{1}{2}\rho^2\Delta(u_s(\xi)-u_s(0))+(u_s(\xi)-u_s(0))\\ &+h(-\nabla(u_s(\xi)-u_s(0))=g(\xi)+(u_s(\xi)-u_s(0))-su_s(\xi)\end{aligned}$$

and using the Hölder interior estimates of the gradient and the second derivatives concerning quasilinear elliptic equations of divergence form (see [18, Chap. 4, § 6]), we obtain (5.25), the constant C depending on β, ρ, s_0, R, g, but being independent of s. It is then possible, using the Arzela-Ascoli theorem, to construct by recurrence a sequence $(\sigma_n)_{n\in\mathbb{N}}$ in $]0, s_0]$, converging to 0, and a function $v_0\in C^{2,\beta}(\mathbb{R}^m)$ such that

$$(5.27)\qquad v_0/\overline{B_R}=\lim_{n\uparrow+\infty}(u_{\sigma_n}-u_{\sigma_n}(0))/\overline{B_R}\quad\text{in } C^{2,\beta}(\overline{B_R})$$

for each $R>0$. The sequence $(\sigma_n u_{\sigma_n}(0))_{n\in\mathbb{N}}$ being bounded, there exist a subsequence $(s_n)_{n\in\mathbb{N}}$ of $(\sigma_n)_{n\in\mathbb{N}}$ and $\lambda_0\in\mathbb{R}_+$ such that

$$(5.28)\qquad \lambda_0=\lim_{n\uparrow+\infty}s_nu_{s_n}(0).$$

Since $\lim_{n\uparrow+\infty}(s_n\nabla u_{s_n})/\overline{B_R}=0$ in $C^0(\overline{B_R})$, $(s_nu_{s_n})/\overline{B_R}$ converges to λ_0 in $C^1(\overline{B_R})$. Finally, (λ_0, v_0) is a solution of (3.5). For the regularity property $v_0\in C^3(\mathbb{R}^m;\mathbb{R})$, see [12].

D. *The radiative condition.* From (5.24), we have (with the method used in Lemma 3)

$$(5.29)\qquad \nu\nabla h(-\nabla u_s(\lambda\nu))\leqq\mu_0+\mu_1+\frac{\mathscr{C}}{c_0}-\frac{\mathscr{C}'}{c_0'}\lambda.$$

(3.6) follows immediately.

5d. We are now in position to deduce Theorem 2 from Lemmas 4, 5, 6: let (λ_0, v_0) be a solution of (3.5) satisfying (3.6); ∇v_0 has polynomial growth, and hence v_0 too; therefore

$$(5.30)\qquad \lim_{t\uparrow+\infty}\frac{1}{t}E(v_0(\xi_{x,p}(t)))=0;$$

Lemma 4 completes the proof.

6. Appendix: the Neumann problem in B_r.

LEMMA 7. *For some $\alpha\in]0,1[$, the Neumann problem* (2.2) *in B_r and*

$$(6.1)\qquad \frac{\partial u}{\partial\nu}(r\nu)=C\qquad (C\text{ being a constant})$$

has a unique $C^{2,\alpha}(\bar{B}_r)$ solution.

Proof. The uniqueness proceeds directly from the maximum principle. The existence property is reduced to the establishment of certain a priori estimates for possible solutions. This reduction is achieved through the application of a topological fixed point theorem in an appropriate function space:

a) The fixed point theorem: Let T be a compact mapping of a Banach space $\mathscr{B}$ into itself, and suppose there exists a constant M such that

$$(6.2)\qquad \|x\|_{\mathscr{B}}\leqq M\quad\text{for all } x\in\mathscr{B}\text{ and }\sigma\in[0,1]\text{ satisfying } x=\sigma T(x).$$

Then T has a fixed point (see [26] or [12, Thm. 10.3]).

b) For all $v \in C^{1,\alpha}(\bar{B}_r)$, the operator T is defined by letting $u = T(v)$ be the unique solution in $C^{2,\alpha}(\bar{B}_r)$ of the linear Neumann problem (6.1),

$$-\tfrac{1}{2}\rho^2 \Delta u(x) + su(x) = g(x) - h(-\nabla v(x)) \quad \text{for each } x \in B_r. \tag{6.3}$$

The unique solvability of this problem is guaranteed by the linear existence and uniqueness result (see [18, Chap. 3, Thm. 3.2] or [12, Thm. 6.31 and remarks on p. 124]) using the Fredholm alternative.

The solvability of (6.1), (2.2) in $C^{2,\alpha}(\bar{B}_r)$ is thus equivalent to the solvability of the equation $u = T(u)$ in the Banach space $\mathcal{B} = C^{1,\alpha}(\bar{B}_r)$.

c) We will now show that the operator T is continuous and compact: By virtue of the global Schauder estimate about the Neumann problem for linear elliptic equations (see [18, Chap. 3, Thm. 3.1], or [12, Thm. 6.30]), T maps bounded sets in $C^{1,\alpha}(\bar{B}_r)$ into bounded sets in $C^{2,\alpha}(\bar{B}_r)$ which (by the Arzela-Ascoli theorem) are relatively compact in $C^2(\bar{B}_r)$ and $C^{1,\alpha}(\bar{B}_r)$. In order to show the continuity of T, we let $(v_n)_{n\geqq 1}$, converge to v in $C^{1,\alpha}(\bar{B}_r)$; $\{T(v_n)\}$ is relatively compact in $C^2(\bar{B}_r)$; let $(T(v_{n_p}))_{p\geqq 1}$ be a convergent subsequence of $(T(v_n))_{n\geqq 1}$ with limit $u \in C^2(\bar{B}_r)$; let $w_p = T(v_{n_p})$. Then since

$$\begin{aligned} &-\Delta u(x) + su(x) + h(-\nabla v(x)) - g(x) \\ &\quad = \lim_{p\to\infty} (-\Delta w_p(x) + sw_p(x) + h(-\nabla v_{n_p}(x)) - g(x)) = 0, \end{aligned}$$

we must have $u = T(v)$, and hence the sequence $(T(v_n))_{n\geqq 1}$ itself converges to u.

d) It only remains to show that, for some $\alpha \in]0, 1[$, there exists a constant M such that

$$|u|_{1,\alpha,B_r} \leqq M \quad \text{for all } u \in C^{1,\alpha}(\bar{B}_r) \text{ and } \sigma \in [0, 1] \text{ satisfying} \tag{6.4}$$

$$-\Delta u(x) + su(x) + \sigma h(-\nabla u(x)) = \sigma g(x) \text{ and (6.1):} \tag{6.5}$$

We establish first that (6.1) and (6.5) implies

$$0 \leqq u(\xi) \leqq \frac{1}{s}(\sup g - h(0)) \quad \text{for each } \xi \in \bar{B}_r \tag{6.6}$$

(the constants 0 and $(1/s)(\sup g - h(0))$ are respectively a subsolution and a supersolution of (6.5); then the application of the comparison principle gives (6.6)). By virtue of the global estimate about the solutions of the Neumann problem for quasilinear elliptic equations of divergence form (see [18, Chap. 10, Thm. 2.1]), it follows from (6.6) that $|u|_{1,\alpha,B_r} \leqq M$ for some constants $\alpha \in]0, 1[$ and $M \geqq 0$ independent of σ.

Remarks. a) Because of $g \in C^2(\mathbb{R}^m; \mathbb{R}_+)$, we have

$$u \in C^{2,\alpha}(\bar{B}_r) \cap C^{3,\alpha}(B_r)$$

(see for example [12, Thm. 6.17]).

b) Lemma 7 can be proved by the regularity of weak solutions; moreover, the solution of the Neumann problem admits a stochastic interpretation similar to that of Theorem 1, for a reflected diffusion. The same remark applies to the Neumann problem associated to (3.5): see the works of J. M. Lasry and P. L. Lions.

Acknowledgments. The author is grateful to Professor J. M. Lasry for useful discussions.

REFERENCES

[1] J. P. AUBIN AND F. H. CLARKE, *Shadow prices and duality for a class of optimal control problems*, this Journal, 17 (1979), pp. 567-586.

[2] V. E. BENĚS AND I. KARATZAS, *Optimal stationary linear control of the Wiener process*, J. Optim. Theory Apl., 35 (1981), pp. 611-633.

[3] A. BENSOUSSAN, E. G. HURST, JR. AND B. NASLUND, *Management Applications of Modern Control Theory*, North-Holland, Amsterdam, 1974.

[4] A. BENSOUSSAN AND J. L. LIONS, *Applications of Variational Inequalities in Stochastic Control*, North-Holland, Amsterdam, 1982.

[5] ———, *Contrôle impulsionnel et inéquations quasi variationnelles*, Dunod, Paris, 1982.

[6] D. BLACKWELL, *Discrete dynamic programming*, Ann. Math. Stat., 33 (1962), pp. 719-726.

[7] I. EKELAND AND J. A. SCHEINKMAN, *Transversality conditions for some infinite horizon discrete time optimization problems*, to appear.

[8] W. H. FLEMING, *Optimal continuous-parameter stochastic control*, SIAM Rev., 11 (1969), pp. 470-509.

[9] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[10] A. FRIEDMAN, *Stochastic Differential Equations and Applications*, Vol. 1 and 2, Academic Press, New York, 1975.

[11] I. I. GIKHMAN AND A. V. SKOROKHOD, *Stochastic Differential Equations*, Springer-Verlag, New York, 1972.

[12] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, New York, 1977.

[13] R. HOWARD, *Dynamic Programming and Markov Processes*, Technology Press and John Wiley, New York, 1960.

[14] H. J. KUSHNER, *Optimality conditions for the average cost per unit time problem with a diffusion model*, this Journal, 16 (1978), pp. 330-346.

[15] ———, *Introduction to Stochastic Control*, Holt, Rinehart and Winston, New York, 1971.

[16] ———, *Optimal discounted stochastic control for diffusion processes*, this Journal, 5 (1967), pp. 520-531.

[17] G. S. LADDE, V. LAKSHMIKANTHAM AND P. T. LIU, *Differential inequalities and Ito type stochastic differential equations*, Proc. Equations différentielles et fonctionnelles non linéaires, P. Janssens, J. Mawhin and N. Rouche, eds., Hermann, Paris, 1973, pp. 611-640.

[18] O. A. LADYZHENSKAYA AND N. N. URAL'CEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[19] J. M. LASRY, *Evolution of problems of stochastic control when the discount vanishes*, Cahier de mathématiques de la décision n° 7519, Univ. Paris-Dauphine, 1975.

[20] ———, *Thèse*, Univ. Paris-Dauphine.

[21] ———, *Contrôle stationnaire asymptotique*, Proc. Congrès de Contrôle Optimal, IRIA, 1974, pp. 296-313, in Lecture Notes in Economics and Mathematical Systems 107, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, New York.

[22] P. L. LIONS, *Thèse de 3ème cycle*, Univ. Paris VI, 1978.

[23] ———, *Control of diffusion processes in* $\mathbb{R}^n$, Cahier de mathématiques de la décision n° 7936, Univ. Paris-Dauphine, 1979.

[24] ———, *Résolution de problèmes elliptiques quasilinéaires*, Arch. Rational Mech. Anal., 74 (1980), pp. 335-353.

[25] P. MANDL, *Analytic Treatment of One Dimensional Markov Processes*, Springer-Verlag, New York, 1968.

[26] H. SCHAEFER, *Über die Methode der a priori Schranken*, Math. Ann., 129 (1955), pp. 415-416.

[27] R. TARRES, *Asymptotic evolution of a stochastic control problem when the discount vanishes*, Journées sur l'analyse des systèmes, septembre 1978, Univ. Bordeaux 1, Astérisque (Société mathématiques de France) 75-76 (1980), pp. 227-237.

[28] ———, *Optimal non-explosive control of a non constrained diffusion and behaviour when the discount vanishes*, Proc. Workshop on Stochastic Control and Stochastic Differential Systems, Univ. of Bonn, January 1979, Lecture Notes in Control and Information Sciences 16, M. Kohlmann and W. Vogel, eds., Springer-Verlag, Berlin, 1979, pp. 588-597.

[29] ———, *Comportement asymptotique d'un problème de contrôle stochastique*, Cahier de mathématiques de la décision n° 8215, Univ. Paris-Dauphine, 1982.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 4, July 1985

009

NONDIFFERENTIABLE OPTIMIZATION PROBLEMS FOR ELLIPTIC SYSTEMS*

ANDRZEJ MYŚLIŃSKI† AND JAN SOKOŁOWSKI†

Abstract. The paper is devoted to analysis of optimization problems in coefficients of fourth order elliptic boundary value problems. Similar problems were investigated in the framework of shape optimal design of thin plates. Since in general such problems have no optimal solution, G-convergence theory of elliptic operators is exploited in order to define and to characterize generalized optimal solutions. Necessary optimality conditions for nonsmooth optimization problems are derived. Results of computations for two examples are presented.

Key words. distributed parameter system, control in coefficients, nondifferentiable optimization, generalized optimal solution, G-convergence, optimum design problem

1. Introduction. The paper is devoted to the analysis of some nondifferentiable optimization problems in coefficients of fourth order elliptic systems. Since optimization problems in coefficients of partial differential equations in general have no optimal solutions [15], [23], a family of auxiliary regularized nondifferentiable optimization problems depending on a parameter $\varepsilon \in (0, \delta]$, $\delta > 0$ is introduced. For every fixed $\varepsilon \in (0, \delta]$ existence of an optimal solution to the regularized problem is assured and necessary optimality conditions for this problem are obtained.

In order to pass to the limit with $\varepsilon \downarrow 0$ the notion of so-called G-convergence [17] is introduced. G-convergence allows us to define and characterize a generalized optimal solution of the initial optimization problem. In the paper the method described above is applied to two optimal design problems of a clamped plate of variable thickness $u(x) \geqq c_1 > 0$, $x \in \Omega \subset R^2$ shown in Fig. 1.

In the first problem (P1), formulated in § 2, we are looking for a plate with fixed volume, variable thickness bounded from below and above and with required static properties. Maximal deflection of a loaded plate is minimized over a set U_{ad} of admissible elements $u(\cdot)$. In the second problem (P2), formulated in § 3, the smallest eigenvalue of an appropriate eigenvalue problem is maximized over the set U_{ad}. Problems (P1) and (P2) have already been studied by many authors in the framework of optimal design [2], [6], [7], [10], [16], [18], [19], [20], [22], [28], [29], [30]. In [19] an anisotropic model of axisymmetric plate is proposed. Using this model numerical results for an optimization problem similar to our problem (P2) were obtained in [20]. However this approach in general does not assure the existence of an optimal solution. Problems (P1) and (P2) involve nonsmooth cost functionals. Differentiability properties of some nonsmooth functionals were investigated in an abstract setting by J. P. Zolesio [26]. In particular he obtained [27] the form of directional derivative[1] of the multiple eigenvalue of a fourth order elliptic eigenvalue problem with respect to the variations of coefficients of an elliptic operator. The method of J. P. Zolesio is applied in our paper in order to obtain directional differentiability of nonsmooth cost functionals. Lemarechal's method [11] of nonsmooth optimization combined with the finite element method was used for computations. It is confirmed by our numerical results that from the numerical viewpoint problem (P2) is more complex than problem (P1).

* Received by the editors January 26, 1983, and in revised form July 30, 1984.

† Systems Research Institute of the Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland.

[1] See, however, Rousselet [29] for a result on directional derivatives of eigenvalues for shape sensitivity, and Haug-Rousselet [28] for coefficient sensitivity.

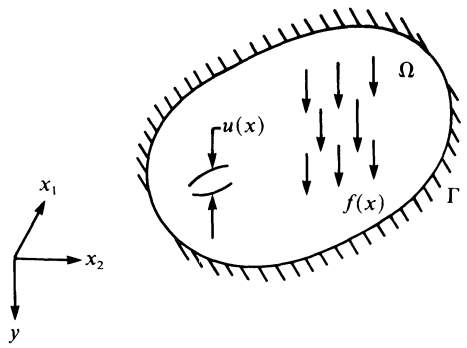


FIG. 1. *Clamped plate of variable thickness.*

The outline of the paper is the following. In the second part of this section notation is introduced, a linear isotropic model of the clamped plate is described and differentiability properties of some nonsmooth functionals are given. In § 2 a method of regularization for problem (P1) is introduced, necessary optimality conditions for a family of regularized problems $(P1)_\varepsilon$, $\varepsilon \in (0, \delta]$, $\delta > 0$, are presented and the notion of a generalized solution to problem (P1) is proposed. Section 3 describes similar results obtained for problem (P2). In § 4 numerical results are presented.

Let us introduce the following notation. Let Ω be an open, bounded and connected set in R^2 which satisfies the cone condition [1] with boundary $\partial\Omega \in C^0$ manifold of dimension 1. $L^2(\Omega)$ is the Hilbert space of (equivalence classes of) Lebesgue square-summable real-valued functions with the scalar product

$$(y, z)_{L^2(\Omega)} = \int_\Omega y(x)z(x)\, dx.$$

$L^\infty(\Omega)$ is the Banach space of (equivalence classes of) essentially bounded functions with the norm:

$$\|y\|_{L^\infty(\Omega)} = \operatorname{ess\,sup} \{|y(x)| \mid x \in \Omega\}.$$

$H^1(\Omega)$ is the Sobolev space [1] with the norm

$$\|y\|_{H^1(\Omega)} = \left(\|y\|^2_{L^2(\Omega)} + \sum_{i=1}^{2} \left\| \frac{\partial y}{\partial x_i} \right\|^2_{L^2(\Omega)} \right)^{1/2}.$$

$H_0^2(\Omega)$ is the Sobolev space, the closure of the space C_0^∞ in the norm

$$\|y\|_{H_0^2(\Omega)} = \left(\sum_{i,j=1}^{2} \left\| \frac{\partial^2 y}{\partial x_i\, \partial x_j} \right\|^2_{L^2(\Omega)} \right)^{1/2}.$$

We denote by $H^{-2}(\Omega)$ the dual space to $H_0^2(\Omega)$ and by $\langle \cdot, \cdot \rangle$ duality pairing between $H^{-2}(\Omega)$ and $H_0^2(\Omega)$.

Let us describe briefly a linear model of the clamped plate [4]. We assume that tensor $\{a_{ijkl}\}$, $a_{ijkl} \in L^\infty(\Omega)$, $i, j, k, l = 1, 2$ which characterizes the rigidity of the plate [4] has the form

$$a_{ijkl}(x) = u^3(x) b_{ijkl}, \qquad x \in \Omega, \quad i, j, k, l = 1, 2 \tag{1.1}$$

where b_{ijkl}, $i, j, k, l = 1, 2$ are given constants which depend on the material of the plate. Denote $b_{\alpha\beta} = b_{ijkl}$, $\alpha = \{i, j\}$, $\beta = \{k, l\}$, $i, j, k, l = 1, 2$ and assume that 4×4 matrix $[b_{\alpha\beta}]$

possesses inverse $[b_{\alpha\beta}]^{-1}$. Furthermore assume that the following symmetry conditions are satisfied:

$$b_{ijkl} = b_{jikl} = b_{klij}, \qquad i, j, k, l = 1, 2$$

and that there exists a constant $\upsilon > 0$ such that:

$$\sum_{i,j=1}^{2} \sum_{k,l=1}^{2} b_{ijkl} e_{ij} e_{kl} \geqq \upsilon \sum_{i,j=1}^{2} e_{ij}^2 \tag{1.2}$$

for every symmetric 2×2 matrix $[e_{ij}]$. Assume that the plate can be of variable bounded thickness $u(x)$, $x \in \Omega$ but of a constant volume i.e.: it can be treated as an element of the set

$$U_{\text{ad}} = \left\{ u \in L^{\infty}(\Omega) \mid 0 < c_1 \leqq u(x) \leqq c_2 \text{ for a.e. } x \in \Omega, \int_{\Omega} u(x)\, dx = c \right\} \tag{1.3}$$

where c_1, c_2, c are given constants such that $U_{\text{ad}} \neq \emptyset$.

Suppose that the plate is loaded by a given perpendicular force $f = f(x)$, $x \in \Omega$ and assume $f \in H^{-2}(\Omega)$.

Denote by $y = y(u; x)$, $u \in U_{\text{ad}}$, $x \in \Omega$ the deflection of the plate which can be described by the elliptic boundary value problem of the form:

$$\begin{aligned} &y = y(u, \cdot) \in H_0^2(\Omega), \\ &a_u(y, z) = \langle f, z \rangle \quad \forall z \in H_0^2(\Omega) \end{aligned} \tag{1.4}$$

where

$$a_u(y, z) = \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_{\Omega} u^3(x) b_{ijkl} \frac{\partial^2 y}{\partial x_i\, \partial x_j}(x) \frac{\partial^2 z}{\partial x_k\, \partial x_l}(x)\, dx \quad \forall y, z \in H_0^2(\Omega). \tag{1.5}$$

It is well known [4] that under our assumptions for any fixed $u \in U_{\text{ad}}$ and $f \in H^{-2}(\Omega)$ there exists a unique weak solution to the equation (1.4) such that $y(u; \cdot) \in C(\bar{\Omega})$.

Let us recall two theorems due to J. P. Zolesio [27] concerning directional differentiability of nonsmooth functionals which we shall need in the sequel. Let $E \subset L^{\infty}(\Omega)$ be an open set, W be a compact topological space and let there be given the mapping:

$$F(\cdot, \cdot)\colon W \times E \mapsto R.$$

THEOREM 1. *Assume*

(i) $F(\cdot, \cdot)$ *is upper semicontinuous on* $E \times W$;

(ii) $F(w, \cdot)$ *is continuous on* E *for every element* $w \in W$;

(iii) $F(w, \cdot)$ *is differentiable on* E *for every element* $w \in W$, *i.e., there exists the limit*

$$dF(w, u; v) = \lim_{t \downarrow 0} (F(w, u + tv) - F(w, u))/t$$

such that the mapping:

$$L^{\infty}(\Omega) \ni v \mapsto dF(w, u; v) \in R$$

is linear and continuous for every pair $(w, u) \in W \times E$;

(iv) $dF(\cdot, \cdot; \cdot)$ *is upper semicontinuous on* $W \times E \subset L^{\infty}(\Omega)$.

Then the functional

$$f(u) = \sup \{F(w, u) \mid w \in W\}$$

is differentiable in every direction $v \in L^\infty(\Omega)$ *and*

$$df(u; v) = \sup \{dF(w, u; v) \,|\, w \in W, f(u) = F(w, u)\}.$$

THEOREM 2. *Assume that the conditions* (ii), (iii) *of Theorem* 1 *are satisfied and suppose*

(v) $F(\cdot, \cdot)$ *is lower semicontinuous on* $W \times E$.

Define the functional

$$g(u) = \inf \{F(w, u) | w \in W\}$$

and assume that $dF(\cdot, \cdot; \cdot)$ *has the following property*:

(vi) *for every sequence* $\{w_n\} \subset W$, $\{u_n\} \subset E$, $\{v_n\} \subset L^\infty(\Omega)$ *such that*

$$w_n \in M(u_n) = \{w \in W \,|\, g(u_n) = F(w, u_n)\},$$

$$w_n \to w \quad \text{in } W,$$

$$u_n \to u \quad \text{in } L^\infty(\Omega),$$

$$v_n \to v \quad \text{in } L^\infty(\Omega).$$

It follows that

$$\lim_{n \to \infty} dF(w_n, u_n; v_n) \geqq dF(w, u; v).$$

Then the functional $g(\cdot)$ *is differentiable in every direction* $v \in L^\infty(\Omega)$ *and*

$$dg(u; v) = \inf \{dF(w, u; v) \,|\, w \in M(u)\}.$$

2. Minimization of maximal deflection of the clamped plate. Consider the following optimal design problem:

$$\text{(P1)} \qquad \inf \left\{ \sup_{x \in \Omega} |y(u; x)| \,|\, u \in U_{\text{ad}} \right\}$$

where $y(u; \cdot) \in H_0^2(\Omega) \subset C(\bar{\Omega})$, $u \in U_{\text{ad}}$, is a unique solution to (1.4). Problem (P1) has no optimal solution i.e.: in general there is no element $u \in U_{\text{ad}}$ at which functional

$$J(u) = \sup_{x \in \Omega} |y(u; x)| \tag{2.1}$$

attains minimal value.

In order to assure existence of an optimal solution we define a regularized problem in the following way. Let $\varepsilon > 0$ be a parameter, $\varepsilon \in (0, \delta]$, $\delta > 0$. Consider the following family of regularized problems:

$$\text{(P1)}_\varepsilon \qquad \inf \left(J(u) + \frac{\varepsilon}{2} \|u\|^2_{H^1(\Omega)} \,|\, u \in U_{\text{ad}} \cap H^1(\Omega) \right).$$

We prove existence of an optimal solution $u_\varepsilon \in U_{\text{ad}} \cap H^1(\Omega)$ to the problem $(\text{P1})_\varepsilon$ for every $\varepsilon \in (0, \delta]$. We obtain the form of necessary conditions of optimality which are satisfied by u_ε. Furthermore we investigate existence of a limit of the sequence $\{u_\varepsilon\}$ for $\varepsilon \downarrow 0$.

THEOREM 3. *For* $\varepsilon \in (0, \delta]$ *there exists an optimal solution* $u_\varepsilon \in U_{\text{ad}} \cap H^1(\Omega)$ *to* $(\text{P1})_\varepsilon$ *such that the following necessary condition of optimality is satisfied*:

$$\max_{s \in \Omega^*(u_\varepsilon)} \int_\Omega g(s, x)(u(x) - u_\varepsilon(x))\, dx + \varepsilon (u_\varepsilon, u - u_\varepsilon)_{H^1(\Omega)} \geqq 0 \quad \forall u \in U_{\text{ad}} \cap H^1(\Omega), \tag{2.2}$$

where

$$\Omega^*(u) = \{x \in \Omega \,|\, J(u) = y(u; x)\}$$

and the element $g(s, \cdot) \in L^1(\Omega)$ *is defined by*

$$g(s, x) = 3u_\varepsilon^2(x) \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} b_{ijkl} \frac{\partial^2 y_\varepsilon}{\partial x_i \, \partial x_j}(x) \frac{\partial^2 p}{\partial x_k \, \partial x_l}(s, x) \tag{2.3}$$

and elements y_ε, $p(s, \cdot) = p(s) \in H_0^2(\Omega)$ *satisfy the equations*

$$a_{u_\varepsilon}(y_\varepsilon, \theta) = \langle f, \theta \rangle \quad \forall \theta \in H_0^2(\Omega), \tag{2.4}$$

$$a_{u_\varepsilon}(p(s), \theta) = \langle \operatorname{sign} y(u; s) \delta(s) \, \theta \rangle \quad \forall \theta \in H_0^2(\Omega) \tag{2.5}$$

where $\delta(s)$ *is the Dirac functional evaluated at the point* $s \in \Omega$.

For the proof of Theorem 3 we need the following lemmas.

LEMMA 1. *Let there be given a sequence* $\{u_n\} \subset U_{\text{ad}}$ *and an element* $u \in L^\infty(\Omega)$ *such that*

$$u_n(x) \to u(x) \quad \text{a.e. in } \Omega, \text{ for } n \to \infty. \tag{2.6}$$

Then

$$\lim_{n \to \infty} J(u_n) = J(u). \tag{2.7}$$

The proof of Lemma 1 is omitted as a standard one.

Define a set $E \subset L^\infty(\Omega)$ of the form

$$E = \{u \in L^\infty(\Omega) \,|\, 0 < \tilde{c}_1 < u(x) < \tilde{c}_2 \text{ a.e. in } \Omega\},$$

where $\tilde{c}_1$, $\tilde{c}_2$ are given constants such that $0 < \tilde{c}_1 < c_1 < c_2 < \tilde{c}_2$.

LEMMA 2. *The directional differential of the function* $J(u)$ *at* $u \in E$ *has the form*

$$dJ(u; h) = \max_{s \in \Omega^*(u)} \int_\Omega g(s, x) h(x) \, ds, \qquad h \in L^\infty(\Omega), \tag{2.8}$$

where for fixed $s \in \bar{\Omega}$ *the element* $g(s, \cdot) \in L^1(\Omega)$ *has the form*

$$g(s, x) = 3u^2(x) \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \frac{\partial^2 y}{\partial x_i \, \partial x_j} y(u; x) \frac{\partial^2 p}{\partial x_k \, \partial x_l}(s, x), \qquad x \in \Omega. \tag{2.9}$$

Proof. We shall apply Theorem 1. To do that we have to verify all assumptions of this theorem.

Denote

$$F(x, u) \equiv y(u; x), \qquad u \in E \subset L^\infty(\Omega), \quad x \in W \equiv \bar{\Omega} \subset R^2$$

where $y(u, \cdot) \in H_0^2(\Omega) \subset C(\bar{\Omega})$, $u \in E$ is a unique solution to (1.4). It can be shown that there exists a constant: $L < \infty$ such that

$$\|y(u_1; \cdot) - y(u_2; \cdot)\|_{C(\bar{\Omega})} \leqq L \|u_1 - u_2\|_{L^\infty(\Omega)} \quad \forall u_1, u_2 \in E; \tag{2.10}$$

hence assumptions (i), (ii) of Theorem 1 are satisfied. In order to verify assumption (iii) let us note that by application of implicit function theorem it follows that for fixed $s \in \bar{\Omega}$, $y(\cdot\,; s)$ is differentiable on the set $E \subset L^\infty(\Omega)$. Directional differential $dy(u, s; v)$, $u \in E$, $s \in \bar{\Omega}$, $v \in L^\infty(\Omega)$ has the form:

$$dy(u, s; v) = \langle \delta(s), z \rangle \tag{2.11}$$

where the element $z \in H_0^2(\Omega)$ is a unique solution to the elliptic equation:

$$(2.12)\quad a_u(z, \theta) = -\sum_{i,j=1}^{2}\sum_{k,l=1}^{2}\int_\Omega 3u^2(x)v(x)b_{ijkl}\frac{\partial^2 y}{\partial x_i\,\partial x_j}(u; x)\frac{\partial^2 \theta}{\partial x_k\,\partial x_l}(x)\,dx \quad \forall \theta \in H_0^2(\Omega).$$

It can be verified that $dy(\cdot, \cdot; \cdot)$ is continuous on $E \times \bar{\Omega} \times L^\infty(\Omega)$ hence the assumption (iv) of Theorem 1 is satisfied. From Theorem 1 it follows that

$$(2.13)\qquad dJ(u; v) = \max_{s \in \Omega^*(u)} \langle \operatorname{sign} y(u; s)\delta(s), z\rangle.$$

Define the adjoint state $p(s) \in H_0^2(\Omega)$, $s \in \Omega$ which satisfies the equation:

$$a_u(p(s), \phi) = \langle \operatorname{sign} y(u; s)\delta(s), \phi\rangle \quad \forall \phi \in H_0^2(\Omega).$$

Simple calculations show that (2.8), (2.9) follow from (2.13).

Proof of Theorem 3. The proof of existence of an optimal solution to the problem $(P1)_\varepsilon$, $\varepsilon \in (0, \delta]$ is classical and is omitted. The form (2.2) of necessary conditions of optimality follows from Lemma 2.

Denote by $L_u \in \mathscr{L}(H_0^2(\Omega); H^{-2}(\Omega))$, $u \in E$ the elliptic operator which is associated with the form $a_u(\cdot, \cdot)$, i.e.:

$$(2.14)\qquad \langle L_u y, z\rangle = a_u(y, z) \quad \forall y, z \in H_0^2(\Omega).$$

We define a generalized optimal solution to the problem (P1) in the form of the so-called G-limit for $\varepsilon \downarrow 0$ of matrices $\{u_\varepsilon^3[b_{\alpha\beta}]\}$ of coefficients of elliptic operators $\{L_{u_\varepsilon}\}$.

We recall the notion of G-convergence in the case of 4th order elliptic operators.

DEFINITION [17], [31]. Let $\{u_n\} \subset U_{\text{ad}}$, $n = 1, 2, \cdots$ be a given sequence. Denote by $\{u_n^3[b_{\alpha\beta}]\}$ the sequence of 4×4 matrices of coefficients of elliptic operators $\{L_{u_n}\}$. The sequence $\{u_n^3[b_{\alpha\beta}]\}$ G-converges to the matrix $[q_{\alpha\beta}]$, $q_{\alpha\beta} = q_{ijkl} \in L^\infty(\Omega)$, $i, j, k, l = 1, 2$, what we denote;

$$(2.15)\qquad u_n^3[b_{\alpha\beta}] \overset{G}{\Longrightarrow} [q_{\alpha\beta}] \quad \text{in } \Omega,$$

if for any sequence $\{y_n\} \subset H_0^2(\Omega)$ such that

$$(2.16)\qquad \begin{aligned} &y_n \to y \quad \text{weakly in } H_0^2(\Omega),\\ &L_{u_n} y_n \to \sum_{i,j=1}^{2}\sum_{k,l=1}^{2}\frac{\partial^2}{\partial x_i\,\partial x_j} q_{ijkl}\frac{\partial^2 y}{\partial x_k\,\partial x_l} \quad \text{strongly in } H^{-2}(\Omega),\end{aligned}$$

the following convergences take place:

$$(2.17)\qquad u_n^3 \sum_{i,j=1}^{2} b_{ijkl}\frac{\partial^2 y_n}{\partial x_i\,\partial x_j} \to \sum_{i,j=1}^{2} q_{ijkl}\frac{\partial^2 y}{\partial x_i\,\partial x_j} \quad \text{weakly in } L^2(\Omega), \quad k, l = 1, 2.$$

THEOREM 4. *There exist a subsequence* $\{u_\varepsilon\} \in U_{\text{ad}} \cap H^1(\Omega)$ *and a matrix of coefficients* $[q_{\alpha\beta}]$, $q_{\alpha\beta} = q_{ijkl} \in L^\infty(\Omega)$, $i, j, k, l = 1, 2$ *such that*

$$(2.18)\qquad u_\varepsilon^3[b_{\alpha\beta}] \overset{G}{\Longrightarrow} [q_{\alpha\beta}] \quad \textit{in } \Omega,$$

$$(2.19)\qquad y(u_\varepsilon; \cdot) \to \bar{y}(\cdot) \quad \textit{weakly in } H_0^2(\Omega),$$

$$(2.20)\qquad \lim_{\varepsilon \downarrow 0}\left\{J(u) + \frac{\varepsilon}{2}\|u_\varepsilon\|^2_{H^1(\Omega)}\right\} = \inf\{J(u) | u \in U_{\text{ad}}\} = \sup\{\bar{y}(x) | x \in \Omega\}$$

where $\bar{y}(\cdot) \in H_0^2(\Omega)$ is a unique weak solution to the equation

$$\sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_\Omega q_{ijkl}(x) \frac{\partial^2 \bar{y}}{\partial x_i \partial x_j}(x) \frac{\partial^2 \phi}{\partial x_k \partial x_l}(x)\, dx = \langle f, \phi \rangle \quad \forall \phi \in H_0^2(\Omega). \tag{2.21}$$

For the proof of Theorem 4 we need the following lemma:

LEMMA 3. *Let $\{u_n\} \subset U_{ad}$ be a given sequence. There exist a subsequence, still denoted $\{u_n\}$ and a matrix of coefficients $[q_{\alpha\beta}]$, $q_{\alpha\beta} = q_{ijkl} \in L^\infty(\Omega)$, $i, j, k, l = 1, 2$ such that:*

$$u_n^3[b_{\alpha\beta}] \overset{G}{\Longrightarrow} [q_{\alpha\beta}] \quad \text{in } \Omega$$

and for every $f \in H^{-2}(\Omega)$

$$L_{u_n}^{-1} f \to L^{-1} f \quad \text{weakly in } H_0^2(\Omega) \tag{2.22}$$

where the operator $L \in \mathcal{L}(H_0^2(\Omega); H^{-2}(\Omega))$ is defined in the following way:

$$\langle Ly, z \rangle = \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_\Omega q_{ijkl}(x) \frac{\partial^2 y}{\partial x_i \partial x_j}(x) \frac{\partial^2}{\partial x_k \partial x_l}(x)\, dx \qquad \forall y, z, \in H_0^2(\Omega). \tag{2.23}$$

Furthermore the following estimation holds, for a.e. $x \in \Omega$:

$$\begin{aligned} \psi_1(x) \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} b_{ijkl} e_{ij} e_{kl} &\leqq \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} q_{ijkl}(x) e_{ij} e_{kl} \\ &\leqq \psi_2(x) \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} b_{ijkl} e_{ij} e_{kl} \end{aligned} \tag{2.24}$$

for every symmetric 2×2 matrix $[e_{ij}]$, where elements $1/\psi_1(\cdot)$, $\psi_2(\cdot)$ are cluster points, in the $L^\infty(\Omega)$ weak-(*) topology of sequences $\{1/u_n^3(\cdot)\}$ and $\{u_n^3(\cdot)\}$, respectively.

The proof of Lemma 3 is given in the Appendix.

Remark. Lemma 3 is a direct extension to the higher order elliptic boundary value problems of the method of variational estimations proposed by Tartar [24] in case of second order elliptic problems.

Proof of Theorem 4. Let the element $f \in H^{-2}(\Omega)$ be fixed. For given element $u \in U_{ad}$ denote $y(u) = y(u; \cdot) \in H_0^2(\Omega)$ a unique solution to (1.4). Define functional

$$I(y) = \sup \{|y(x)| \,|\, x \in \Omega\}, \qquad y \in H_0^2(\Omega) \tag{2.25}$$

and denote by Z, $\bar{Z}$ the following subsets of the space $H_0^2(\Omega)$:

$$Z = \{y \in H_0^2(\Omega) | u \in U_{ad}, y = y(u)\},$$

$$\bar{Z} = \{y \in H_0^2(\Omega) \,|\, \{y_k\} \subset Z, k = 1, 2, \cdots, y_k \to y \text{ weakly in } H_0^2(\Omega)\}.$$

There exist sequences $\{v_k\} \subset U_{ad} \cap H^1(\Omega)$ and $\{\varepsilon_k\} \subset R^+$, $k = 1, 2, \cdots$ such that

$$\lim_{k \to \infty} J(v_k) = \inf \{J(u) \,|\, u \in U_{ad}\} \overset{\text{def}}{=} J^*,$$

$$\lim_{k \to \infty} \varepsilon_k = 0,$$

$$\lim_{k \to \infty} \varepsilon_k \|v_k\|_{H^1(\Omega)}^2 = 0.$$

Since $\|y(u)\|_{H_0^2(\Omega)} \leqq C \|f\|_{H^{-2}(\Omega)}$, for every $u \in U_{ad}$, there exists an element $\hat{y} \in \bar{Z}$ such that for a subsequence

$$y(v_k) \to \hat{y} \quad \text{weakly in } H_0^2(\Omega); \tag{2.26}$$

furthermore

$$J^* = I(\hat{y}).$$

On the other hand u_{ε_k} is the optimal solution $(P1)_{\varepsilon_k}$ and hence

$$J(v_k)+\tfrac{1}{2}\varepsilon_k\|v_k\|^2_{H^1(\Omega)} \geqq J(u_{\varepsilon_k})+\tfrac{1}{2}\varepsilon_k\|u_{\varepsilon_k}\|^2_{H^1(\Omega)} \tag{2.27}$$

whence

$$0 \leqq \varepsilon_k\|u_{\varepsilon_k}\|^2_{H^1(\Omega)} \leqq C.$$

Denote

$$r = \lim_{k\to\infty} \sup \tfrac{1}{2}\varepsilon_k\|u_{\varepsilon_k}\|^2_{H^1(\Omega)} \leqq C$$

and observe that

$$J^* \geqq I(\hat{y}) + r = \inf\{I(y) \mid y \in Z\} + r;$$

hence $r = 0$.

Define $\bar{y} = \hat{y}$ and note that by Lemma 3 it follows that (2.20), (2.21) hold for a subsequence of the sequence $\{\varepsilon_k\}$, which completes the proof.

Remark. In case of a linear model (1.4) of an isotropic plate, rigidity of the plate is characterized by the tensor $\{u^3(\cdot)b_{ijkl}\}$, $i, j, k, l = 1, 2 \in U_{\text{ad}}$.

The elliptic boundary value problem (2.21) can be formally interpreted as a model of an anisotropic plate. Rigidity of this anisotropic plate model is characterized by the tensor $\{q_{ijkl}(\cdot)\}$, $i, j, k, l = 1, 2$.

A generalized optimal solution to the problem (P1) can be defined in the form of the tensor $\{q_{ijkl}(\cdot)\}$, $i, j, k, l = 1, 2$. Several results concerning the existence of optimal solutions to optimization problems in coefficients of partial differential equations are given in [9], [13], [15], [21], [23], [25], [30].

3. Maximization of the smallest eigenvalue. Let us consider an optimal design problem [18], [7] of maximization of the smallest eigenvalue of an elliptic boundary value problem. Let $\lambda(u)$, $u \in U_{\text{ad}}$ denote the smallest eigenvalue of the following eigenvalue problem:

$$a_u(\eta, \phi) = \lambda(u)\int_\Omega u(x)\eta(x)\phi(x)\,dx \quad \forall \phi \in H_0^2(\Omega) \tag{3.1}$$

where $\eta \in H_0^2(\Omega)$, $\eta \neq 0$ is the eigenfunction corresponding to the eigenvalue $\lambda(u)$.

Consider the following optimization problem:

$$\sup\{\lambda(u) \mid u \in U_{\text{ad}}\}. \tag{P2}$$

Define for $u \in E \subset L^\infty(\Omega)$, $\eta \in H_0^2(\Omega)$, $\eta \neq 0$ the functional $F(u, \eta)$ of the form

$$F(u, \eta) = a_u(\eta, \eta) \Big/ \left(\int_\Omega u(x)\eta^2(x)\,dx\right). \tag{3.2}$$

It is well known [8] that

$$\lambda(u) = \inf\{F(u, \eta) \mid \eta \in H_0^2(\Omega), \eta \neq 0\}. \tag{3.3}$$

Denote by $M(u) \subset H_0^2(\Omega)$ the set of all normalized eigenfunctions, corresponding to the eigenvalue $\lambda(u)$:

$$M(u) = \{\eta \in H_0^2(\Omega) \mid \lambda(u) = F(u, \eta), \|\eta\|_{L^2(\Omega)} = 1\}. \tag{3.4}$$

In a similar way as in § 2 for $\varepsilon > 0$ we define the following family of regularized problems:

$$(\mathrm{P2})_\varepsilon \qquad \sup\{\lambda(u) - \tfrac{1}{2}\varepsilon\|u\|^2_{H^1(\Omega)} \mid u \in U_{\mathrm{ad}} \cap H^1(\Omega)\}.$$

Note that problem $(\mathrm{P2})_0$, for $\varepsilon = 0$ does not coincide with (P2); however, it is true that

$$\sup\{\lambda(u) \mid u \in U_{\mathrm{ad}}\} = \sup\{\lambda(u) \mid u \in U_{\mathrm{ad}} \cap H^1(\Omega)\}.$$

We prove existence of an optimal solution $v_\varepsilon \in U_{\mathrm{ad}} \cap H^1(\Omega)$ to the problem $(\mathrm{P2})_\varepsilon$ and we obtain necessary conditions of optimality for this problem.

THEOREM 5. *There exists an optimal solution $v_\varepsilon \in U_{\mathrm{ad}} \cap H^1(\Omega)$ to the problem $(\mathrm{P2})_\varepsilon$ for $\varepsilon > 0$. Necessary conditions of optimality take on the form*:

$$\inf_{\eta \in M(v_\varepsilon)} \int_\Omega g(\eta, x)(u(x) - v_\varepsilon(x))\, dx - \varepsilon(v_\varepsilon, u - v_\varepsilon)_{H^1(\Omega)} \leqq 0 \quad \forall u \in U_{\mathrm{ad}} \cap H^1(\Omega) \tag{3.5}$$

where

$$g(\eta, x) = \left(3v_\varepsilon^2(x) \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} b_{ijkl} \frac{\partial^2 \eta}{\partial x_k\, \partial x_l}(x) \frac{\partial^2 \eta}{\partial x_k\, \partial x_l}(x) - \lambda(v_\varepsilon)\eta^2(x)\right) \Big/ \left(\int_\Omega v_\varepsilon(x)\eta^2(x)\, dx\right), \qquad \eta \in H_0^2(\Omega), \quad \eta \neq 0.$$

For the proof of Theorem 5 we need the following lemmas:

LEMMA 4. *Let there be given a sequence $\{u_n\} \subset U_{\mathrm{ad}}$ and an element $u \in L^\infty(\Omega)$ such that for $n \to \infty$*

$$u_n(x) \to u(x) \quad \text{a.e. in } \Omega.$$

Then

$$\lim_{n \to \infty} \lambda(u_n) = \lambda(u). \tag{3.6}$$

The proof of Lemma 4 is omitted.

LEMMA 5. *There exists a weakly compact set $W \subset H_0^2(\Omega)$ such that for every element $u \in U_{\mathrm{ad}}$ the smallest eigenvalue of* (3.1) *can be defined as*:

$$\lambda(u) = \inf\{F(u, \eta) \mid \eta \in W\}. \tag{3.7}$$

Proof. It is sufficient to prove that there exists a constant $R < \infty$ such that:

$$\forall u \in U_{\mathrm{ad}}, \quad \forall \eta \in M(u): \quad \|\eta\|_{H_0^2(\Omega)} \leqq R. \tag{3.8}$$

If (3.8) holds, we can define the set W as

$$W = \{\eta \in H_0^2(\Omega) \mid \|\eta\|_{H_0^2(\Omega)} \leqq R, \|\eta\|_{L^2(\Omega)} = 1\}. \tag{3.9}$$

It can be verified that the set (3.9) is a weakly compact subset of $H_0^2(\Omega)$. In order to prove (3.8) let us recall [4], that there exist constants $0 < M_1 \leqq M_2 < \infty$ such that:

$$M_1\|\phi\|^2_{H_0^2(\Omega)} \leqq a_u(\phi, \phi) \leqq M_2\|\phi\|^2_{H_0^2(\Omega)} \quad \forall u \in U_{\mathrm{ad}}, \quad \forall \phi \in H_0^2(\Omega); \tag{3.9}'$$

furthermore the following estimation holds:

$$\frac{M_1}{C_2}\|\phi\|^2_{H_0^2(\Omega)} \leqq F(u, \phi) \leqq \frac{M_2}{C_1}\|\phi\|^2_{H_0^2(\Omega)} \quad \forall u \in U_{\mathrm{ad}}, \quad \forall \phi \in H_0^2(\Omega), \quad \|\phi\|_{L^2(\Omega)} = 1. \tag{3.10}$$

Let $\eta_0 \in H_0^2(\Omega)$ be any element such that $\|\eta_0\|_{L^2(\Omega)} = 1$. From (3.10) it follows that if we put

$$R = \sqrt{\frac{M_2}{M_1}\frac{C_2}{C_1}}\|\eta_0\|_{H_0^2(\Omega)}, \tag{3.11}$$

then (3.8) holds.

LEMMA 6. *Let there be given sequences* $\{u_k\} \subset U_{ad}$, $\{\eta_k\} \subset H_0^2(\Omega)$, $\eta_k \subset M(u_k)$, $k = 1, 2, \cdots$, *and an element* $u \in L^\infty(\Omega)$ *such that for* $k \to \infty$:

$$u_k(x) \to u(x) \quad \text{a.e. } in\ (\Omega). \tag{3.12}$$

Then there exist an element $\eta \in M(u)$ *and a subsequence of* $\{\eta_k\}$, *still denoted* $\{\eta_k\}$ *such that for* $k \to \infty$

$$\eta_k \to \eta \quad \textit{strongly in } H_0^2(\Omega). \tag{3.13}$$

Proof. By (3.8) it follows that $\|\eta_k\|_{H_0^2(\Omega)} \leqq R$. Hence there exists an element $\eta \in H_0^2(\Omega)$ such that for a subsequence, still denoted $\{\eta_k\}$ we have

$$\eta_k \to \eta \quad \text{weakly in } H_0^2(\Omega) \tag{3.14}$$

and by compact imbedding $H_0^2(\Omega) \subset C(\bar{\Omega})$ [12] it implies that:

$$\eta_k(x) \to \eta(x) \quad \text{uniformly on } \bar{\Omega}. \tag{3.15}$$

We have to prove that $\eta \in M(u)$.

Let there be given an element $\bar{\eta} \in M(u)$. By (3.12) it follows that:

$$\forall \phi \in H_0^2(\Omega): \quad L_{u_k}\phi \to L_u\phi \quad \text{strongly in } H^{-2}(\Omega) \tag{3.16}$$

and

$$\lim_{k \to \infty} a_{u_k}(\bar{\eta}, \bar{\eta}) = a_u(\bar{\eta}, \bar{\eta}), \tag{3.17}$$

$$\lim_{k \to \infty} \inf a_{u_k}(\eta_k, \eta_k) \geqq a_u(\eta, \eta); \tag{3.18}$$

hence

$$\lim_{k \to \infty} F(u_k, \bar{\eta}) = F(u, \bar{\eta}), \tag{3.19}$$

$$\lim_{k \to \infty} \inf F(u_k, \eta_k) \geqq F(u, \eta). \tag{3.20}$$

On the other hand by definition of η_k

$$F(u_k, \eta) \geqq F(u_k, \eta_k), \qquad k = 1, 2, \cdots. \tag{3.21}$$

Hence (3.19) and (3.20) imply that

$$F(u, \bar{\eta}) \geqq F(u, \eta) \tag{3.22}$$

which shows that $\eta \in M(u)$.

In order to prove (3.13) let us recall that by (3.16) it follows [17] that

$$\forall f \in H^{-2}(\Omega) \quad L_{u_k}^{-1} f \to L_u^{-1} f \quad \text{strongly in } H_0^2(\Omega); \tag{3.23}$$

hence

$$L_{u_k}^{-1} f_k \to L_u^{-1} f \quad \text{strongly in } H_0^2(\Omega) \tag{3.24}$$

whenever

$$f_k \to f \quad \text{strongly in } H^{-2}(\Omega). \tag{3.25}$$

To conclude the proof let us note that

$$\eta_k = L_{u_k}^{-1} f_k \tag{3.26}$$

where $f_k = \lambda(u_k) u_k \eta_k$. By Lemma 4 and by (3.12), (3.15) it follows that $f_k \to f = \lambda(u) u \eta$ strongly in $L^2(\Omega)$, thus strongly in $H^{-2}(\Omega)$.

Proof of Theorem 5. The proof of existence of an optimal solution $v_\varepsilon \in U_{\text{ad}} \cap H^1(\Omega)$ is standard and it is omitted.

In order to obtain the form (3.5) of necessary conditions of optimality we can apply Theorem 2, where

$$g(u) = \inf\{F(u, \eta) \mid \eta \in W \subset H_0^2(\Omega)\}, \qquad u \in E;$$

the set W has the form (3.9) and the set $E \subset L^\infty(\Omega)$ is defined as in Lemma 1.

THEOREM 6. *There exist a subsequence, still denoted by* $\{v_\varepsilon\}$, *an element* $v_0 \in L^\infty(\Omega)$, *a matrix* $Q = [q_{\alpha\beta}]_{4\times 4}$ *of coefficients* $q_{\alpha\beta} = q_{ijkl} \in L^\infty(\Omega)$, $i, j, k, l = 1, 2$, *and a number* $\lambda^* > 0$ *such that for* $\varepsilon \downarrow 0$

$$v_\varepsilon^3[b_{\alpha\beta}] \overset{G}{\Longrightarrow} [q_{\alpha\beta}] \quad \text{in } \Omega, \tag{3.27}$$

$$\lambda(v_\varepsilon) \to \lambda^*, \tag{3.28}$$

$$\lim_{\varepsilon \downarrow 0} \left\{ \lambda(v_\varepsilon) - \frac{\varepsilon}{2} \|v_\varepsilon\|^2_{H^1(\Omega)} \right\} = \sup\{\lambda(u) \mid u \in U_{\text{ad}}\} = \lambda^*, \tag{3.29}$$

where λ^* *is the smallest eigenvalue of the problem*

$$\sum_{i,j,k,l=1}^{2} \int_\Omega q_{ijkl}(x) \frac{\partial^2 \eta}{\partial x_i \, \partial x_j}(x) \frac{\partial^2 \phi}{\partial x_k \, \partial x_l}(x) = \lambda \int_\Omega v_0(x) \eta(x) \phi(x) \, dx \quad \forall \phi \in H_0^2(\Omega) \tag{3.30}$$

and $\eta \in H_0^2(\Omega)$ *is the eigenfunction corresponding to the eigenvalue* λ^*.

Proof. Sequences $\{\eta_\varepsilon\}$, $\{v_\varepsilon\}$, $\{\lambda(v_\varepsilon)\}$ are uniformly bounded in spaces $H_0^2(\Omega)$, $L^\infty(\Omega)$, R respectively; hence there exist elements:

$$\bar{\eta} \in H_0^2(\Omega), \quad v_0 \in L^\infty(\Omega), \quad \bar{\lambda} \in R$$

such that for subsequences:

$$\eta_\varepsilon \to \bar{\eta} \quad \text{weakly in } H^2(\Omega), \tag{3.31}$$

$$v_\varepsilon \to v \quad \text{weakly-(*) in } L^\infty(\Omega), \tag{3.32}$$

$$\lambda(v_\varepsilon) \to \bar{\lambda} \quad \text{in } R. \tag{3.33}$$

Furthermore by Lemma 3 it follows that there exists a matrix of coefficients $Q = [q_{\alpha\beta}]$ such that

$$v_\varepsilon^3[b_{\alpha\beta}] \overset{G}{\Longrightarrow} [q_{\alpha\beta}] \quad \text{in } \Omega. \tag{3.34}$$

By (3.31) and by compact imbedding [12] $H_0^2(\Omega) \subset C(\bar{\Omega})$ it follows that

$$\eta_\varepsilon(x) \to \bar{\eta}(x) \quad \text{uniformly on } \bar{\Omega}. \tag{3.35}$$

On the other hand (3.35), (3.32), (3.33) implies that

$$\lambda(v_\varepsilon)v_\varepsilon\eta_\varepsilon \to \bar{\lambda}v_0\bar{\eta} \quad \text{weakly in } L^2(\Omega). \tag{3.36}$$

Denote $f_\varepsilon = \lambda(v_\varepsilon)v_\varepsilon\eta_\varepsilon$, $f = \bar{\lambda}v_0\bar{\eta}$, $f_\varepsilon, f \in L^2(\Omega)$. By compact imbedding $L^2(\Omega) \subset H^{-2}(\Omega)$ it follows that:

$$f_\varepsilon \to f \quad \text{strongly in } H^{-2}(\Omega). \tag{3.37}$$

For given matrix $Q = [q_{\alpha\beta}]$ denote:

$$a_Q(y, z) = \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_\Omega q_{ijkl}(x) \frac{\partial^2 y}{\partial x_i\, \partial x_j}(x) \frac{\partial^2 z}{\partial x_k\, \partial x_l}(x)\, dx \quad \forall y, z \in H_0^2(\Omega), \tag{3.38}$$

$$F_Q(y) = a_Q(y, y) \Big/ \int_\Omega v_0(x) y^2(x)\, dx, \qquad y \in H_0^2(\Omega), \quad y \neq 0, \tag{3.39}$$

$$\lambda^* = \inf\{F_Q(y) \mid y \in H_0^2(\Omega), \|y\|_{L^2(\Omega)} = 1\}, \tag{3.40}$$

$$M^*(Q) = \{y \in H_0^2(\Omega) \mid F_Q(y) = \lambda^*, \|y\|_{L^2(\Omega)} = 1\}. \tag{3.41}$$

Since bilinear form $a_u(\cdot, \cdot)$ is uniformly continuous and coercive [4] on the space $H_0^2(\Omega)$ for $u \in U_{\text{ad}}$, then (2.24) and (3.9)′ imply that

$$M_1\|y\|^2_{H_0^2(\Omega)} \leqq a_Q(y, y) \leqq M_2\|y\|^2_{H_0^2(\Omega)} \quad \forall y \in H_0^2(\Omega);$$

hence the set $M^*(Q)$ is not empty and λ^* is the smallest eigenvalue of the eigenvalue problem (3.30).

Note that $L_{v_\varepsilon}\eta_\varepsilon = f_\varepsilon$ in Ω i.e.:

$$a_{v_\varepsilon}(\eta_\varepsilon\phi) = \langle f_\varepsilon, \phi\rangle \quad \forall \phi \in H_0^2(\Omega). \tag{3.42}$$

By (3.31), (3.34) and by the definition of G-convergence it follows that:

$$\lim_{\varepsilon \downarrow 0} a_{v_\varepsilon}(\bar{\eta}_\varepsilon, \phi) = a_Q(\bar{\eta}, \phi). \tag{3.43}$$

Thus by (3.37), (3.42):

$$a_Q(\bar{\eta}, \phi) = \langle f, \phi\rangle = \bar{\lambda} \int_\Omega v_0(x)\bar{\eta}(x)\phi(x)\, dx \quad \forall \phi \in H_0^2(\Omega). \tag{3.44}$$

On the other hand by (3.32), (3.35), (3.43) we have

$$\lim_{\varepsilon \downarrow 0} F(v_\varepsilon, \eta_\varepsilon) = F_Q(\bar{\eta}) \tag{3.45}$$

and by definition of the smallest eigenvalue

$$\lambda(v_\varepsilon) \leqq F(v_\varepsilon, \eta) \quad \forall \eta \in M^*(Q), \quad \varepsilon > 0;$$

hence for $\varepsilon \downarrow 0$ we obtain:

$$\bar{\lambda} = F_Q(\bar{\eta}) \leqq F_Q(\eta) = \lambda^* \quad \forall \eta \in M^*(Q);$$

thus $\bar{\eta} \in M^*(Q)$ and $\bar{\lambda} = \lambda^*$.

4. Numerical results. Problems $(\text{P1})_\varepsilon$, $(\text{P2})_\varepsilon$, for $\varepsilon > 0$, are nonsmooth infinite dimensional optimization problems. The finite element method was used for discretization of the elliptic boundary value problem (1.4). In order to solve the resulting finite dimensional, nonconvex and nonsmooth optimization problem, the method of Lemarechal [11] combined with the shifted penalty function method [5] was applied.

The QZ algorithm [14] was used for computation of the smallest eigenvalue to the eigenvalue problem (3.1). The domain Ω had the form:

$$\Omega = \{(x_1, x_2) \in R^2 \,|\, x_i \in (0, 1),\ i = 1, 2\}.$$

The computations were carried out for a model rectangular clamped plate divided into 64 rectangular finite elements of the Bogner-Fox-Schmidt type [3]. The constants c_1, c_2 in (1.3) were equal to 0.8 and 1.2 respectively. Constant c was equal to 1.0 and $\varepsilon = 0.0001$ was used in the computations.

The optimal values of the cost functionals were referred to the value of the cost functionals computed for the plate with constant thickness $\bar{u}(x) \equiv 1$, $x \in \Omega$ and the following ratios were computed:

$$r_1 = |J_\varepsilon(\bar{u}) - J_\varepsilon(u_\varepsilon)| / J_\varepsilon(\bar{u}) \quad \text{for } (\mathrm{P1})_\varepsilon,$$

$$r_2 = |\lambda_\varepsilon(\bar{u}) - \lambda_\varepsilon(v_\varepsilon)| / \lambda_\varepsilon(\bar{u}) \quad \text{for } (\mathrm{P2})_\varepsilon$$

where

$$J_\varepsilon(u) = \sup_{x \in \Omega} y(u; x) + \tfrac{1}{2}\varepsilon \|u\|^2_{H^1(\Omega)}, \qquad \lambda_\varepsilon(v) = \lambda(v) - \tfrac{1}{2}\varepsilon \|v\|^2_{H^1(\Omega)}.$$

The following numerical results were obtained:

(i) For problem $(\mathrm{P1})_\varepsilon$ the results are shown in Fig. 2 and in Table 1 for one quarter of the plate loaded by distributed force $f(x) = 1$, $x \in \Omega$. Note that it is sufficient to present results for one quarter of the plate since the optimal solution is symmetric in this case. In this case $r_1 = 0.63$. The thickness of the plate is maximal in the middle of the rectangular Ω and in the middle of its boundaries.

(ii) For problem $(\mathrm{P2})_\varepsilon$ the results are shown in Table 2 and in Fig. 3. In this case $r_2 = 1.01$. The smallest eigenvalue is double for the optimum thickness plate. The material concentrates in the middle of the boundaries and around the middle of the plate.

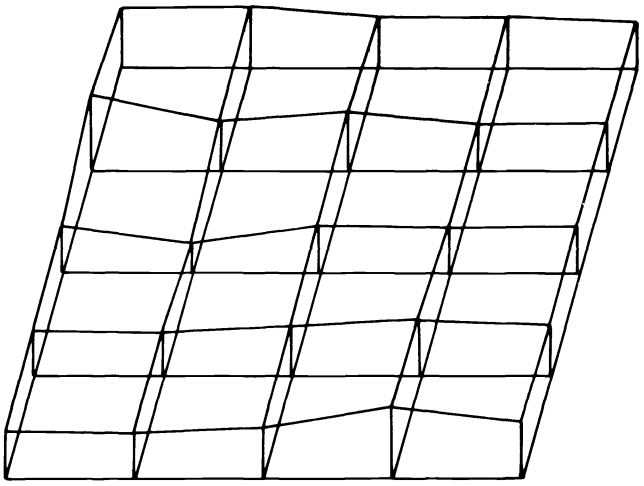


FIG. 2. *Problem* (P1), *one quarter of optimal shape plate.*

TABLE 1
Numerical results for problem (P1). *Thickness of one quarter of optimal shape plate.*

	x_1-coordinate				
x_2-coord.	0.000	0.125	0.250	0.375	0.500
0.500	1.04	1.04	1.00	1.00	0.98
0.375	1.11	0.99	1.03	0.99	0.99
0.250	0.99	0.89	0.99	0.99	0.99
0.125	0.89	0.97	0.99	1.02	1.01
0.000	0.99	0.98	1.00	1.08	1.02

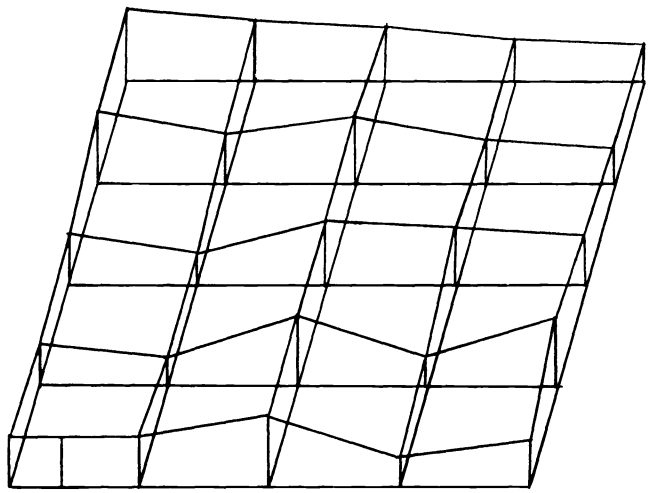


FIG. 3. *Problem* (P2), *one quarter of optimal shape plate.*

TABLE 2
Numerical results for problem (P2). *Thickness of one quarter of optimal shape plate.*

	x_1-coordinate				
x_2-coord.	0.000	0.125	0.250	0.375	0.500
0.500	1.18	1.10	1.00	0.88	0.84
0.375	1.20	1.02	1.12	0.94	0.86
0.250	1.04	0.80	1.11	1.06	0.99
0.125	0.91	0.83	1.19	0.81	1.20
0.000	0.98	0.96	1.18	0.80	0.95

Appendix.

Proof of Lemma 3. By [17, Thms. 9, 13] it follows that there exist a 4×4 symmetric matrix of coefficients $[q_{\alpha\beta}]$, $q_{\alpha\beta}=q_{ijkl}\in L^{\infty}(\Omega)$, $i,j,k,l=1,2$ and a subsequence $\{u_n^3[b_{\alpha\beta}]\}$, $n=1,2,\cdots$ such that (2.15) and (2.22) hold. In order to obtain the estimate (2.24) consider an auxiliary elliptic boundary value problem of the form:

Given an element $w \in C^\infty(R^2)$, determine an element $y_n \in H^2(\Omega)$ such that:

$$y_n - w \in H_0^2(\Omega), \tag{1}$$

$$\sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_\Omega u_n^3(x) b_{ijkl} \frac{\partial^2 y_n}{\partial x_i \partial x_j}(x) \frac{\partial^2 z}{\partial x_k \partial x_l}(x)\, dx = \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_\Omega q_{ijkl}(x) \frac{\partial^2 w}{\partial x_i \partial x_j}(x) \frac{\partial^2 z}{\partial x_k \partial x_l}(x)\, dx \quad \forall z \in H_0^2\Omega. \tag{2}$$

It can be shown in a way similar to [17, Lemma 5, p. 86; Thm. 12, p. 91] that the following convergences take place for $n \to \infty$:

$$y_n \to w \quad \text{weakly in } H^2(\Omega), \tag{3}$$

$$\sum_{i,j=1}^{2} u_n^3 b_{ijkl} \frac{\partial^2 y_n}{\partial x_i \partial x_j} \to \sum_{i,j=1}^{2} q_{ijkl} \frac{\partial^2 w}{\partial x_i \partial x_j}, \tag{4}$$

weakly in $L^2(\Omega)$ for every $k, l = 1, 2$,

$$\sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_\Omega \theta(x) b_{ijkl} \frac{\partial^2 y_n}{\partial x_i \partial x_j}(x) \frac{\partial^2 y_n}{\partial x_k \partial x_l}(x) u_n^3(x)\, dx \to \sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_\Omega \theta(x) q_{ijkl}(x) \frac{\partial^2 w}{\partial x_i \partial x_j}(x) \frac{\partial^2 w}{\partial x_k \partial x_l}(x)\, dx \quad \forall \theta \in D(\Omega). \tag{5}$$

Given $\lambda_1, \lambda_2, \lambda \in R$, define

$$w(x) = \tfrac{1}{2}\lambda_1 x_1^2 + \tfrac{1}{2}\lambda_2 x_2^2 + \lambda_3 x_1 x_2, \qquad x \in R^2.$$

Note that the following inequality holds:

$$\forall \theta \in D(\Omega), \quad \theta(x) \geqq 0:$$

$$\sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_\Omega \theta(x) u_n^3(x) b_{ijkl} \frac{\partial^2}{\partial x_i \partial x_j}(y_n(x) - w(x)) \frac{\partial^2}{\partial x_k \partial x_l}(y_n(x) - w(x))\, dx \geqq 0. \tag{6}$$

Since $0 < c_1^3 \leqq u_n^3(x) \leqq c_2^3$ a.e. in Ω, then there exist elements $\psi_1, \psi_2 \in L^\infty(\Omega)$ such that

$$u_n^3(\cdot) \to \psi_2(\cdot) \qquad \text{weakly-}(*) \text{ in } L^\infty(\Omega), \tag{7}$$

$$1/u_n^3(\cdot) \to 1/\psi_1(\cdot) \quad \text{weakly-}(*) \text{ in } L^\infty(\Omega); \tag{8}$$

furthermore $(\psi_2(x), 1/\psi_1(x)) \in \overline{\operatorname{conv} X}$ for a.e. $x \in \Omega$ where $X = \{(a, b) \in R^2 \mid ab = 1, c_1^3 \leqq a \leqq c_2^3\}$.

Taking into account (3), (4), (5), (7), we can pass to the limit in (6) and we obtain the right-hand side inequality in (2.24), with

$$[e_{ij}] = \begin{bmatrix} \lambda_1 & \lambda_3 \\ \lambda_3 & \lambda_2 \end{bmatrix}.$$

In order to obtain the left-hand side inequality in (2.24), denote

$$[d_{\alpha\beta}] = [b_{\alpha\beta}]^{-1} \tag{9}$$

where $d_{\alpha\beta} = d_{ijkl}$, $i, j, k, l = 1, 2$, and define

$$\mu_{ij}(x) = \psi_1(x) \sum_{k,l=1}^{2} b_{ijkl} \frac{\partial^2 w}{\partial x_k \partial x_l}(x), \tag{10}$$

$$a_{ij,n}(x) = u_n^3(x) \sum_{k,l=1}^{2} b_{ijkl} \frac{\partial^2 y_n}{\partial x_k \partial x_l}(x), \qquad x \in \Omega, \quad i,j = 1,2. \tag{11}$$

Note that for $n = 1, 2, \cdots$ the following inequality is satisfied: $\forall \theta \in D(\Omega)$, $\theta(x) \geqq 0$

$$\sum_{i,j=1}^{2} \sum_{k,l=1}^{2} \int_{\Omega} \theta(x) \frac{d_{ijkl}}{u_n^3(x)} (a_{ij,n}(x) - \mu_{ij}(x))(a_{kl,n}(x) - \mu_{kl}(x))\, dx \geqq 0. \tag{12}$$

Taking into account (3), (4), (5), (8), we can pass to the limit in (12) and we obtain the left inequality in (2.24).

Acknowledgments. The authors are very indebted to Professors Claude Lemarechal, Francois Murat, Kazimierz Malanowski and Jean-Paul Zolesio for stimulating discussions and valuable comments concerning the subject of this paper. Lemarechal's implementation of his method was used for computations.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] N. V. BANICHUK, *Optimization of Shape of Elastic Bodies*, Nauka, Moscow, 1980. (In Russian.)
[3] P. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[4] G. FICHERA, *Existence theorems in elasticity*, Handbuch Der Physik, Vol. VI a/2 (1972), pp. 347-389.
[5] W. FINDEISEN, J. SZYMANOWSKI AND A. WIERZBICKI, *Theory and Numerical Methods of Optimization*, Polish Scientific Publisher, Warsaw, 1980. (In Polish.)
[6] E. J. HAUG AND J. S. ARORA, *Applied Optimal Design*, Wiley Interscience, New York, 1979.
[7] E. J. HAUG AND J. CEA, *Optimization of Distributed Parameter Structures*, Sijthoff and Nøordhoff, Alphen aan den Rijn, Netherlands, 1981.
[8] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[9] R. V. KOHN AND G. STRANG, *Structural design optimization, homogenization and relaxation of variational problems*, Proc. Conference on Disordered Media, New York Univ., June 1981, Lectures Notes in Physics 154, Springer, Berlin, 1982, pp. 131-147.
[10] W. KOZLOWSKI AND Z. MROZ, *Optimal design of solid plates*, Int. J. Solids Structures, 5 (1969), pp. 781-794.
[11] C. LEMARECHAL, *An extension of Davidon's method to nondifferentiable problems*, in Mathematical Programming Study 3: Non-Differentiable Optimization, North-Holland, Amsterdam, 1975, pp. 95-109.
[12] J. L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1972.
[13] K. A. LURIE AND A. V. CHERKAEV, *G-closure of some particular sets of admissible material characteristics for the problem of bending of thin plates*, DCAMM Report No. 214, Technical University of Denmark, 1981.
[14] C. B. MOLER AND G. W. STEWART, *An algorithm for generalized matrix eigenvalue problems*, SIAM J. Numer. Anal., 10 (1973), pp. 241-256.
[15] F. MURAT, *Contre-example pour divers problems où le contrôle intervient dans les coefficients*, Ann. Math. Pures Appl., 112 (1977), pp. 143-168.
[16] A. MYŚLIŃSKI, *A nonsmooth optimal design problem of bending thin plates*, Control and Cybernetics, 11 (1982), pp. 121-134.
[17] V. V. ZHIKOV, S. M. KOZLOV, O. A. OLEINIK AND KHA TIENG NGOAN, *Homogenization and G-convergence of differential operators*, Uspekhi Matem. Nauk, 34, No. 5 (209), (1979), pp. 65-133. (In Russian.)
[18] N. OLHOFF, *Optimal design of vibrating rectangular plates*, Int. J. Solids Structures, 10 (1974), pp. 93-109.
[19] N. OLHOFF, K. A. LURIE, A. V. CHERKAEV AND A. V. FEDOROV, *Sliding regimes and anisotropy in optimal design of vibrating axisymmetric plates*, Int. J. Solids Structures, 17 (1981), pp. 931-948.
[20] N. OLHOFF AND K. T. CHENG, *An investigation concerning optimal design of solid elastic plates*, Int. J. Solids Structures, 17 (1981), pp. 305-323.
[21] U. E. RAITUM, *On optimal control problems for linear elliptic equations*, Soviet Math. Dokl., 20 (1979), pp. 129-132.

[22] A. SAWCZUK AND Z. MROZ, *Optimization in Structural Design*, Proc. 1973 IUTAM Symposium, Warsaw, Poland, Springer-Verlag, 1975.

[23] J. SOKOŁOWSKI, *Optimal control in coefficients for weak variational problems in Hilbert space*, Appl. Math. Optim., 7 (1981), pp. 283-293.

[24] L. TARTAR, *Estimation de coefficients homogeneises*, Lecture Notes in Mathematics 204, Springer-Verlag, New York, pp. 364-373.

[25] ———, *Problèmes de contrôle des coefficients dans des équations aux derivées partielles*, Lecture Notes in Computer Sciences 41, Springer-Verlag, New York, pp. 420-426.

[26] J. P. ZOLESIO, *Identification de domaines par déformation*, Thesis, Nice Univ., 1979.

[27] ———, *Semiderivatives of repeated eigenvalues*, in Optimization of Distributed Parameter Structures, E. J. Haug and J. Cea, eds., Sijthoff and Nøørdhoff, Alphen aan den Rihn, Netherlands, 1981.

[28] E. J. HAUG AND B. ROUSSELET, *Design sensitivity analysis in structural mechanics.* I: *Statistic response variations*, II: *Eigenvalue variations*, J. Struct. Mech., 8 (1980), pp. 17-41; pp. 161-186.

[29] B. ROUSSELET, *Etude de la regularité des valeurs propres par rapport à des déformations bilipschitziennes du domaine*, CRAS Paris, 283 (1976), p. 507.

[30] J. CEA AND K. MALANOWSKI, *An example of maximum problem in partial differential equations*, this Journal, 8 (1970), pp. 305-316.

[31] V. V. ZHIKOV, S. M. KOZLOV AND O. A. OLEINIK, *On G-convergence of the parabolic operators*, Uspekhi Matem. Nauk, 36, No. 1 (217) (1981), pp. 11-58. (In Russian.)

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

A CHARACTERIZATION OF PROPERLY MINIMAL ELEMENTS OF A SET*

JOHANNES JAHN†

Abstract. In this paper properly minimal elements of a set are characterized as minimal solutions of appropriate approximation problems without any convexity assumptions.

Key words. vector optimization, scalarization, approximation

1. Introduction and problem formulation. Properly minimal elements play an important role in vector optimization (in connection with duality investigations see, for instance [11], [12] and [5]). This notion was first introduced by Kuhn-Tucker [15] and modified by Geoffrion [7], and later it was formulated in a more general framework (e.g., see [3] but also [2], [18], [19], [8], [1], [4], [16] and [9]).

Borwein [3, Thm. 2] gave an interesting characterization of properly minimal elements as minimal solutions of appropriate scalar optimization problems. For this characterization the convexity of the vector optimization problem is assumed. In [13] suitable characterizations of minimal and weakly minimal elements are presented without assuming any convexity. It is the aim of this paper to extend this theory to the notion of properly minimal elements and to formulate a similar characterization without any convexity assumption. Since we are concerned with approximation problems, the following theory is developed in a normed linear space.

Throughout this paper let $(Y, \|\cdot\|_Y)$ be a real normed space, and let C_Y be a convex cone in Y. Obviously C_Y induces a partial ordering in Y. Further, let a nonempty subset V of Y be given. Then we ask for properly minimal elements of the set V.

Before we present the definition of these minima we list some useful notations. The *algebraic sum* of two nonempty subsets S and T of the real linear space Y is denoted by

$$S+T := \{s+t \,|\, s \in S \text{ and } t \in T\}.$$

A nonempty set $S \subset Y$ is called *starshaped* at $\bar{s} \in S$, if

$$\lambda s + (1-\lambda)\bar{s} \in S \text{ for all } s \in S \text{ and all } \lambda \in [0,1].$$

The *core* of a set $S \subset Y$ is given as

$$\operatorname{cor}(S) := \{s \in S \,|\, \text{for each } y \in Y \text{ there is some } \bar{\lambda} > 0 \text{ with } s + \lambda y \in S \text{ for all } \lambda \in (0, \bar{\lambda}]\}.$$

$\bar{S}$ (and $\bar{S}^{\sigma(Y,Y^*)}$, respectively) denotes the *closure* of a set $S \subset Y$ in the norm topology (and the weak topology $\sigma(Y, Y^*)$, respectively), and $\operatorname{int}(S)$ means the *interior* of the set S. For some $\bar{y} \in Y$ and some $\eta > 0$ the notation

$$N(\bar{y}, \eta) := \{y \in Y \,|\, \|y - \bar{y}\|_Y \leqq \eta\}$$

is used for a closed *ball* around $\bar{y}$ with radius η. The cone *generated* by a nonempty

* Received by the editors October 4, 1983, and in revised form April 25, 1984.

† Technische Hochschule Darmstadt, Fachbereich Mathematik, Schlossgartenstrasse 7, 6100 Darmstadt, West Germany. This paper was written when the author was a visitor at the Department of Mathematics of North Carolina State University, Raleigh, North Carolina 27695-8205.

subset S of Y is denoted

$$\operatorname{cone}(S) := \{\lambda s \mid \lambda \geqq 0 \text{ and } s \in S\}.$$

Furthermore, we recall that the (Bouligand) *tangent cone* $T(S, \bar{s})$ to a set $S \subset Y$ at some $\bar{s} \in S$ is defined as

$$T(S, \bar{s}) := \{\lim_{n\to\infty} \lambda_n(s_n - \bar{s}) \mid \lambda_n \geqq 0 \text{ and } s_n \in S \text{ for all } n \in \mathbb{N} \text{ and } \bar{s} = \lim_{n\to\infty} s_n\}$$

(the convergence is to be understood with respect to the norm $\|\cdot\|_Y$). The tangent cone $T(S, \bar{s})$ is a local approximation of the set S at $\bar{s}$. $T(S, \bar{s})$ is closed (in a normed setting), and it is even convex, if the set S is convex.

DEFINITION 1.1. Let $(Y, \|\cdot\|_Y)$ be a real normed space, let C_Y be a convex cone in Y, and let V be a nonempty subset of Y.

a) An element $\bar{y} \in V$ is called a *minimal* element of the set V, if

$$(\{\bar{y}\} - C_Y) \cap V = \{\bar{y}\}.$$

b) An element $\bar{y} \in V$ is called a *properly minimal* element of the set V, if $\bar{y}$ is a minimal element of the set V and the zero element 0_Y is a minimal element of the tangent cone $T(V + C_Y, \bar{y})$.

Definition 1.1b was given by Borwein [3] in the case of a topological linear space Y. Benson [1] gave a slightly different definition: An element $\bar{y} \in V$ is called a properly minimal element of the set V (in the sense of Benson), if $\bar{y}$ is a minimal element of the set V and the zero element 0_Y is a minimal element of the set $\overline{\operatorname{cone}(V + C_Y - \{\bar{y}\})}$. If the set $V + C_Y$ is starshaped at $\bar{y}$, a well-known result states that

$$T(V + C_Y, \bar{y}) = \overline{\operatorname{cone}(V + C_Y - \{\bar{y}\})}$$

(e.g., see [14, p. 155]); in this case the definitions of Borwein and Benson coincide.

In [4] Borwein presented an interesting extension of his concept of proper minimality: An element $\bar{y} \in V$ is called a properly minimal element of the set V, if

$$(-C_Y) \cap \overline{\operatorname{cone}(V - \{\bar{y}\})} \subset C_Y.$$

If C_Y is pointed (i.e. $(-C_Y) \cap C_Y = \{0_Y\}$), then this inclusion reduces to

$$(-C_Y) \cap \overline{\operatorname{cone}(V - \{\bar{y}\})} = \{0_Y\}.$$

Since we assume implicitly in the theorems of the next section that the ordering cone C_Y is pointed (although this assumption is not formulated explicitly), the main result of this paper does not hold for Borwein's extended concept of proper minimality [4].

2. A characterization result. First, we formulate a sufficient condition for properly minimal elements of a set. For this purpose we need

DEFINITION 2.1. Let S be a nonempty subset of a real linear space Y, and let C_Y be a convex cone in Y. A functional $\varphi : S \to \mathbb{R}$ is called *strongly monotonically increasing on* S, if for each $\bar{s} \in S$

$$s \in (\{\bar{s}\} - C_Y) \cap S,\ s \neq \bar{s} \Rightarrow \varphi(s) < \varphi(\bar{s}).$$

In the following we consider an additional norm $\|\cdot\|$ on Y which is strongly monotonically increasing on C_Y. If such a norm exists, then the ordering cone C_Y is necessarily pointed.

The next theorem states that certain approximation problems are especially qualified for the determination of properly minimal elements of a set where no convexity assumptions are required (for a convex version of this theorem see [13, Thm. 2.7.]).

THEOREM 2.2. *Let $(Y, \|\cdot\|_Y)$ be a real normed space, and let C_Y be a convex cone in Y which has a nonempty core. Let $\|\cdot\|$ be any (additional) norm on Y which is strongly monotonically increasing on C_Y and for which there exists some $\alpha > 0$ with*

$$\|y\| \leqq \alpha \|y\|_Y \quad \text{for all } y \in Y. \tag{1}$$

Further, let an element $\hat{y} \in Y$ with $V \subset \{\hat{y}\} + \operatorname{cor}(C_Y)$ be given. If there exists an element $\bar{y} \in V$ with the property

$$\|\bar{y} - \hat{y}\| \leqq \|y - \hat{y}\| \quad \text{for all } y \in V, \tag{2}$$

then $\bar{y}$ is a properly minimal element of the set V.

Proof. First, we show that $\bar{y}$ is a minimal element of the set V. If we assume that $\bar{y}$ is not a minimal element of V, then there exists some $y \in (\{\bar{y}\} - C_Y) \cap V$ with $y \neq \bar{y}$. Consequently, we obtain

$$y - \hat{y} \in (\{\bar{y} - \hat{y}\} - C_Y) \cap (V - \{\hat{y}\}) \subset (\{\bar{y} - \hat{y}\} - C_Y) \cap C_Y.$$

Since the norm $\|\cdot\|$ is strongly monotonically increasing on C_Y, we get

$$\|y - \hat{y}\| < \|\bar{y} - \hat{y}\|$$

which contradicts the assumption (2). So, $\bar{y}$ is a minimal element of the set V.

Next, we prove that 0_Y is a minimal element of the tangent cone $T(V + C_Y, \bar{y})$. Since the norm $\|\cdot\|$ is assumed to be strongly monotonically increasing on C_Y, we obtain from (2)

$$\|\bar{y} - \hat{y}\| \leqq \|y - \hat{y}\| \leqq \|y + c - \hat{y}\| \quad \text{for all } y \in V \text{ and all } c \in C_Y$$

and

$$\|\bar{y} - \hat{y}\| \leqq \|y - \hat{y}\| \quad \text{for all } y \in V + C_Y, \tag{3}$$

respectively. It is evident that the functional $\|\cdot - \hat{y}\|$ is convex. But it is also continuous in the topology generated by the norm $\|\cdot\|_Y$ because with the inequality (1) we conclude

$$|\|y_1 - \hat{y}\| - \|y_2 - \hat{y}\|| \leqq \|y_1 - y_2\| \leqq \alpha \|y_1 - y_2\|_Y \quad \text{for all } y_1, y_2 \in Y.$$

Then a well-known result from optimization theory states that the inequality (3) implies

$$\|\bar{y} - \hat{y}\| \leqq \|\bar{y} - \hat{y} + y\| \quad \text{for all } y \in T(V + C_Y, \bar{y}) \tag{4}$$

(e.g., see [14, p. 156]). With $S := T(V + C_Y, \bar{y}) \cap (\{\hat{y} - \bar{y}\} + C_Y)$ the inequality (4) is also true for all $y \in S$. With the same arguments as in the first part of this proof we conclude that 0_Y is a minimal element of the set S.

Now, we assume that 0_Y is not a minimal element of the tangent cone $T(V + C_Y, \bar{y})$. Then there exists some $y \in (-C_Y) \cap T(V + C_Y, \bar{y})$ with $y \neq 0_Y$. Because of the inclusion $V \subset \{\hat{y}\} + \operatorname{cor}(C_Y)$ there exists some $\lambda > 0$ with $\lambda y \in \{\hat{y} - \bar{y}\} + C_Y$. Consequently, we get

$$\lambda y \in (-C_Y) \cap T(V + C_Y, \bar{y}) \cap (\{\hat{y} - \bar{y}\} + C_Y),$$

and therefore we have $\lambda y \in (-C_Y) \cap S$ which contradicts the fact that 0_Y is a minimal element of the set S. Hence, 0_Y is a minimal element of the tangent cone $T(V + C_Y, \bar{y})$, and the assertion is obvious. □

Dinkelbach/Dürr [6] proved for $Y = \mathbb{R}^n$ and $C_Y = \mathbb{R}^n_+$ that certain l_p-norms are qualified for the determination of properly minimal elements in the sense of Geoffrion [7]. This result is extended by Theorem 2.2.

With the following theorem we answer the question under which assumptions a properly minimal element of a set is a minimal solution of an approximation problem. For a better insight we give the definition of a base of a cone (e.g., see [17, p. 25]).

DEFINITION 2.3. Let Y be a real linear space, and let $C \subset Y$ be a nontrivial convex cone (i.e. $C \neq \{0_Y\}$). A nonempty convex subset B of C is called a *base* for C, if each nonzero element $y \in C$ has a unique representation of the form $y = \lambda b$ for some $\lambda > 0$ and some $b \in B$.

If the ordering cone C of a real linear space Y has a base, then it is nontrivial and pointed (i.e. $(-C) \cap C = \{0_Y\}$) by definition. Now we establish the promised necessary condition for properly minimal elements.

THEOREM 2.4. *Let $(Y, \|\cdot\|_Y)$ be a real normed space, and let C_Y be a convex cone in Y which has a weakly compact base. Let V be a subset of Y, and for some $\bar{y} \in V$ let the cone generated by $T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})$ be weakly closed. If $\bar{y}$ is a properly minimal element of the set V, then for each $\hat{y} \in \{\bar{y}\} - C_Y$, $\hat{y} \neq \bar{y}$, there exists an (additional) norm $\|\cdot\|$ on Y which is strongly monotonically increasing on C_Y, for which there exists some $\alpha > 0$ with*

$$\|y\| \leqq \alpha \|y\|_Y \quad \textit{for all } y \in Y$$

and which has the property

$$1 = \|\bar{y} - \hat{y}\| < \|y - \hat{y}\| \quad \textit{for all } y \in V \setminus \{\bar{y}\}.$$

Proof. The proof of this theorem is rather technical, and therefore a short overview is given first in order to examine the geometry. In part 1 it is shown that the base of the convex cone $-C_Y$ and the cone S generated by $T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})$ have a positive "distance" ε. This allows us to construct another cone C in the second part which is "larger" than the ordering cone C_Y but for which $(-C) \cap S = \{0_Y\}$. It can be shown that C is convex, closed, pointed and that it has a nonempty interior. In part 3 we define the desired norm $\|\cdot\|$ as the Minkowski functional with respect to an appropriate order interval. Moreover, in part 4 several properties of the norm are proved.

1. In the following let B denote the base of the convex cone C_Y and let S denote the cone generated by $T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})$, i.e.

$$S := \operatorname{cone}(T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})).$$

Since the base B is weakly compact and for each $x \in S$ the functional $\|x - \cdot\|_Y : Y \to \mathbb{R}$ is weakly lower semicontinuous, for each $x \in S$ the scalar optimization problem

$$\inf_{y \in -B} \|x - y\|_Y$$

is solvable, i.e., there is a $y(x) \in -B$ with the property that

$$\|x - y(x)\|_Y \leqq \|x - y\|_Y \quad \text{for all } y \in -B.$$

Next, we consider the scalar optimization problem

$$\varepsilon := \inf_{x \in S} \|x - y(x)\|_Y.$$

If we assume $\varepsilon = 0$, then there exists an infimal net

$$\|x_n - y(x_n)\|_Y \to 0 \quad \text{with } x_n \in S. \tag{5}$$

Since B is weakly compact and S is weakly closed, the set $S + B$ is weakly closed, and

the condition (5) implies

$$0_Y \in \overline{S+B} \subset \overline{S+B}^{\sigma(Y,Y^*)} = S+B. \tag{6}$$

$\bar{y}$ is assumed to be a properly minimal element of the set V. Consequently, 0_Y is a minimal element of the tangent cone $T(V+C_Y, \bar{y})$ and a minimal element of the set $V-\{\bar{y}\}$, and we obtain

$$\begin{aligned}\{0_Y\} &= (-C_Y) \cap T(V+C_Y, \bar{y}) \cup (-C_Y) \cap (V-\{\bar{y}\}) \\ &= (-C_Y) \cap (T(V+C_Y, \bar{y}) \cup (V-\{\bar{y}\}))\end{aligned}$$

and

$$\begin{aligned}\{0_Y\} &= (-C_Y) \cap \operatorname{cone}(T(V+C_Y, \bar{y}) \cup (V-\{\bar{y}\})) \\ &\supset (-B) \cap S.\end{aligned}$$

Since $0_Y \notin B$, we conclude $(-B) \cap S = \varnothing$ which contradicts the condition (6). Thus, we get

$$0 < \varepsilon = \inf_{x \in S} \inf_{y \in -B} \|x-y\|_Y,$$

i.e. the sets S and $-B$ have a positive "distance" ε.

2. Now, we "separate" the sets $-B$ and S by a cone $-C$. Since the base B is weakly compact and $0_Y \notin B$ we obtain

$$0 < \delta := \inf_{y \in B} \|y\|_Y.$$

For

$$\beta := \min\left\{\frac{\varepsilon}{2}, \frac{\delta}{2}\right\} > 0$$

we define the set

$$U := B + N(0_Y, \beta)$$

($N(0_Y, \beta)$ denotes a closed ball around 0_Y with radius β). It is evident that U is a convex set. Consequently, the cone generated by U and its closure

$$C := \overline{\operatorname{cone}(U)}$$

is a convex cone. By definition, this cone has a nonempty interior. In order to see that C is pointed, we investigate the cone

$$\tilde{C} := \operatorname{cone}(B + N(0_Y, \tfrac{3}{2}\beta))$$

which is a superset of C. If we assume that there is a $\tilde{y} \in (-\tilde{C}) \cap \tilde{C}$ with $\tilde{y} \neq 0_Y$, then there exists a $\lambda > 0$ and a $y \in B + N(0_Y, \frac{3}{2}\beta)$ with $\tilde{y} = \lambda y$. Because of $-\tilde{y} = \lambda(-y) \in \tilde{C}$ we obtain for some $\mu > 0$

$$-\mu y \in B + N(0_Y, \tfrac{3}{2}\beta).$$

So, y and $-\mu y$ are elements of the convex set $B + N(0_Y, \frac{3}{2}\beta)$ which implies $0_Y \in B + N(0_Y, \frac{3}{2}\beta)$. But this is a contradiction to the choice of $\beta \leqq \delta/2$. Consequently, $\tilde{C}$ is pointed and with $C \subset \tilde{C}$ the cone C is pointed as well.

3. Next, we choose an arbitrary $\hat{y} \in \{\bar{y}\} - C_Y$ with $\hat{y} \neq \bar{y}$ and we define the order interval (with respect to the partial ordering induced by C)

$$[\hat{y}-\bar{y}, \bar{y}-\hat{y}] := (\{\hat{y}-\bar{y}\}+C) \cap (\{\bar{y}-\hat{y}\}-C).$$

Because of the construction of C and the set U, respectively, $\bar{y}-\hat{y}$ belongs to the interior of C. Furthermore, C is closed and pointed. Consequently, the Minkowski functional $\|\cdot\|: Y \to \mathbb{R}$ given by

$$\|y\| := \inf_{\lambda>0}\left\{\lambda \,\middle|\, \frac{1}{\lambda} y \in [\hat{y}-\bar{y}, \bar{y}-\hat{y}]\right\} \quad \text{for all } y \in Y$$

is a norm on Y and

$$[\hat{y}-\bar{y}, \bar{y}-\hat{y}] = \{y \in Y \mid \|y\| \leqq 1\}. \tag{7}$$

4. We have to show several properties of the norm $\|\cdot\|$. Since 0_Y belongs to the interior of the order interval $[\hat{y}-\bar{y}, \bar{y}-\hat{y}]$, there exists some $\alpha > 0$ with

$$N(0_Y, \alpha) \subset [\hat{y}-\bar{y}, \bar{y}-\hat{y}]$$

which implies with (7)

$$\|y\| \leqq \alpha \|y\|_Y \quad \text{for all } y \in Y.$$

In order to see that the norm $\|\cdot\|$ is strongly monotonically increasing on C_Y, observe that the norm is C-monotone on C, i.e.

$$\tilde{y} \in C, y \in (\{\tilde{y}\} - C) \cap C \Rightarrow \|y\| \leqq \|\tilde{y}\|.$$

For each $\tilde{y} \in C_Y \subset C$ and each $y \in (\{\tilde{y}\} - (C_Y \backslash \{0_Y\})) \cap C_Y$ we have with $C_Y \backslash \{0_Y\} \subset \operatorname{int}(C)$

$$\|y\| < \|\tilde{y}\|.$$

So, $\|\cdot\|$ is strongly monotonically increasing on C_Y. Finally, we prove that $\bar{y}$ is a unique solution of a certain approximation problem. Since $\bar{y}-\hat{y}$ belongs to the closure of the unit ball, we obtain $\|\bar{y}-\hat{y}\| = 1$. Furthermore, we assert that

$$(-C) \cap S = \{0_Y\}. \tag{8}$$

Because of the construction of the set U and the choice of $\beta \leqq (\varepsilon/2)$, respectively, for each $y \in S \backslash \{0_Y\}$ there exists some $\eta > 0$ with

$$N(y, \eta) \cap \operatorname{cone}(U) = \varnothing$$

which implies $y \notin \overline{\operatorname{cone}(U)} = C$. So $(-C) \cap (S \backslash \{0_Y\}) = \varnothing$ and the set equality (8) is evident. Moreover, with (8) and (7) we conclude

$$[\hat{y}-\bar{y}, \bar{y}-\hat{y}] \cap (\{\bar{y}-\hat{y}\} + S) = \{\bar{y}-\hat{y}\}$$

and

$$1 = \|\bar{y}-\hat{y}\| < \|\bar{y}-\hat{y}+y\| \quad \text{for all } y \in S \backslash \{0_Y\}.$$

Since $V - \{\bar{y}\} \subset S$, we get

$$1 = \|\bar{y}-\hat{y}\| < \|y-\hat{y}\| \quad \text{for all } y \in V \backslash \{\bar{y}\}.$$

This completes the proof. □

In the proof of the preceding theorem the cones $-C_Y$ and $\operatorname{cone}(T(V+C_Y, \bar{y}) \cup (V - \{\bar{y}\}))$ are "separated". For a finite dimensional linear space Henig [10, Thm. 2.1] presented a cone separation theorem. But in this abstract case we need stronger assumptions since Henig's proof depends on the finite dimensionality of the linear space. Here we use the notion of the base of a cone in order to obtain the desired result.

The assumption that the ordering cone C_Y has a weakly compact base implies the closedness of C_Y. In Theorem 2.4 we do not need the assumptions $\operatorname{cor}(C_Y) \neq \emptyset$ and $\hat{y} \in \{\bar{y}\} - \operatorname{cor}(C_Y)$ which play an important role in Theorem 2.2. On the other hand in Theorem 2.2 it is not required that $\bar{y}$ is uniquely determined by the inequality (2). With Theorem 2.2 and Theorem 2.4 we get immediately the main result of this paper.

COROLLARY 2.5. *Let $(Y, \|\cdot\|_Y)$ be a real normed space, and let C_Y be a convex cone in Y which has a nonempty core and a weakly compact base. Let V be a subset of Y, and let an element $\hat{y} \in Y$ with $V \subset \{\hat{y}\} + \operatorname{cor}(C_Y)$ be given. Further, for some $\bar{y} \in V$ let the cone generated by $T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})$ be weakly closed. Then $\bar{y}$ is a properly minimal element of the set V if and only if there exists an (additional) norm $\|\cdot\|$ on Y which is strongly monotonically increasing on C_Y, for which there exists some $\alpha > 0$ with*

$$\|y\| \leqq \alpha \|y\|_Y \quad \text{for all } y \in Y$$

and which has the property

$$1 = \|\bar{y} - \hat{y}\| < \|y - \hat{y}\| \quad \text{for all } v \in V \setminus \{\bar{y}\}.$$

In the preceding corollary we assume that the ordering cone C_Y has a weakly compact base B and a nonempty core. But this implies that the convex hull of the set $B \cup \{0_Y\}$ is also weakly compact and it has a nonempty core. For any element e of this core we define a norm $\|\cdot\|$ on Y with the aid of the order interval $[-e, e]$. Since this unit ball is weakly compact as well, the real normed space $(Y, \|\cdot\|)$ is even reflexive.

With the following propositions we give sufficient conditions under which various assumptions of Corollary 2.5 are fulfilled. The first proposition is standard.

PROPOSITION 2.6. *Let $(Y, \|\cdot\|_Y)$ be a reflexive Banach space with a closed ordering cone C_Y. The convex cone C_Y has a weakly compact base if and only if there exists a linear functional t with*

$$t(y) > 0 \quad \text{for all } y \in C_Y \setminus \{0_Y\}$$

such that the set $\{y \in C_Y \mid t(y) = 1\}$ is bounded.

PROPOSITION 2.7. *Let $(Y, \|\cdot\|_Y)$ be a real normed space, and let V and C_Y be nonempty subsets of Y with $0_Y \in C_Y$. If the set $V + C_Y$ is starshaped at some $\bar{y} \in V$ and the tangent cone $T(V + C_Y, \bar{y})$ is weakly closed, then the cone generated by $T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})$ is also weakly closed.*

Proof. Since the set $V + C_Y$ is starshaped at $\bar{y}$ we conclude

$$V - \{\bar{y}\} \subset V + C_Y - \{\bar{y}\} \subset T(V + C_Y, \bar{y}).$$

So, we obtain

$$\operatorname{cone}(T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})) = \operatorname{cone}(T(V + C_Y, \bar{y})) = T(V + C_Y, \bar{y})$$

which leads to the assertion. □

If the set $V + C_Y$ is starshaped at $\bar{y} \in V$, then Corollary 2.5 remains also valid for the notion of properly minimal elements in the sense of Benson. In this case the cone generated by the set $T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})$ can be replaced by the tangent cone $T(V + C_Y, \bar{y})$. With the following proposition we investigate the special case that $V + C_Y$ is a convex set.

PROPOSITION 2.8. *Let $(Y, \|\cdot\|_Y)$ be a real normed space, and let V and C_Y be nonempty subsets of Y with $0_Y \in C_Y$. If the set $V + C_Y$ is convex, then for each $\bar{y} \in V$ the cone generated by $T(V + C_Y, \bar{y}) \cup (V - \{\bar{y}\})$ is weakly closed.*

Proof. The tangent cone $T(V+C_Y, \bar{y})$ is closed and also convex because of the convexity of the set $V+C_Y$. Consequently, the tangent cone $T(V+C_Y, \bar{y})$ is weakly closed. Thus the assertion follows from Proposition 2.7. □

3. Conclusion. Properly minimal elements of an arbitrary set are characterized as minimal solutions of certain approximation problems. Even in the nonconvex case, it turns out that under suitable assumptions these approximation problems are equivalent to a vector optimization problem, if the notion of proper minimality is used. This shows the importance of the approximation theory in vector optimization.

Acknowledgment. The author gratefully acknowledges the helpful comments of the referees which improved the presentation of this paper.

REFERENCES

[1] H. P. BENSON, *An improved definition of proper efficiency for vector maximization with respect to cones*, J. Math. Anal. Appl., 71 (1979), pp. 232-241.
[2] H. P. BENSON AND T. L. MORIN, *The vector maximization problem: proper efficiency and stability*, SIAM J. Appl. Math., 32 (1977), pp. 64-72.
[3] J. BORWEIN, *Proper efficient points for maximizations with respect to cones*, this Journal, 15 (1977), pp. 57-63.
[4] ———, *The geometry of Pareto efficiency over cones*, Math. Oper. Statist., Ser. Optim., 11 (1980), pp. 235-248.
[5] J. M. BORWEIN AND J. W. NIEUWENHUIS, *Two kinds of normality in vector optimization*, Math. Programming, 28 (1984), pp. 185-191.
[6] W. DINKELBACH AND W. DÜRR, *Effizienzaussagen bei Ersatzprogrammen zum Vektormaximumproblem*, Oper. Res. Verfahren, XII (1972), pp. 69-77.
[7] A. M. GEOFFRION, *Proper efficiency and the theory of vector maximization*, J. Math. Anal. Appl., 22 (1968), pp. 618-630.
[8] R. HARTLEY, *On cone-efficiency, cone-convexity and cone-compactness*, SIAM J. Appl. Math., 34 (1978), pp. 211-222.
[9] M. I. HENIG, *Proper efficiency with respect to cones*, J. Optim. Theory Appl., 36 (1982), pp. 387-407.
[10] ———, *A cone separation theorem*, J. Optim. Theory Appl., 36 (1982), pp. 451-455.
[11] J. JAHN, *Duality in vector optimization*, Math. Programming, 25 (1983), pp. 343-353.
[12] ———, *Zur vektoriellen linearen Tschebyscheff-Approximation*, Math. Oper. Statis., Ser. Optim., 14 (1983), pp. 577-591.
[13] ———, *Scalarization in vector optimization*, Math. Programming, 29 (1984), pp. 203-218.
[14] W. KRABS, *Optimization and Approximation*, John Wiley, Chichester, 1979.
[15] H. W. KUHN AND A. W. TUCKER, *Nonlinear programming*, in Proc. Second Berkeley Symposium on Mathematical Statistics and Probability, J. Neyman, ed., Univ. California Press, Berkeley, 1951, pp. 481-492.
[16] J. W. NIEUWENHUIS, *Properly efficient and efficient solutions for vector maximization problems in Euclidean space*, J. Math. Anal. Appl., 84 (1981), pp. 311-317.
[17] A. L. PERESSINI, *Ordered Topological Vector Spaces*, Harper and Row, New York, 1967.
[18] W. VOGEL, *Vektoroptimierung in Produkträumen*, Verlag Anton Hain, Mathematical Systems in Economics 35, Meisenheim am Glan, 1977.
[19] R. E. WENDELL AND D. N. LEE, *Efficiency in multiple objective optimization problems*, Math. Programming, 12 (1977), pp. 406-414.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

A COMPUTATIONAL COMPARISON OF THE ELLIPSOID ALGORITHM WITH SEVERAL NONLINEAR PROGRAMMING ALGORITHMS*

J. G. ECKER† AND M. KUPFERSCHMID‡

Abstract. A computational comparison of several general purpose nonlinear programming algorithms is presented. This study was motivated by the preliminary results in [12] which show that the recently developed ellipsoid algorithm is competitive with a widely used augmented Lagrangian algorithm. To provide a better perspective on the value of ellipsoid algorithms in nonlinear programming, the present study includes some of the most highly regarded nonlinear programming algorithms and is a much more comprehensive study than [12]. The algorithms considered here are chosen from four distinct classes and 50 well-known test problems are used. The algorithms used represent augmented Lagrangian, ellipsoid, generalized reduced gradient, and iterative quadratic programming methods. Results regarding robustness and relative efficiency are presented.

Key words. nonlinear programming, algorithm evaluation, ellipsoid algorithm

1. Introduction. In [12], Ecker and Kupferschmid provide computational evidence that the ellipsoid algorithm is extremely robust and, relative to efficiency, is competitive with the augmented Lagrangian algorithm implemented in subroutine EO4VAF of the Mark 7 NAG Subroutine Library [16], hereafter referred to as NAG7. The results in [12] are surprising because for most of the test problems the ellipsoid algorithm was superior to NAG7 with regard to both the computer time and the number of function evaluations required to reduce the solution error to a certain level. In [13], the comparison was extended, for geometric programming problems, to include a special-purpose geometric programming algorithm and the general-purpose algorithms GRG2 and NAG8, described below, in addition to NAG7. That study confirmed the results of [12] with regard to the robustness of the ellipsoid algorithm and, among the general purpose algorithms, the ellipsoid algorithm was found to be most efficient at some levels of solution error.

To obtain a better perspective on the performance of the ellipsoid algorithm in nonlinear programming, a much more extensive computational study was warranted. The present study includes a wider selection of test problems than either of the earlier studies [12] and [13], and a wider selection of algorithms chosen from four distinct classes. To represent the ellipsoid, generalized reduced gradient, iterative quadratic programming, and augmented Lagrangian methods, we chose the following implementations:

EA3: The variant of the ellipsoid algorithm implemented by Kupferschmid and Ecker [27].

GRG2: The generalized reduced gradient algorithm of Lasdon, Waren, Jain and Ratner [28].

IQP: The Han-Powell [21], [30], iterative quadratic programming algorithm implemented by Crane, Hillstrom and Minkoff [8].

NAG8: The augmented Lagrangian algorithm of Gill and Murray [18], as implemented in subroutine EO4VAF of the Mark 8 NAG Subroutine Library [17].

* Received by the editors October 12, 1982, and in revised form June 15, 1984. This research was supported in part by National Science Foundation under grant MCS82-01790.

† Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181.

‡ Voorhees Computing Center, and Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York 12181.

RQP: The recursive (iterative) quadratic programming method of Biggs [4], as implemented in subroutine OPRQP of the Hatfield Subroutine Library [22].

In [33], Schittkowski performed a computational comparison of 26 nonlinear programming algorithms and ranked them according to various criteria. Relative to efficiency and reliability, the Schittkowski rankings of the above algorithms are as shown in Table 1.

TABLE 1
Schittkowski rankings.

Method	Efficiency	Reliability
GRG2	7	1
IQP	2	3 (see Remark 1)
NAG8	14	22 (see Remark 2)
RQP	3	7

Remark 1. IQP is a slightly different (and earlier) version of the Harwell subroutine VF02AD used in [33]. In [8], however, the authors state that their implementation "leaves Powell's original algorithm unchanged" but the program was made easier to use and incorporates some recently developed LINPACK subprograms. IQP uses the quadratic programming subroutine of Fletcher [15].

Remark 2. The routine SALQDR used in [33] is, according to the NAG documentation [17], essentially identical to the subroutine EO4VAF used here.

The above rankings are given here only to provide some evidence that the algorithms chosen for comparison with EA3 in this study include some of the better nonlinear programming algorithms. Of course, one can never claim that the version of an algorithm used in a particular study is the best or the most current version because of the continuing nature of algorithm development.

Because the ellipsoid algorithm is not as well-known as the other algorithms used in this study, we provide a brief, but detailed, description of the method. The algorithm was first proposed by Shor in [35] and it is a simple method for solving nonlinear programming problems of the form

$$\text{minimize } f_0(x)$$

$$\text{subject to } x \in S = \{x \in R^n \mid f_i(x) \leqq 0,\ i = 1, 2, \cdots, m\},$$

where each f_i is a convex function. The method assumes that there exists an optimal point $x^* \in S \neq \varnothing$ and that an initial ellipsoid

$$E_0 = \{x \in R^n \mid (x - x^0)^T Q_0^{-1} (x - x^0) \leqq 1\}, \tag{1}$$

centered at x^0 with a symmetric positive definite matrix Q_0, can be found such that $x^* \in E_0$ and such that $E_0 \cap S$ has computationally positive volume in R^n. The method is now known as the ellipsoid algorithm and the basic steps of the algorithm are as follows.

THE ELLIPSOID ALGORITHM.

Step 0. Select x^0 and a symmetric positive definite matrix Q_0 so that the ellipsoid E_0 in (1) contains an optimal point x^*.
Set $k = 0$.

Step 1. Let f_v be a constraint function for which $f_v(x^k) > 0$ or the objective function if $f_v(x^k) \leqq 0$, $i = 1, \cdots, m$.

Calculate a subgradient g_k of f_v at x^k. If $g_k = 0$, STOP because x^k is a minimizing point.
Calculate $d = -Q_k g_k / \sqrt{g_k^T Q_k g_k}$ provided $g_k^T Q_k g_k > 0$, otherwise STOP.

Step 2. Let $x^{k+1} = x^k + (1/(n+1))d$ and $Q_{k+1} = (n^2/(n^2-1))(Q_k - (2/(n+1))dd^T)$
Return to Step 1 with $k+1$ replacing k.

Starting with E_0, the algorithm generates a sequence of successively smaller ellipsoids so that if E_k denotes the current ellipsoid with matrix Q_k and center x^k, then the next ellipsoid has its center at x^{k+1} and a matrix Q_{k+1} as given in Step 2. The ratio q_n of the volume of E_{k+1} to the volume of E_k is given by

$$q = \left(\frac{n}{n+1}\right)\left(\frac{n^2}{n^2-1}\right)^{(n-1)/2} < 1$$

and thus the volumes decrease as a geometric series with ratio q_n depending on the dimension n [35].

Geometrically, the ellipsoid E_{k+1} is generated as follows. Using the subgradient g_k of Step 1, construct the hyperplane $g_k^T(x - x^k) = 0$ passing through the center x^k of E_k. Because each f_i is convex, x^* belongs to the half-space

$$H_k = \{x | -g_k^T(x - x^k) \geqq 0\}$$

and so $x^* \in E_k \cap H_k$. The ellipsoid E_{k+1} defined by the update formulae in Step 2 is actually the unique ellipsoid of minimum volume containing $E_k \cap H_k$ and so $x^* \in E_{k+1}$. For a complete derivation of these update formulae see the survey of ellipsoid algorithms given by Bland, Goldfarb and Todd [5].

In Step 1, if f_v is differentiable the subgradient g_k of f_v at x^k is the gradient and in the implementation of the ellipsoid algorithm used in this paper we normalize g_k in Step 1 by dividing it by the Chebyshev norm of g_k. Analytically, this does not alter the calculation of d.

In the above form, the ellipsoid algorithm stops with the current point x^k being optimal or when the current matrix Q_k is seen numerically to be not positive definite. Of course, the latter could be replaced by other standard convergence criteria but for the results of this paper we simply let the algorithm run until Q_k is no longer positive definite.

In [36], Shor shows for the convex unconstrained case that the best objective function value obtained in the first k iterations of the ellipsoid algorithm converges to the optimal value as the terms of a geometric series with a ratio q_n. Another similar convergence result is given by Goffin [19].

The test problems used in this study are chosen from several major collections of nonrandomly generated problems. A total of 50 test problems are used, including 25 geometric programming problems and 25 general nonlinear problems. The collection includes 13 convex and 37 nonconvex problems. Table 2 summarizes the geometric programming problems, and Table 3 summarizes the general problems.

All of the test problems have the form

$$\min f_0(x) \quad \text{subject to } f_i(x) \leqq 0, \qquad i = 1, 2, \cdots, m,$$

where $x \in R^n$, and we let x^0 denote the starting point for the algorithms.

After Eason and Fenton [11] we use error versus effort curves to display the convergence trajectories of the various algorithms. The error measure that we use here

TABLE 2
Geometric programming test problems.

Problem	Convex	n§	m	Reference
Dembo 1b	yes	12	3	[9]
Dembo 2†	no	5	9	[9]
Dembo 3	no	7	15*	[9]
Dembo 4a	no	8	4	[9]
Dembo 5	no	8	6	[9]
Dembo 6	no	13	18*	[9]
Dembo 7	no	16	25*	[9]
Dembo 8a	yes	7	4	[9]
RM 10‡	no	3	1	[32]
RM 11	no	4	2	[32]
RM 12	no	9	4	[32]
RM 13	no	9	6	[32]
RM 14	no	10	7	[32]
RM 15	no	10	7	[32]
RM 16	no	10	7	[32]
RM 17	no	11	9	[32]
RM 18	no	13	9	[32]
RM 19	no	9	5	[32]
RM 20	no	13	11	[32]
RM 21	no	10	22	[32]
RM 22	no	9	10	[32]
Beck 15	yes	8	7	[2]
Avriel	no	5	8	[1]
Smeers	yes	33	36	[38]
Tower	yes	36	33	[39]

* These include as constraints some explicit bounds on variables.

† The statement of this problem in [9] is imprecise and contains some typographical errors; see [25] for a correct problem statement.

‡ Here "RM" denotes a problem from Rijckaert and Martens.

§ n = number of variables, and m = number of constraints.

is computed as follows. First, the combined error measure

$$e(x^k) = |f_0(x^k) - f_0(x^*)| + \sum_{i=1}^{m} \lambda_i^* |f_i(x^k)|$$

is computed for each iterate x^k in the solution process, where the λ_i^* are the Lagrange multipliers at optimality and x^* denotes the optimal solution. These values are then normalized to obtain the relative error measure

$$E(x^k) = e(x^k)/e(x^0),$$

and the common logs of the $E(x^k)$ are plotted versus the measure of effort used so far. In [13], details regarding the calculation of the optimal Lagrange multipliers are given.

The measure of effort used for the error curves in this paper is the problem state central processing unit (PSCPU) time used by an algorithm. In [12] and [25], complete details are given regarding the determination of PSCPU time used by an algorithm in the solution process. In summary, we turn a timer on and off so as to measure only the effort used in performing the steps of the algorithm, thereby excluding from the measurements any time used for input and output operations, for other tasks performed

TABLE 3
General nonlinear test problems.*

Problem	Convex	n	m	Reference
Colville 1	yes	5	15	[6]
Colville 2	no	15	20	[6]
Colville 3	no	5	16	[6]
Colville 4	no	4	8	[6]
Colville 5†	no	6	4	[6]
Colville 8†	no	3	20	[6]
Himmelblau 3	no	2	7	[23]
Himmelblau 9	no	4	6	[23]
Himmelblau 12†	no	5	48	[23]
Himmelblau 13†	no	5	16	[23]
Himmelblau 16	no	6	14	[23]
Himmelblau 17	no	10	20	[23]
Himmelblau 21	no	6	6	[23]
Himmelblau 22	no	6	16	[23]
Himmelblau 24	no	2	2	[23]
BBZ 1‡	yes	3	3	[3]
BBZ 2	yes	5	7	[3]
BBZ 3	yes	16	13	[3]
Hald & Madsen 5	no	5	40	[20]
Swiss 5	no	2	5	[26]
Dixon	no	2	4	[10]
Hearn Dual	yes	2	2	[26]
Rosen-Suzuki	yes	4	3	[20]
Quad 2	yes	5	4	[14]
Shapiro 2	yes	10	1	[34]

* This set of general nonlinear problems contains all of the inequality constrained Colville problems and all of the inequality constrained Himmelblau problems except Himmelblau 23 (which has 100 variables and is therefore too big to conveniently handle using our test programs).

† These problems involve functions for which analytical derivatives cannot be given.

‡ BBZ denotes a problem from Ben-Israel, Ben-Tal, and Zlobec.

only as conveniences to the experimenter, and for the performance measurement process itself. Extensive experiments, see [25], have shown that our method of measuring PSCPU time is accurate, reproducible, and substantially uncontaminated by system-load effects and other influences external to the experiments.

The construction of meaningful error curves using the process described above requires the optimal solution to be known to considerably more precision than is usually reported in the literature. We therefore use very accurate solutions x^* in the construction of the error curves. These solutions are also chosen to be strictly feasible; see [26] for exact problem statements and the best strictly feasible point known to us for each of the test problems used in this study.

To supplement the error curves, we provide a tabular summary showing the overall performance of each algorithm on all of the problems. On some problems we also report the number of function and gradient evaluations used to reach the reasonable error level of 10^{-3}. When effort is reported in terms of function and gradient evaluations, we report a pair of numbers, (FE, GE), where FE is the number of times the objective function or a constraint is evaluated and GE is the number of times the gradient of the objective function or the gradient of a constraint is evaluated.

2. Experimental procedure. Given a test problem and a starting point, each algorithm was allowed to run until its best possible solution was obtained. The best strictly feasible point obtained by any of the algorithms was then declared as the optimal solution x^*. Most of the algorithms have parameters that can be adjusted to affect algorithm performance. Given an algorithm, we selected one set of parameters to be used for all of the test problems. Through a trial and error process involving several runs, one might optimize the performance of an algorithm on each problem. However, it is difficult to quantify the effort required for such a "tuning" process, and such tuning was not included as part of our experimental process. For GRG2, we used the parameters suggested by Lasdon et al. when they used GRG2 to solve the geometric programming test problems in [31]. Following the advice of Lasdon [29], these same parameters were also used for the general nonlinear programming problems. For NAG8 we used the parameter values suggested in the documentation [17], although we did try to find other parameters to improve its performance but were unsuccessful. For IQP and RQP, we also set the parameters as suggested in the documentation. These latter two algorithms solve a quadratic programming subproblem at each iteration and their overall performance is affected by the performance of this subproblem solver. We did not alter these algorithms by providing for a more efficient quadratic programming solver, although this is an area of current research.

The experiments reported below were conducted using an IBM System/370 Model 3033 computer with a type UO6 central processor. All calculations were done in IBM double-precision (with 56 fraction bits of accuracy). The programs were executed in an interactive multiprocessing environment with virtual memory under the Michigan Terminal System (MTS) operating program, release 5.OC. The computer has a 56 ns average cycle time and executes about 5.0 million machine instructions per second.

At each iteration of an algorithm, the following quantities are written in a file: the current iterate x^k, the current objective function value $f_0(x^k)$, the PSCPU time used so far by the algorithm, and the numbers of function and gradient evaluations used so far. After the experiment is over, this performance measurements file is used in analyzing the performance of the algorithm and in constructing the error versus effort curves.

For each test problem, we assume that a vector, x^u, of upper bounds and a vector, x^l, of lower bounds on the variables is known, and the starting point x^0 is chosen as the midpoint of these bounds. Many of the test problems have published upper and lower bounds and, when available, we use these bounds to generate the starting point. If published bounds are not available, then reasonably wide bounds are chosen so as to include the optimal vector. The ellipsoid algorithm requires that an initial ellipsoid E_0 be given which contains the optimal point, and we select E_0 as the ellipsoid of minimum volume containing

$$\{x \mid x^l \leqq x \leqq x^u\}.$$

In § 5, we consider the sensitivity of EA3 to the starting ellipsoids along with the sensitivity of the other algorithms to the starting points resulting from the chosen bounds.

In order to guarantee that each algorithm solves the same problems, the data necessary to define a particular problem is given in a single data structure accessed by each of the algorithms.

3. Results of the computational experiments. For some of the problems, convergence behavior of the various algorithms is displayed by means of the error versus

effort curves described above. Unfortunately, because of space limitations, error curves cannot be presented for all of the problems. Therefore, computational results for all of the problems are summarized in Tables 4 and 5. We discuss the error curves first.

Figures 1-3 give the error curves for each algorithm on three of the geometric programming problems. The convergence behavior displayed by the various algorithms in Fig. 1 is rather typical of their behavior on most of the geometric programming problems. In particular, the error curves for Dembo 3 in Fig. 1 show EA3 to be the most efficient in reducing the combined relative error to the 10^{-1} and 10^{-2} range, and for error levels in the 10^{-3} to 10^{-8} range, GRG2 and IQP are the most efficient. On Dembo 3, the behavior of the latter two algorithms is nearly identical down to the 10^{-6} error level and then GRG2 stops making any further reduction in the error while IQP succeeds in eventually reducing the error to about 10^{-11}. The error curve for RQP shows only a negligible reduction in the error for about 3 seconds and then a sudden reduction to the 10^{-14} error level. NAG8 is unsuccessful on Dembo 3 and converges to a point with an objective function value of about 1,388 instead of the true optimal value of about 1,227, and the point to which it converges, before terminating with an exponent overflow in one of its internal routines, is not a Kuhn-Tucker point.

In Fig. 1 the error curve for EA3 departs from its linear trend at the error level of about 10^{-14} and this is typical of the end behavior in EA3's convergence trajectories. The elements of the positive-definite matrix defining the current ellipsoid are extremely small (usually less than 10^{-20}) when this occurs and this is the only evidence of numerical instability that we have observed for EA3.

Dembo 7 is well-known as a very difficult problem [31], partly because the feasible region is extremely small. When Dembo's original bounds [9] are used, EA3 converges

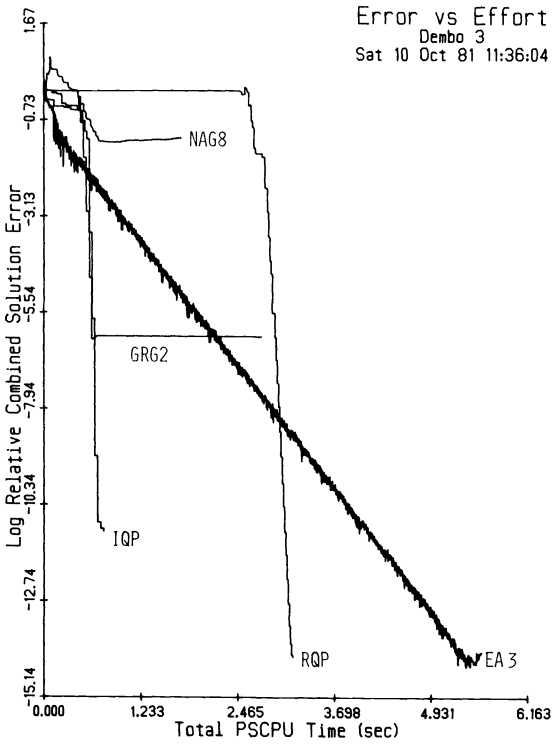


FIG. 1

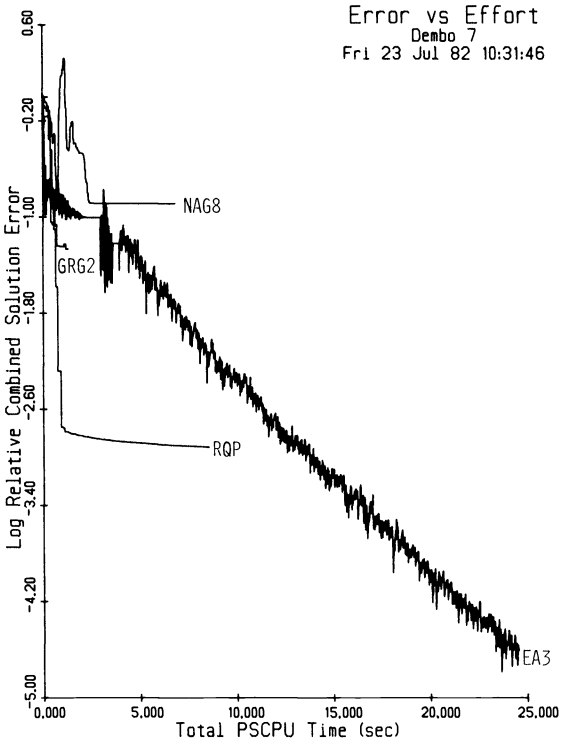

Error vs Effort
Dembo 7
Fri 23 Jul 82 10:31:46
Log Relative Combined Solution Error
0.60
-0.20
-1.00
-1.80
-2.60
-3.40
-4.20
-5.00
NAG8
GRG2
RQP
EA3
0.000
5.000
10.000
15.000
20.000
25.000
Total PSCPU Time (sec)

FIG. 2

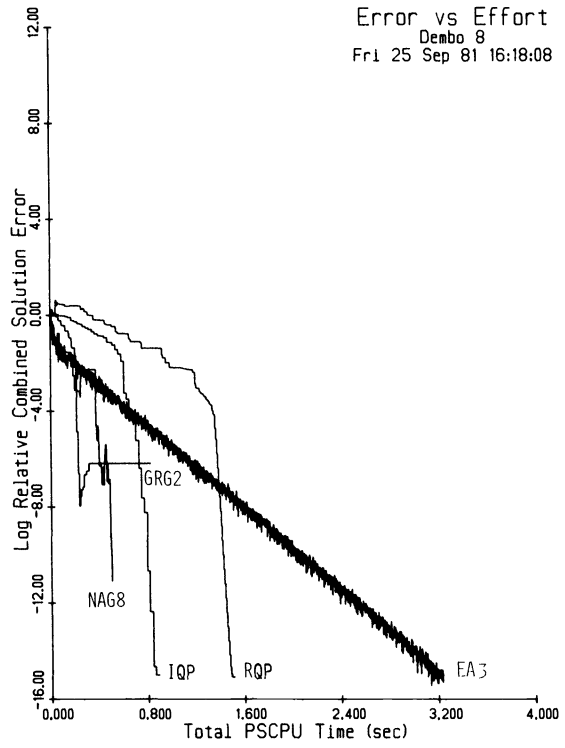

Error vs Effort
Dembo 8
Fri 25 Sep 81 16:18:08
Log Relative Combined Solution Error
12.00
8.00
4.00
0.00
-4.00
-8.00
-12.00
-16.00
GRG2
NAG8
IQP
RQP
EA3
0.000
0.800
1.600
2.400
3.200
4.000
Total PSCPU Time (sec)

FIG. 3

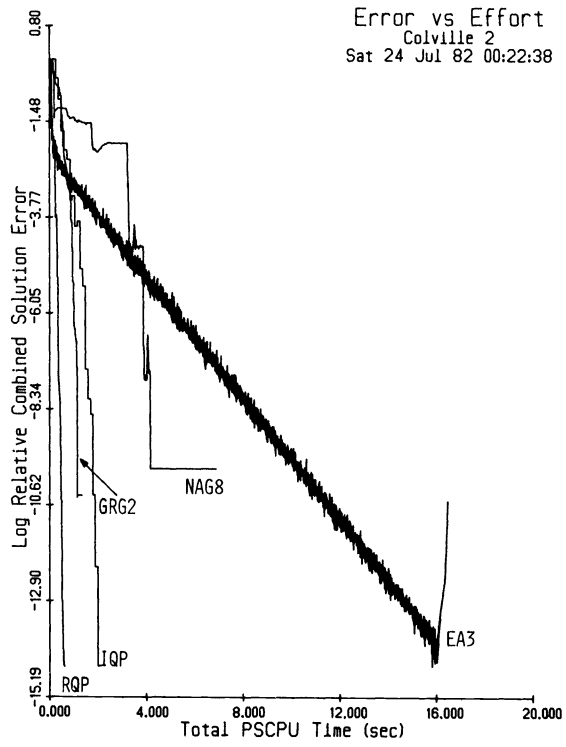


FIG. 4

to a nonstationary point and none of the other algorithms solve the problem either. We therefore use tighter bounds [26] for all of the algorithms. Even with the tighter bounds, the error curves in Fig. 3 show that the only algorithm to make any significant progress is EA3 (the data for EA3 were truncated for plotting, but EA3 ultimately reduces the solution error to about 10^{-7}). No error curve is visible for IQP because it terminates on the second iteration with a floating point overflow in its internal routine HARWQP.

It is interesting to observe the distinct linear trend of the convergence trajectories for EA3 in these figures. Roughly speaking, EA3 makes consistent, but slow, progress independent of its closeness to the solution. In contrast, as displayed nicely in Fig. 3, once the other algorithms get close to the solution there is usually an acceleration of the convergence process. Another observation regarding the linear convergence trend displayed by EA3 is that the vertical axis intercept of the linear trend is usually -1 or less.

Figures 4–6 contain error curves for some of the general problems and again EA3 displays relatively "slow but sure" convergence. Again, however, at error levels of 10^{-1} and 10^{-2}, EA3 is more efficient than the other algorithms.

The departure of EA3's error curves from their linear trend at extremely low error levels is most visible in Figs. 4 and 5. Part of the departure in Fig. 5 is due to the fact that Colville 8 does not have analytical derivatives and near the end of EA3's trajectory, when the ellipsoids are extremely small, the numerically approximated gradients are not sufficiently accurate. This problem also appeared to cause some instability in GRG2.

The last problem for which error curves are given is Hald 5 (see Fig. 6) which is a minimax optimization problem with 5 variables and 40 constraints. Here the ellipsoid

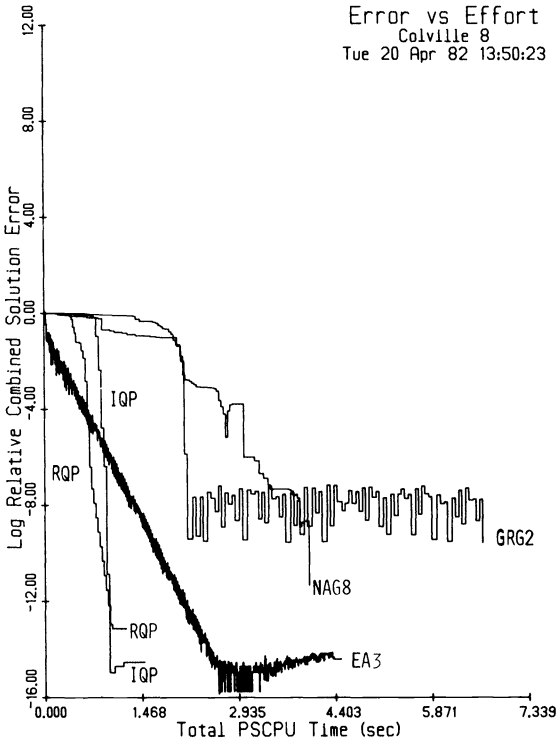

Error vs Effort
Colville 8
Tue 20 Apr 82 13:50:23
Log Relative Combined Solution Error
12.00
8.00
4.00
0.00
-4.00
-8.00
-12.00
-16.00
IQP
RQP
GRG2
NAG8
RQP
EA3
IQP
0.000
1.468
2.935
4.403
5.871
7.339
Total PSCPU Time (sec)

FIG. 5

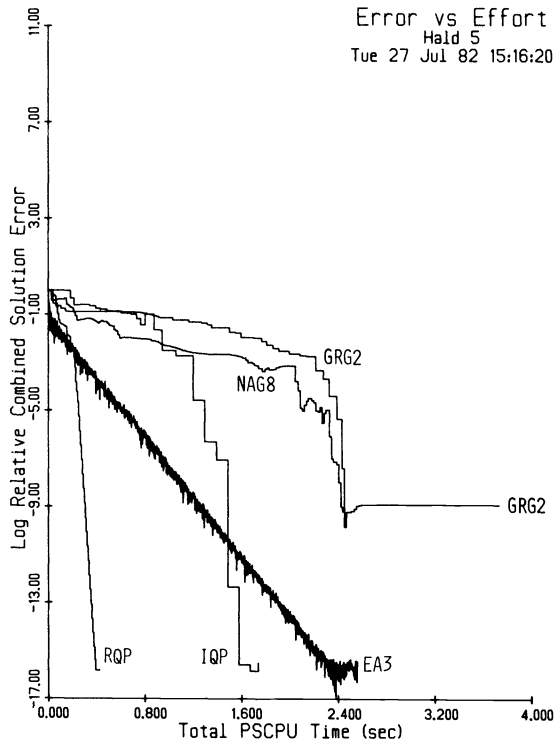

Error vs Effort
Hald 5
Tue 27 Jul 82 15:16:20
Log Relative Combined Solution Error
11.00
7.00
3.00
-1.00
-5.00
-9.00
-13.00
-17.00
GRG2
NAG8
GRG2
RQP
IQP
EA3
0.000
0.800
1.600
2.400
3.200
4.000
Total PSCPU Time (sec)

FIG. 6

algorithm performs remarkably better than most of the other algorithms all the way down to the 10^{-9} error level. This is in keeping with the fact that the volume reduction ratio q_n of the ellipsoid algorithm depends only on the number of variables and not on the number of constraints.

Most of the test problems used in this study involve functions that are relatively easy to evaluate. The one exception is the problem Shapiro 2. Evaluating the objective function and the single constraint for this problem requires considerable effort. Evaluating the objective function or its gradient requires the solution of a Lyapunov equation of order 6. Evaluating the constraint requires finding all of the eigenvalues of a 6 by 6 matrix, and using finite differencing $n+1$ function evaluations are required to numerically determine the constraint gradient. After the first few iterations the stability constraint is inactive and so its gradient is not needed. As in most of the problems considered above, EA3 is again first in reducing the error to 10^{-1} and 10^{-2}, and ultimately obtains the best solution. GRG2 is the only other algorithm that has any success on this problem. NAG8 and RQP terminate at points with relative errors of about 10^{+14} and 10^{+9} respectively while IQP terminates after a few iterations with the message that the quadratic programming subproblem is infeasible. It is interesting to note that when the starting point is chosen to be extremely near the true solution then these algorithms are successful in solving the problem.

4. General discussion of results. To supplement the error curves and to summarize the results for all 50 problems, we have constructed tables which show the efficiency and robustness of the algorithms for the error levels 10^{-1} through 10^{-8}.

The results reported in Table 4 show that EA3 is by far the most efficient of the algorithms in reducing the relative solution error to the 10^{-1} and 10^{-2} levels. However,

TABLE 4

Efficiency at various error levels for all 50 test problems. Fraction of test problems solved first.

	Solution error level required*							
Method	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
EA3	.66*	.44*	.18	.22	.16	.12	.14	.12
GRG2	.14	.12	.26	.26	.32*	.36*	.28*	.14
IQP	.00	.08	.14	.16	.16	.18	.22	.28
NAG8	.08	.10	.10	.06	.08	.08	.08	.10
RQP	.12	.26	.30*	.28*	.26	.24	.24	.30*

* Denotes best performance at this error level.

TABLE 5

Robustness at various error levels for all 50 test problems. Fraction of test problems solved.

	Solution error level required							
Method	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
EA3	.98*	.98*	.96*	.96*	.96*	.96*	.94*	.92*
GRG2	.90	.80	.76	.76	.74	.74	.58	.30
IQP	.80	.76	.76	.76	.76	.76	.76	.76
NAG8	.80	.72	.64	.54	.50	.50	.48	.40
RQP	.74	.68	.66	.60	.60	.60	.58	.56

* Denotes the most robust algorithm at this error level.

at smaller levels of error, the performance of both GRG2 and RQP is superior to that of EA3 and the other algorithms.

The results regarding robustness of the algorithms are summarized in Table 5. In constructing this table, we simply record whether or not the algorithms reduce the relative error to the given error levels. As mentioned in § 2, the relative error is defined using a declared solution to each problem. One thing that does not show in these tables is whether or not the failure of an algorithm to reach a given error level is due to the fact that it converges to a local minimum with a larger objective function value than the declared solution. As part of our experimental process, we routinely check to see if a point to which an algorithm converges is a Kuhn-Tucker point, and for the 50 test problems, there are only 6 for which some algorithm converges to such a point with an objective value different from the declared solution. Those details are not included here and would not alter any of the qualitative behavior that is observed.

TABLE 6
Function and gradient evaluations required to reach the 10^{-3} solution error level.

Method	Iter*	Time†	FE's	GE's	Problem
EA3	203	.105	1,878	203	
GRG2	7	.030	528	112	
IQP	3	.054	64	64	Colville 1
NAG8	61	.218	992	992	
RQP	7	.019	144	144	
EA3	482	.759	4,424	482	
GRG2	38	.773	10,647	798	
IQP	7	.857	147	147	Colville 2
NAG8	518	3.236	12,369	12,327	
RQP	15	.231	336	315	
EA3	230	.120	1,930	230	
GRG2	7	.035	799	119	
IQP	3	.043	51	51	Colville 3
NAG8	66	.238	1,547	1,071	
RQP	9	.027	170	153	
EA3	138	.058	1,242	138	
GRG2	10	.024	509	90	
IQP	13	.140	189	189	Colville 4
NAG8	12	.040	252	243	
RQP	22	.035	666	198	
EA3	53	.426	190	53	
GRG2	6	.497	480	30	
IQP	9	.610	70	70	Colville 5
NAG8	‡	‡	‡	‡	
RQP	‡	‡	‡	‡	
EA3	60	.312	865	59	
GRG2	25	2.118	5,754	525	
IQP	18	.862	378	378	Colville 8
NAG8	61	2.305	1,470	1,428	
RQP	22	.651	672	462	

* Number of iterations to attain the 10^{-3} error level.

† PSCPU time in seconds.

‡ For this problem the results are for the 10^{-1} error level because no algorithm attained the 10^{-3} level and NAG8 and RQP did not even attain the 10^{-1} error level.

We conclude this section with some information regarding the number of function and gradient evaluations required by the algorithms to reach the relative error level of 10^{-3}. We present results for a subset of the problems, namely the 6 Colville and the 8 Dembo problems. In Tables 6-7, we give the number of function evaluations (FE's) and the number of gradient evaluations (GE's) that are required by the various algorithms to reduce the relative solution error to the 10^{-3} error level. On this subset of test problems, it is clear that IQP is vastly superior to all of the other algorithms with regard to function and gradient evaluations.

TABLE 7
Function and gradient evaluations required to reach the 10^{-3} solution error level.

Method	Iter*	Time*	FE's	GE's	Problem
EA3	165	.979	406	164	
GRG2	11	.380	568	44	
IQP	14	.798	64	64	Dembo 1
NAG8	289	1.532	1,280	1,272	
RQP	21	.451	152	84	
EA3	254	.290	1,693	253	
GRG2	5	.063	480	50	
IQP	3	.037	30	30	Dembo 2
NAG8	92	.255	1,140	790	
RQP	9	.055	100	90	
EA3	442	.838	3,387	441	
GRG2	17	.579	3,728	272	
IQP	13	.536	256	256	Dembo 3
NAG8	†	†	†	†	
RQP	174	2.857	5,440	2,784	
EA3	404	.670	1,178	403	
GRG2	37	.517	2,570	185	
IQP	8	.215	40	40	Dembo 4
NAG8	†	†	†	†	
RQP	969	7.828	9,752	4,845	
EA3	467	.790	1,907	466	
GRG2	51	.820	4,270	357	
IQP	22	.729	287	287	Dembo 5
NAG8	†	†	†	†	
RQP	29	.265	252	203	
EA3	172	.706	1,271	171	
GRG2	27	1.480	6,251	513	
IQP	44	6.140	1,520	1,520	Dembo 6
NAG8	67	.494	1,520	1,311	
RQP	†	†	†	†	
EA3	2,219	13.559	26,462	2,218	
	all other algorithms failed to attain 10^{-3} error level				Dembo 7
EA3	215	.343	614	214	
GRG2	15	.199	910	75	
IQP	25	.606	265	265	Dembo 8
NAG8	84	.215	415	410	
RQP	149	1.201	1,705	745	

* See the footnotes to Table 6.
† The error level of 10^{-3} was not attained.

It is interesting to note from Fig. 4 that EA3, GRG2, and IQP take about the same PSCPU time to reduce the error to 10^{-3} (the precise times are .76, .77, and .86 seconds, respectively), but Table 6 shows that IQP requires vastly fewer numbers of function and gradient evaluations to reach the same error level. The reason for this is that during this same time IQP has a considerable amount of other work that it needs to perform. For example, in the 7 iterations to reach the 10^{-3} error level, IQP solves 7 quadratic programming subproblems to determine search directions and then performs the subsequent line searches. On this problem, the effort required by IQP for this other work accounts for the vast majority of its effort. Clearly, if the time for this other work were negligible relative to the function and gradient evaluations, then IQP would have reached the 10^{-3} error level in much less *time* than all the other algorithms. However, as we discussed above, this study involves problems with functions that are relatively easy to evaluate even though they are, for the most part, taken from real applications and are representative of a large class of nonlinear programming problems. We should note, however, that even for problems involving functions that take a considerable amount of time to evaluate (such as functions defined by computer simulations, differential equation solvers, etc.), PSCPU time is a reasonable measure of effort.

5. Additional discussion of ellipsoid algorithm performance. Before summarizing our conclusions, we consider some additional features of the ellipsoid algorithm that we have observed in the course of our experimental work for this paper and for the experiments reported in [25], [12], and [13].

First, we have observed that the qualitative behavior of the ellipsoid algorithm is relatively insensitive to the size of the starting ellipsoid. Increasing the size of the initial ellipsoid does, of course, alter the sequence of iterates x^k, and the convergence trajectory followed by EA3 is dependent on the starting ellipsoids.

In order to obtain further insight into the sensitivity of EA3 to the starting ellipsoids, as well as the sensitivity of the other algorithms to the starting points, the following experiment was performed. The test problems, Colville 1, 2, 3, 4 and Dembo 2, 4, 6, 8 were arbitrarily chosen and the original upper bounds, used for the results in § 4, were systematically increased but the lower bounds were left unchanged. In particular, these 8 problems were rerun using a new upper bound on each variable that is 10 times the original upper bound, if the upper bound is positive, and 1/10 the upper bound, if it is negative. This process was then repeated with 100 replacing 10. The results of this experiment are summarized below:

(i) For each of these 8 problems, the time required by EA3 to reduce the relative error to each given error level consistently decreases as the bounds are enlarged. These results are consistent with the way in which an ellipsoid algorithm works, and a plausible explanation of this phenomenon is that the larger the starting ellipsoid, the more widely dispersed are the iterates generated. Since the time required for a single EA3 iteration is typically less than the time required for an iteration of the other algorithms, it typically takes a shorter time for EA3 to "sample" widely dispersed points. If the initial absolute error is large, then it is likely that initially EA3 will reduce the relative error more quickly than the other algorithms. In summary, enlarging the bounds does not adversely affect the qualitative behavior displayed by EA3 in Table 4.

(ii) With regard to robustness, the other algorithms appear to be far more sensitive to the starting point (the midpoint of the enlarged bounds) than EA3 is to the size of the starting ellipsoids. Increasing the bounds, and thus the distance between the starting

point and the optimal point, did not affect the robustness of EA3 but it dramatically worsened the robustness of the other algorithms.

Another ellipsoid algorithm phenomenon that we have observed regards local minima. From the beginning of our experimental work, we have observed that on problems with several local minima the ellipsoid algorithm often converges to a global minimum. For example, as we report in [12], on Himmelblau 22 EA3 finds a strictly feasible point with an objective function value of about 3.13 whereas the previously published minimum value for this problem is 4.07 (see [24], [7] and [23]).

To investigate this phenomenon further, the following experiment was performed. The notorious "six hump camel back" function of Dixon et al. [10], was chosen and the five algorithms of our study were run using several different sets of bounds. A contour plot of this two variable problem is given in Fig. 7, and within the bounds

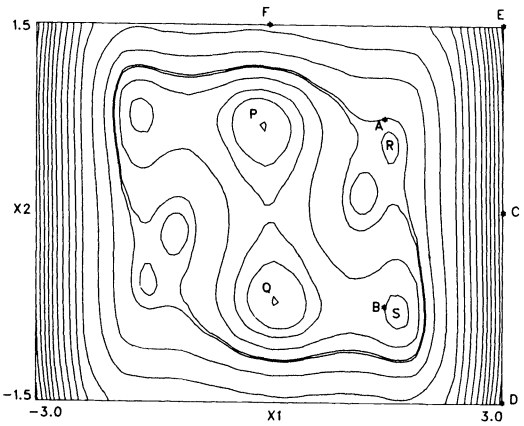


FIG. 7. *The six hump camel problem* [10].

$[-3, +3]$ for x_1 and $[-1.5, +1.5]$ for x_2 there are six local minima, 2 local maxima, and 7 saddle points. Two of the local minimizing points, P and Q, have the same objective function value of -1.032 which is the global minimum value. The local minimizing point S has an objective function of about $-.2155$ and the local minimizing point R has an objective function value of about 2.104.

Six starting points labeled A through F (see Table 8) were generated as the midpoints of bounds chosen so that two of the original (four) bounds are part of each

TABLE 8
Points converged to on the six hump camel problem.

Starting pt.	EA3	GRG2	IQP	NAG8	RQP
A	*Q*	*R*	*R*	*R*	*R*
B	*Q*	*S*	*S*	*S*	*S*
C	*P*	*Q*	*Q*	*S*	*R*
D	*P*	*P*	*	*R*	*R*
E	*Q*	*S*	*	*S*	*S*
F	*P*	*P*	*Q*	*P*	*P*

* For these bounds, IQP converges to the saddle point at 0, 0.

set of bounds and so that the new bounded region contains the original region. (Given the starting points, this uniquely determines the bounds.) The function is symmetric to the diagonals and so starting points were only chosen on one side. The results of this experiment are given in Table 8.

For all of these starting points, EA3 converges to a global minimizing point. Of course, EA3 does not always converge to such a point on this problem. For example, if upper bounds of 1,000 and 500 and lower bounds of -100 and -50 are chosen, then EA3 converges to a local minimum.

The last ellipsoid algorithm phenomenon discussed here concerns the ability of EA3 to sometimes converge to an optimal point that is far outside the starting ellipsoid. We have observed this happen on several problems but only one will be discussed here to illustrate the point. The test problem Hald & Madsen 5 has an optimal value of about 115.7 and an optimal point x^* with components given approximately by

$$(-12.2, 14.0, -.451, -.010, 115.7).$$

For the results of § 4, upper bounds of 20 and lower bounds of -20 were chosen for the first four variables and an upper bound of 150 and a lower bound of 50 was chosen for x_5. If upper bounds of 5 and lower bounds of -5 are chosen for the variables x_1 through x_4, and the bounds on x_5 left unchanged so x^* is far from being in the initial ellipsoid, then EA3 converges to a point z that is essentially the same as x^* in that

$$\|x^* - z\| < 10^{-6}.$$

In part, it is the ability of the ellipsoid algorithm to generate highly aspheric ellipsoids during the solution process that accounts for this phenomenon.

6. Observations and conclusions. Based on the results reported in this paper, we draw the following conclusions.

(1) Of all the algorithms studied, EA3 is clearly the most robust. GRG2 and IQP are also reasonably robust for most levels of solution error but NAG8 and RQP are not.

(2) Relative to computational efficiency, EA3 is superior to all the other algorithms at solution error levels down to 10^{-2}. For relative error between 10^{-3} and 10^{-8}, RQP and GRG2 are the most efficient. Typically, EA3 ultimately gets the best solution.

(3) In terms of using the fewest number of function and gradient evaluations to reach a given error level, IQP is vastly superior to all of the other algorithms.

We observed the following features regarding the performance of EA3.

- Except at the beginning of its trajectory, EA3 displays the linear convergence that is predicted theoretically. Its initial trajectory is usually convex, so that the linear trend intersects the error axis below zero (typically below -1).
- EA3 remains numerically stable until extremely small error levels have been attained.
- The qualitative behavior of EA3 is insensitive to the size of the starting ellipsoid.
- Often, EA3 finds global minima and avoids local minima.
- It is not always necessary for the initial ellipsoid to contain the optimal point in order for EA3 to converge to the optimal point.

A final comparison between the algorithms considered in this study concerns their complexity and the complexity of the computer programs required to implement them. The ellipsoid algorithm is very simple, whereas all of the other algorithms are relatively much more complicated. The magnitude of the contrast can be seen from the following statistics on the number of lines of executable Fortran code contained in the

implementations:

GRG2	2,495
IQP	1,007
RQP	475
EA3	61

The NAG8 code is proprietary and we therefore could not examine it to count lines, but it consists of 54 separate subroutines.

In view of the above conclusions and observations, it appears that, in addition to being used as a stand-alone method, the ellipsoid algorithm could also be used to provide good starting points for algorithms with better convergence rates near optimality.

REFERENCES

[1] M. AVRIEL AND J. D. BARRETT, *Optimal design of pitched laminated wood beams*, Advances in Geometric Programming, M. Avriel, ed. Plenum Press, New York, New York, pp. 407-419, 1980.

[2] P. A. BECK AND J. G. ECKER, *Some computational experience with a modified convex simplex algorithm for geometric programming*, Report ADTC-72-20, Armament Development and Test Center, USAF-Systems Command, Eglin AFB, FL, 1972.

[3] A. BEN-ISRAEL, A. BEN-TAL AND S. ZLOBEC, *Optimality in Nonlinear Programming*, John Wiley, New York, 1981.

[4] M. C. BIGGS, *Constrained minimization using recursive quadratic programming*, Toward Global Optimization, L. C. W. Dixon and G. P. Szëgo, eds., North-Holland, Amsterdam, 1975, pp. 341-349.

[5] R. G. BLAND, D. GOLDFARB AND M. J. TODD, *The ellipsoid method: a survey*, Oper. Res., 29 (1981), pp. 1039-1091.

[6] A. R. COLVILLE, *A comparative study of nonlinear programming codes*, IBM New York Scientific Center Report 320-2949, International Business Machines Corporation, New York, 1968.

[7] L. W. CORNWELL, P. A. HUTCHISON, M. MINKOFF AND H. K. SCHULTZ, *Test problems for constrained nonlinear mathematical programming algorithms*, Applied Mathematics Division Technical Memorandum 320, Argonne National Laboratory, Argonne, IL, 1978.

[8] R. L. CRANE, K. E. HILLSTROM AND M. MINKOFF, *Solution of the general nonlinear programming problem with subroutine* VMCON, Argonne National Laboratory, Report No. ANL-80-64, Argonne, IL, 1980.

[9] R. S. DEMBO, *A set of geometric programming test problems and their solutions*, Math. Programming, 10 (1976), pp. 192-213.

[10] L. C. W. DIXON, J. GOMULKA AND S. E. HERSOM, *Reflections on the global optimization problem*, Optimization in Action, L. C. W. Dixon, ed., Academic Press, London, 1976, pp. 398-435.

[11] E. D. EASON AND R. G. FENTON, *Testing and evaluation of numerical methods for design optimization*, UTME-TP 7204, Univ. Toronto, Toronto, Ontario, Canada, 1972.

[12] J. G. ECKER AND M. KUPFERSCHMID, *An ellipsoid algorithm for nonlinear programming*, Math. Programming, 27 (1983), pp. 83-106.

[13] J. G. ECKER, M. KUPFERSCHMID AND R. S. SACHER, *Comparison of a special purpose algorithm with general purpose algorithms for solving geometric programming problems*, J. Optim. Theory Appl., 43 (1984), pp. 237-263.

[14] J. G. ECKER AND R. D. NIEMI, *A dual method for quadratic programs with quadratic constraints*, SIAM J. Appl. Math., 28 (1975), pp. 568-576.

[15] R. FLETCHER, *A Fortran program for general quadratic programming*, Report No. R6370, Atomic Energy Research Establishment, Harwell, Berkshire, UK, 1970.

[16] L. FOX AND J. H. WILKINSON, NAG *Fortran Library Manual*, Mark 7, Numerical Algorithms Group Ltd., Oxford, England, 1978.

[17] ———, NAG *Fortran Library Manual* Mark 8, Numerical Algorithms Group Ltd., Oxford, England, 1981.

[18] P. E. GILL AND W. MURRAY, *Numerical Methods for Constrained Minimization*, Academic Press, New York, 1974.

[19] J. L. GOFFIN, *Convergence rates of the ellipsoid method on general convex functions*, Math. Oper. Res., 8 (1983), pp. 135-150.

[20] J. HALD AND K. MADSEN, *Methods for minimax optimization*, Math. Programming, 20 (1981), pp. 49–62.
[21] S.-P. HAN, *A globally convergent method for nonlinear programming*, J. Optim. Theory Appl., 22 (1977), pp. 297–309.
[22] Hatfield Subroutine Library, *The Optima User's Manual*, Numerical Optimisation Centre, Hatfield Polytechnic Institute, Hatfield, Hertfordshire, UK, 1976.
[23] D. M. HIMMELBLAU, *Applied Nonlinear Programming*, McGraw-Hill, New York, 1972.
[24] A. HOLZMAN, SRCC Report 113, U. S. Steel Company, Pittsburgh, PA, 1969.
[25] M. KUPFERSCHMID, *An ellipsoid algorithm for convex programming*, Ph.D. Dissertation, Rensselaer Polytechnic Institute, Troy, NY, 1981.
[26] M. KUPFERSCHMID AND J. G. ECKER, *Test problems for nonlinear programming*, Report VCC-82-1, Voorhees Computing Center, Rensselaer Polytechnic Institute, Troy, NY, 1982.
[27] ———, EA3: *A practical implementation of an ellipsoid algorithm for nonlinear programming*, Dept. Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY, 1984.
[28] L. S. LASDON, A. WAREN, A. JAIN AND M. W. RATNER, *Design and testing of a generalized reduced gradient code for nonlinear programming*, Act Trans. Math. Software, 4 (1978), pp. 34–50.
[29] L. S. LASDON, private communication.
[30] M. J. D. POWELL, *Algorithms for nonlinear constraints that use Lagrangian functions*, Math. Programming, 14 (1978), pp. 224–248.
[31] M. RATNER, L. S. LASDON AND A. JAIN, *Solving geometric programs using* GRG: *results and comparisons*, J. Optim. Theory and Appl., 26 (1978), pp. 253–264.
[32] M. J. RIJCKAERT AND X. M. MARTENS, *Comparison of generalized geometric programming algorithms*, Advances in Geometric Programming, M. Avriel, ed., Plenum Press, New York and London, 1980, pp. 283–320.
[33] K. SCHITTKOWSKI, *Nonlinear Programming Codes: Information, Tests, Performance*, Springer-Verlag, New York, 1980.
[34] E. Y. SHAPIRO, D. A. FREDERICKS AND R. H. ROONEY, *Suboptimal constant output feedback and its application to modern flight control system design*, Intern. J. Control, 33 (1981), pp. 505–517.
[35] N. Z. SHOR, *Cut-off method with space extension in convex programming problems*, Cybernetics, 12 (1977), pp. 94–96.
[36] ———, *New development trends in nondifferentiable optimization*, Cybernetics, 13 (1977), pp. 881–886.
[37] N. Z. SHOR AND V. I. GERSHOVICH, *Family of algorithms for solving convex programming problems*, Cybernetics, 15 (1980), pp. 502–508.
[38] Y. SMEERS AND D. TYTECA, *A geometric programming model for the optimal design of wastewater treatment plants*, Oper. Res., 32 (1984), pp. 314–342.
[39] R. D. WIEBKING, *Deterministic and stochastic geometric programming models for optimal engineering design problems in electric power generation and computer solutions*, Technical Report TR73LS132, General Electric Company, Schenectady, NY, 1974.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

003

STOCHASTIC APPROXIMATIONS VIA LARGE DEVIATIONS: ASYMPTOTIC PROPERTIES*

PAUL DUPUIS† AND HAROLD J. KUSHNER‡

Abstract. Asymptotic properties of Robbins-Munro and Kiefer-Wolfowitz type stochastic approximation algorithms are obtained via the theory of large deviations. The conditions are weak and can even yield w.p.l. convergence results. The probability of escape of the iterates from a neighborhood of a stable point of the algorithm is estimated and shown to be considerably smaller than suggested by the classical "asymptotic normality of local normalized errors" method of getting the asymptotic properties. The escape probabilities are a natural quantity of interest. In many applications, they are more useful than the "local normalized mean square errors." Other large deviations estimates are also obtained. Typically, if $a_n = 1/n^\rho$, $\rho \leqq 1$, then the probability of escape from a neighborhood of a stable point in some (normalized) time interval $[n, m]$: $\sum_n^m a_i \sim T$ is $\exp - n^\rho V_\rho$, where V_ρ does not depend on ρ for $\rho < 1$ and is the solution to an optimal control problem. If the noise is Gaussian, then the optimal control problem is relatively easy. Under quite broad conditions, in the Kiefer-Wolfowitz case the control problem has the Gaussian form, whether or not the noise is Gaussian. The techniques are expected to be quite useful in the analysis of the asymptotic properties of recursive algorithms generally.

Key words. stochastic approximation, large deviations, recursive algorithms, asymptotic properties

1. Introduction. This paper concerns a useful approach to the asymptotic behavior of stochastic approximation algorithms (SA) of the Robbins-Munro (RM) type

$$X_{n+1} = X_n + a_n \bar{b}(X_n) + a_n b(X_n, \xi_n), \qquad a_n \geqq 0, \quad \sum a_n = \infty, \quad a_n \to 0, \tag{1.1}$$

or the Kiefer-Wolfowitz (KW) type

$$X_{n+1} = X_n + a_n \bar{b}(X_n) + a_n b(X_n, \xi_n)/c_n, \tag{1.2}$$

where $Eb(x, \xi_n) \equiv 0$ and $0 < c_n \to 0$, and $x \in R^r$, Euclidean r-space.

Let θ denote a stable point of

$$\dot{x} = \bar{b}(x). \tag{1.3}$$

Under quite broad conditions $X_n \to \theta$ w.p.l. (at least if X_n is in a neighborhood of θ often enough or if θ is globally asymptotically stable). The classical rate of convergence theory (for (1.1), for example) [1], [2], concerns asymptotic normality of $\{(X_n - \theta)/\sqrt{a_n}\}$. It is a "local" result, and the only "dynamical" information which it uses is the gradient matrix $\bar{b}_x(\theta)$. This asymptotic result is useful, but does not provide enough information, and does not fully exploit the dynamical properties.

Here, we take an alternative point of view. Let G denote a bounded open set containing θ, and let P_x denote probability given that $X_n = x$. If $X_n \to \theta$ w.p.l., then the probability that the tail $\{X_j, j \geqq n\}$ leaves G goes to zero as $n \to \infty$. The dependence of this rate on $\bar{b}(\cdot)$ and the other data of the problem is of interest. Suitably normalized estimates can yield considerable information, and provide a useful and informative alternative to the classical approach. For many problems escape probabilities are of

* Received by the editors March 6, 1984, and in revised form August 1, 1984.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by the National Science Foundation under grant Eng-ECS 82-11476.

‡ Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The research of this author was supported in part by the Air Force Office of Scientific Research under grant AF-AFOSR 81-0116, by the National Science Foundation under grant ENG-ECS 82-11476 and the Office of Naval Research under grant #N00014-76-C-0279-P6.

greater interest than the "localized" asymptotic normalized variances of the errors. We now formulate the precise problem. It will be necessary to deal with (1.1) and (1.2) both in their original discrete parameter form, and in an interpolated continuous parameter form. To facilitate this, define $t_0 = 0$, $t_n = \sum_0^{n-1} a_i$, and $m(t) = \max\{n: t_n \leqq t\}$. Thus $m(t_n) = n$. Define $m_n(t) = m(t_n + t)$. Then $0 \leqq t - \sum_n^{m_n(t)-1} a_i \leqq a_{m_n(t)}$. For each n and x define the process $\{X_j^n, j \geqq n\}$ by $X_n^n = x$ and

$$(1.4_{\mathrm{RM}}) \qquad X_{j+1}^n = X_j^n + a_j(\bar{b}(X_j^n) + b(X_j^n, \xi_j)),$$

$$(1.4_{\mathrm{KW}}) \qquad X_{j+1}^n = X_j^n + a_j\bar{b}(X_j^n) + a_j b(X_j^n, \xi_j)/c_n, \qquad j \geqq n.$$

Thus, we have fixed the initial time at n, and initial condition at x. Define the piecewise constant function $x^n(\cdot)$ by

$$x^n(t) = X_j^n \quad \text{on } [t_j - t_n, t_{j+1} - t_n),$$

and let $\bar{x}^n(\cdot)$ denote the piecewise linear interpolation. Henceforth, for notational simplicity, we *drop* the superscript n on X_j^n.

Fix $T < \infty$ and define $\tau_G^n = \min\{t: \bar{x}^n(t) \notin G\}$ and let $A \subset C_x[0, T]$, the set of R^r-valued continuous functions on $[0, T]$ with initial value x. We seek normalizing sequences $\{\lambda_n\}$ such that the limits below exist and can be evaluated:

$$(1.5) \qquad \lim_n \lambda_n \log P_x\{\tau_G^n \leqq T\}, \qquad \lim_n \lambda_n \log P_x\{\bar{x}^n(\cdot) \in A\}.$$

The limits (1.5) are often continuous in x. In particular, under quite broad conditions, the first limit is continuous in x in G.

The limits in (1.5) are of considerable value in studying the asymptotic properties of $\{X_n\}$. The analysis yields information on the locations or distributions of the X_n for large n, and on the most likely escape routes from G. It exploits more of the structure of the algorithm, and results can be obtained even when $\bar{b}(\cdot)$ is "flat" near θ. Also (unlike the classical theory) $\bar{b}(\cdot)$ need not be differentiable—Lipschitz continuity is enough. This is useful in dealing with cases where the algorithm arises in a minimax problem—or where the derivatives of $\bar{b}(\cdot)$ or $b(\cdot, \cdot)$ are discontinuous at θ. The limits are obtained via the variational approach of the theory of large deviations [3], [4], [5], ([5] is our main reference). The values of the limits provide useful information on the dependence of the performance of the algorithm on the data $\bar{b}(\cdot)$, the sequences $\{a_n, c_n\}$ and the statistics of $\{\xi_n\}$, and some of this information is obtainable even without solving the variational problem.

The type of theory used and developed here should also be of considerable value in the analysis of the asymptotic behavior of other types of recursive and tracking algorithms (e.g., adaptive control and communications systems where $a_n \equiv a$, a constant gain). Commonly, for such systems, estimates of path excursions and passage times and moments (from a "stability" set, for example) are of at least as much interest as the usual "localized" mean square error estimates. Applications to other problems in probability and statistics appear in [6], [7]. In [8], large deviations theory is used to approximate the probability of breakdown of a form of an ALOHA-type communications system.

The basic assumptions are introduced in § 2. In (A2) and (A2′) the H-functional (the normalized log of an exponential moment) is introduced in the form in which it is needed, and ways of obtaining the functional and its structure from the structure of the $\{\xi_j\}$ noise are discussed. Various ways of approximating it (required in the subsequent development) are also discussed. In § 3 the Legendre (or Cramer) transformation of H is introduced, together with the *action functional* $S(T, \phi)$, and the main

limit theorems proven. The basic result is of the following "large deviations" type. Let $A \subset C_x[0, T]$, with $\bar{A}$ = closure of A and A^0 = interior of A. Then, under our conditions (we usually write x^n for $x^n(\cdot)$)

$$\begin{aligned} -\inf_{\phi \in A^0} S(T, \phi) &\leqq \underline{\lim}_n \lambda_n \log P_x\{x^n \in A\} \leqq \overline{\lim}_n \lambda_n \log P_x\{x^n \in A\} \\ &\leqq -\inf_{\phi \in \bar{A}} S(T, \phi). \end{aligned} \tag{1.6}$$

Because of its importance in applications, we are particularly concerned with the case where A is the closed set

$$\begin{aligned} &A = \{\phi(\cdot)\colon \phi(0) = x \text{ and } \phi(t) \notin G \text{ for some } t \leqq T\}, \\ &P_x\{\bar{x}^n \in A\} = P_x\{\tau_G^n \leqq T\}, \end{aligned} \tag{1.7}$$

and with the dependence of $\{\lambda_n\}$ and $S(T, \phi)$ on the problem data. Appendix 1 proves an auxiliary "large deviations" estimate. Appendix 2 obtains some "exponential-type" estimates of the probability of various sets, for use in § 3, when the noise is unbounded.

If A is defined by (1.7), then even if there is not equality in (1.6), there will be for a *slight perturbation* of G. For $\rho > 0$, let G_ρ denote a ρ-neighborhood of G. For $\rho < 0$, define $G_\rho = \{x \in G\colon d(x, \partial G) > -\rho\}$. Let A_ρ denote the set (1.7) corresponding to G_ρ. Then $S_\rho = \inf_{\phi \in A_\rho} S(T, \phi)$ decreases as ρ decreases, and S_ρ is continuous at all but a countable number of ρ. Let S_ρ be continuous at $\rho = 0$ ($G_0 = G$). Then

$$\inf_{\phi \in A^0} S(T, \phi) = \inf_{\phi \in \bar{A}} S(T, \phi) \tag{*}$$

and there is equality in (1.6). This continuity, as well as the continuity of $\inf_{\phi \in A} S(T, \phi)$ in the initial condition x can be proved via the methods in [10]. In particular, if the inf in (1.6) are finite, then (*) holds under the degeneracy or nondegeneracy assumptions (A4.2), (A4.4) (for bounded noise) and the controllability-type assumptions (A4.7) in [10]. In an earlier paper [9], the special case $b(x, \xi) = \xi$, with $\{\xi_n\}$ i.i.d. and Gaussian was treated. The λ_n sequence used there is proportional to the one used here, but is not the same. Korostelev [12], [13] has also considered this type of noise sequence with $b(x, \xi) = b_0(x)\xi$.

We now cite some specific results. Let $a_n = 1/n^\rho$, $\rho \leqq 1$, $c_n = 1/n^\gamma$ where $\rho - 2\gamma > 0$ (if $a_n = A/n^\rho$, etc., absorb the A into the other data). Then (asymptotic logarithmic equivalence)

$$P_x\{\tau_G^n \leqq T\} \sim \exp - n^\rho V_\rho, \tag{1.8$_{\text{RM}}$}$$

$$P_x\{\tau_G^n \leqq T\} \sim \exp - n^{\rho - 2\gamma} V'_\rho, \tag{1.8$_{\text{KW}}$}$$

where for $\rho < 1$, $V_\rho = V'_\rho$ and does not depend on ρ. The V_ρ are obtained from the solution to a variational or optimal control problem, and the solution also yields the "most likely" exit paths from G—information which is useful in applications.

If the $\{\xi_n\}$ are mean zero stationary, Gaussian and $b(x, \xi) = b_0(x)\xi$, then the action functional is (if the inverse exists)

$$\begin{aligned} S(T, \phi) &= \int_0^T \tfrac{1}{2}(\dot{\phi} - \bar{b}(\phi))'[b_0(\phi)\bar{R}b_0'(\phi)]^{-1}(\dot{\phi} - \bar{b}(\phi))h(s)\, ds \\ &= \int_0^T L(\dot{\phi}, \phi, s)\, ds, \end{aligned} \tag{1.9}$$

where $\bar{R} = \sum_{-\infty}^{\infty} E\xi_j\xi_0'$ and $0 < h(s)$ is a continuous function. $h(s) = 1$ if $\rho < 1$ and equals

(RM case) e^{-s} if $\rho = 1$. Also $V_\rho = \inf\{S(T, \phi)\colon \phi(0) = x, \phi(t) \notin G \text{ for some } t \leqq T\}$. If the inverse does not exist, then replace the integrand in (1.9) by

$$\min_\alpha [\alpha'(\dot\phi - \bar b(\phi)) - \tfrac{1}{2}\alpha' b_0(\phi)\bar R b_0'(\phi)\alpha]h(s) = L(\dot\phi, \phi, s).$$

Alternatively we can minimize $\frac{1}{2}\int_0^T u'u ds$ subject to $\dot\phi = \bar b(\phi) + b_0(\phi)\bar R^{1/2}u$ and $\phi(0) = x$, $\phi(t) \notin G$ for some $t \leqq T$. Whether $\{\xi_n\}$ is Gaussian or not, the action functional for the KW case is of the form (1.9).

Note that (1.8) is a much faster rate of decrease of the escape probability than is implied by the "asymptotic normality" way of dealing with the asymptotic properties.

In the general case covered by Theorem 2.4 (and for the corresponding KW case), the equivalent optimal control problem can be shown to be the following

$$S(T, \phi) = \int_0^T h_0(s) L_2(u(s))\, ds$$

subject to

$$\dot\phi = \bar b(\phi) + b_0(\phi) g u,$$

$\phi(0) = x$, $\phi(t) \notin G$ for some $t \leqq T$.

Note. The book [14] came to the authors' attention after this paper was completed. It uses large deviations methods to obtain weak conditions for w.p.l. convergence of the algorithm $X_{n+1} = X_n + a_n b(X_n) + d_n \xi_n$, where d_n and ξ_n do not depend on X_n. For the more complex systems dealt with in this paper, the methods which we use can also yield w.p.l. convergence under much weaker conditions than are used, but our main interest is in obtaining explicit formulas for a rate of convergence or an escape probability, so as too better understand the algorithm.

2. The *H*-functional and its properties. The first assumption is on $\{a_n\}$, and essentially says that a_n does not decrease "too fast." Examples appear below.

A1. *$a_n \downarrow 0$ as $n \to \infty$, and $\sum a_n = \infty$. There is a continuous function $h_1(\cdot)$ (positive for $s < \infty$) and such that*

$$a_{m_n(s+\delta)}/a_n \to h_1(s) \quad \text{as } n \to \infty \text{ and then } \delta \downarrow 0. \tag{2.1}$$

By (2.1), $[m_n(s+\delta) - m_n(s)] \sim \delta / a_{m_n(s)}$. Thus

$$a_n[m_n(s+\delta) - m_n(s)]/\delta \to h_0(s) \equiv h_1^{-1}(s), \tag{2.2}$$

as $n \to \infty$ and then $\delta \to 0$. Let $\alpha(\cdot)$ and $\psi(\cdot)$ be piecewise constant functions; w.l.o.g., we let Δ denote the length of their intervals of constancy (which will depend on the function). The next assumption concerns the existence of an exponential moment of the type used in the theory of large deviations. We state the assumption in the form (A2), because that is the form in which it is generally used. But, under (A3), (A2) is equivalent to (A2′) (stated after Theorem 1), and (A2′) is the usual form used in the theory of large deviations for discrete systems.

See the motivation for (2.3) given after (A3). *For the* RM *case, the natural normalization is $\lambda_n = a_n$ and for the* KW *case $\lambda_n = a_n/c_n^2$.*

The form (2.3) below is somewhat abstract, although it is the correct form for our problem. In order to get a feeling for (2.3), several equivalent forms are developed in the sequel, together with results which show how to calculate it in special cases.

$A2_{RM}$ (RM *case*). *There is a continuous function $H_1(\cdot, \cdot, \cdot)$ with $H_1(\cdot, x, s)$ continuously differentiable for each x and s, and such that for each $\alpha(\cdot)$, $\psi(\cdot)$, the limit in*

(2.3) *exists.* (Here $\delta < \Delta$ and Δ/δ is taken to be an integer.)

$$\lim_{\delta\to 0}\lim_{n\to\infty}\lambda_n \log E \exp\frac{1}{\lambda_n}\left\{\sum_{i=0}^{T/\delta-1}\alpha'(i\delta)a_{m_n(i\delta)}\sum_{j=m_n(i\delta)}^{m_n(i\delta+\delta)-1}b(\psi(i\delta),\xi_j)\right\} = \int_0^T H_1(\alpha(s),\psi(s),s)\,ds. \tag{2.3}$$

$A2_{KW}$. *Replace* $a_{m_n(i\delta)}$ *by* $a_{m_n(i\delta)}/c_{m_n(i\delta)}$ *in* (2.3).

$A3_{RM}$. *Let* $b(x,\xi_n) = b_1(x,\tilde{\xi}_n) + b_0(x)\hat{\xi}_n$, *where* $\{\tilde{\xi}_n\}$ *and* $\{\hat{\xi}_n\}$ *are mutually independent,* $\{\hat{\xi}_n\}$ *is stationary mean zero Gaussian with a summable correlation function and* $\{\tilde{\xi}_n\}$ *is stationary and bounded.* $b_1(\cdot,\xi)$, $b_0(\cdot)$ *and* $\bar{b}(\cdot)$ *are uniformly (in* ξ, x*) Lipschitz and are bounded.*

The limit in (2.3) exists in "typical" situations. The form is also consistent with that used in [5] for the system

$$\dot{x}^\gamma = b(x^\gamma, \xi(t/\gamma)),$$

or the discrete parameter counterpart

$$X^\gamma_{n+1} = X^\gamma_n + \gamma b(X^\gamma_n, \xi_n),$$

where the H-functional is defined by

$$\int_0^T H_1(\alpha(s),\psi(s))\,ds = \lim_\gamma \gamma \log E \exp\int_0^{T/\gamma}\alpha'(\gamma s)b(\psi(\gamma s),\xi(s))\,ds$$

or by

$$\lim_\gamma \gamma \log E \exp\sum_0^{T/\gamma-1}\alpha'(i\gamma)b(\psi(i\gamma),\xi_j).$$

We use $a_{m_n(i\delta)}b(\psi(i\delta),\xi_j)$ in lieu of $a_j b(\psi(i\delta),\xi_j)$ because it is easier to calculate the limits when the coefficients are "piecewise constant." But the limits are the same under our conditions; see the remark after (A2′) below. The functions h_1 and h_0 and the time dependence of $H_1(\cdot,\cdot,\cdot)$ appear due to the "time-varying" scaling (due to the fact that $a_n \to 0$).

Remarks. $H_1(\cdot,\psi,s)$ is convex for each ψ and s. If (A2) holds and $a_j b$ in (2.3) is replaced by $a_j\bar{b} + a_j b$ (or $a_j\bar{b} + a_j b/c_j$ in the KW case) then $H_1(\alpha,\psi,s)$ is replaced by $H(\alpha,\psi,s)$ where

$$H(\alpha,\psi,s) = H_1(\alpha,\psi,s) + \alpha'\bar{b}(\psi). \tag{2.4}$$

In order to simplify the notation, we do the proofs for the "bounded" and Gaussian noise cases separately, and use the notation $b(x,\xi)$ *or* $b(x)\xi$ *as appropriate. The results hold for the "combined" case.*

In the Gaussian case, H_1 can be explicitly evaluated. Using the fact that for Gaussian mean zero ξ, $\log E \exp\xi = E\xi^2/2$, we obtain

$$H_1(\alpha,\psi,s) = \tfrac{1}{2}h_1(s)\alpha' b_0(\psi)\bar{R}b_0'(\psi)\alpha, \tag{2.5}$$

where $\bar{R} = \sum_{j=-\infty}^{\infty} R(j)$ and $R(j) = E\xi_0\xi_j'$.

In the scalar case with $\alpha(s)=\alpha$, $\psi(s)=\psi$, both constants, (2.5) follows from

$$\begin{aligned}\lim_{\gamma} \gamma \log E \exp \sum_{0}^{T/\gamma} \alpha b_0(\psi)\xi_j &= \lim_{\gamma} \gamma \frac{\alpha^2 b_0^2(\psi)}{2} = \sum_{i,j=0}^{T/\gamma} E\xi_i\xi_j \frac{\alpha^2 b_0^2(\psi)}{2} \lim \gamma \sum_{i,j=0}^{T/\gamma} R(i-j) \\ &= T\frac{\alpha^2 b_0^2(\psi)}{2} \sum_{-\infty}^{\infty} R(j),\end{aligned}$$

and the general case is dealt with in the same way.

Examples of (A1). Let $a_n = 1/n = \lambda_n$. Then $m_n(s) \sim ne^s$ ($g_n \sim f_n$ means that $g_n/f_n \to 1$ as $n \to \infty$) and $h_0(s) = e^s$, $h_1(s) = e^{-s}$. If $a_n = 1/n^\rho$, where $\rho \in (0, 1)$, then $m_n(s) \sim n + sn^\rho$ and $h_0(s) = 1$. Let V_ρ denote the infima (on either the r.h.s. or l.h.s. of (1.6)) when $\rho < 1$ and $\rho = 1$, respectively. Then for $\rho < 1$, V_ρ does not depend on ρ. Also if there is equality in (1.6) for the set A of (1.7), then (1.8) holds.

We next present an approximation theorem, and then some examples. In (2.3), we used piecewise constant (on intervals of length δ) coefficients of $b(\psi, \xi_j)$, since it is usually easier to prove that the appropriate limit exists (as compared with the case where $a_j b(\psi, \xi_j)$ is used). For the subsequent work, it is important that the limits hold when the coefficients are varied slightly. In particular, we will need that H_1 can also be defined by the limit in (2.6).

$$\begin{aligned}&\int_0^T H_1(\alpha(s), \psi(s), s)\, ds \\ &\quad = \lim_n \lambda_n \log E \exp\left\{ \sum_{i=0}^{T/\delta-1} \alpha'(i\delta) \sum_{j=m_n(i\delta)}^{m_n(i\delta+\delta)-1} a_j b(\psi(i\delta), \xi_j)/\lambda_n \right\}.\end{aligned} \tag{2.6}$$

THEOREM 2.1. *Assume* (A1), ($A2_{RM}$) *and* ($A3_{RM}$). *Then* (2.6) *holds. Also the limit in* (2.6) *is the same if the* a_j *are replaced by* $\hat{a}_j^\delta$, *where*

$$\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{j \in [n, m_n(T))} |\hat{a}_j^\delta - a_j|/a_j = 0, \tag{2.7}$$

and the limit in (2.6) *is then taken in the order* $\lim_{\delta \to 0} \lim_{n \to \infty}$.

Remark. It is readily seen that if $\hat{a}_j^\delta = a_{m_n(i\delta)}$ in the interval $j \in [m_n(i\delta), m_n(i\delta + \delta))$, then (2.7) holds.

Proof. For notational simplicity, let $\alpha(\cdot) \equiv \alpha$ and $b(x, \xi) = \xi$, a scalar, and let $\hat{a}_j^\delta = \hat{a}_j$. The general proof is almost the same. We will show that

$$\lim_n |\lambda_n \log E \exp \hat{A}_n - \lambda_n \log E \exp \tilde{A}_n| = 0, \tag{2.8}$$

where

$$\hat{A}_n = \frac{\alpha}{\lambda_n} \sum_{n}^{m_n(T)-1} \hat{a}_j \xi_j, \qquad \tilde{A}_n = \frac{\alpha}{\lambda_n} \sum_{n}^{m_n(T)-1} \tilde{a}_j \xi_j$$

$$\sup_{j \in [n, m_n(T))} |\tilde{a}_j - \hat{a}_j|/a_j \overset{n}{\to} 0, \qquad a_j = O(\lambda_n).$$

For the Gaussian case, a direct evaluation (along the lines of the calculation which we would use to get (2.5)) yields the result. For the bounded noise case, let $|\tilde{a}_j - \hat{a}_j|/\lambda_n \leqq$

ε. Then there is a constant k such that

$$\exp -\varepsilon[m_n(T)-n]k \leqq \exp \alpha \sum_{n}^{m_n(T)-1} |(\tilde{a}_j - \hat{a}_j)\xi_j|/\lambda_n \leqq \exp \varepsilon[m_n(T)-n]k. \tag{2.9}$$

Also,

$$(2.8) = \lambda_n[\log E \exp (\hat{A}_n + (\tilde{A}_n - \hat{A}_n)) - \log E \exp \hat{A}_n].$$

The use of (2.9) and (2.2) in (2.8) yields that (2.8) $\to 0$ as $n \to \infty$. Q.E.D.

Under (A1) and ($A3_{RM}$), Theorem 2.1 implies that ($A2_{RM}$) is equivalent to ($A2'_{RM}$) below. In ($A2'_{RM}$) we divide by m/δ, the number of j indices in the sum. In ($A2_{RM}$), the normalization is a multiplication by a_n which is $\int_0^T h_0(s)\,ds$ ($\geqq 1$) times the inverse of the number of j indices in that sum.

($A2'_{RM}$). *There is a continuous function $H_0(\cdot,\cdot)$ such that $H_0(\cdot, x)$ is continuously differentiable and*

$$\int_0^T H_0(\alpha(s), \psi(s))\,ds = \lim_{m\to\infty} \frac{\delta}{m} \log E \exp \left\{ \sum_{i=0}^{T/\delta-1} \alpha'(i\delta) \sum_{j=mi}^{mi+m-1} b(\psi(i\delta), \xi_j) \right\}.$$

Under (A1), ($A2'_{RM}$) and ($A3_{RM}$), it can readily be shown that

$$H_1(\alpha, x, s) = h_0(s) H_0(h_1(s)\alpha, x). \tag{2.10}$$

The general form (2.10) will be quite useful in the sequel.

In order to "roughly" check (2.10), let $c<d$ with d-c small and set $\alpha(s)=\alpha$, $\psi(s)=\psi$ on $[c, d]$ and zero elsewhere. Then under ($A2'_{RM}$)

$$\lambda_n \frac{[m_n(d)-m_n(c)]}{[m_n(d)-m_n(c)]} \log E \exp \alpha' \sum_{m_n(c)}^{m_n(d)-1} a_j b(\psi, \xi_j)/\lambda_n \approx (d\text{-}c) h_0(c) H_0(h_1(c)\alpha, \psi).$$

(The ratios converge as $d \downarrow c$.)

The following corollary is proved by a method very similar to that of Theorem 2.1. Combinations of Theorem 2.1 and the corollary will be used frequently.

COROLLARY 2.2. *Assume* (A1), ($A2_{RM}$) *and* ($A3_{RM}$). *Let $\tilde{a}_j$ equal either zero or a_j and let*

$$\sum_{m_n(i\delta)}^{m_n(\delta+i\delta)-1} |a_j - \tilde{a}_j| \leqq \delta' \quad \text{for } i \leqq T/\delta.$$

Then

$$\lim_{\delta} \lim_{\delta'} \lim_{n} \lambda_n \log E \exp \left\{ \sum_{i=0}^{T/\delta-1} \alpha'(i\delta) \sum_{j=m_n(i\delta)}^{m_n(i\delta+\delta)-1} \tilde{a}_j b(\psi(i\delta), \xi_j)/\lambda_n \right\} = \int_0^T H_1(\alpha(s), \psi(s), s)\,ds. \tag{2.11}$$

Bounded stationary mixing $\{\xi_n\}$. Let E_0 denote conditioning on $\{\xi_j, j<0\}$. For another way of obtaining H_1 or H_0 consider:

$A4_{RM}$. *Let $\{\xi_n, -\infty < n < \infty\}$ be bounded and stationary, and suppose that there is*

a continuous $\hat{H}_0(\cdot,\cdot)$ *with* $\hat{H}_0(\cdot, x)$ *continuously differentiable for each* x *and such that*

$$\lim_N \frac{1}{N} \log E \exp \alpha' \sum_0^{N-1} b(\psi, \xi_j) = \lim_N \frac{1}{N} \log E_0 \exp \alpha' \sum_0^{N-1} b(\psi, \xi_j) = \hat{H}_0(\alpha, \psi), \tag{2.12}$$

where the convergence is uniform in the conditioning data.

Example. Let $\{\xi_n\}$ be a finite state Markov chain with all states communicating with each other. Then $(A4_{RM})$ holds by an argument similar to that of Freidlin [5, Thm. 2.2]. It is also a special case [10, Thm. 3.8].

THEOREM 2.3. *Under* (A1) *and* $(A4_{RM})$, *condition* $(A2_{RM})$ *holds and*

$$H_0(\alpha, \psi) = \hat{H}_0(\alpha, \psi),$$

$$H_1(\alpha, \psi, s) = h_0(s) H_0(h_1(s)\alpha, \psi).$$

Proof. Let $M = T/\delta =$ integer. By Theorem 2.1, (2.3) is equivalent to

$$\begin{aligned} &\int_0^T H_1(\alpha(s), \psi(s), s)\, ds \\ &\quad = \lim_{\delta \to 0} \lim_{n \to \infty} \lambda_n \log E \exp \left\{ \sum_{i=0}^{M-1} \alpha'(i\delta) h_1(i\delta) \sum_{j=m_n(i\delta)}^{m_n(i\delta+\delta)-1} b(\psi(i\delta), \xi_j) \right\}. \end{aligned} \tag{2.13}$$

Thus, we are concerned with $\lim_{\delta,n} \lambda_n \log E_{\delta,n}$, where

$$E_{\delta,n} = E \exp \left\{ \sum_{i=0}^{M-1} \alpha_i' h_i \sum_{j=m_n(i\delta)}^{m_n(i\delta+\delta)-1} b(\psi_i, \xi_j) \right\}$$

where $\alpha_i = \alpha(i\delta)$, $h_i = h_1(i\delta)$ and $\psi_i = \psi(i\delta)$.

Let E_i denote conditioning in $\{\xi_j, j \leqq m_n(i\delta) - 1\}$. Then

$$E_{\delta,n} = E \exp \left\{ \alpha_0' h_0 \sum_{j=m_n(0)}^{m_n(\delta)-1} b(\psi_0, \xi_j) \right\} \cdots E_{M-1} \exp \left\{ \alpha_{M-1}' h_{M-1} \sum_{m_n(T-\delta)}^{m_n(T)-1} b(\psi_{M-1}, \xi_j) \right\}. \tag{2.14}$$

By the hypotheses,

$$\begin{aligned} &\left| E_i \exp \alpha_i' h_i \sum_{j=m_n(i\delta)}^{m_n(i\delta+\delta)-1} b(\psi_i, \xi_j) - \exp \hat{H}_0(\alpha_i h_i, \psi_i)(m_n(i\delta+\delta) - m_n(i\delta)) \right| \\ &\quad \leqq \exp \rho_{ni}[m_n(i\delta+\delta) - m_n(i\delta)] \end{aligned}$$

where $\rho_{ni} \to 0$ as $n \to \infty$, uniformly in i, and in the conditioning data. Substituting this into (2.14), multiplying by λ_n log and taking $n \to \infty$ yields

$$\left| \lambda_n \log E_{\delta,n} - \sum_{i=0}^{M-1} \lambda_n (m_n(i\delta+\delta) - m_n(i\delta)) \hat{H}_0(\alpha_i h_i, \psi_i) \right| \overset{n}{\to} 0$$

from which the theorem readily follows. Q.E.D.

We now treat the special case:

(A5) $\xi_j = \sum_k g_{j-k}\theta_k$, *where the* $\{\theta_k\}$ *are i.i.d. and either Gaussian or bounded, and* $\sum k|g_k| < \infty$ *and* $g_k = 0$ *for* $k < 0$, *and* $b(\psi, \xi) = b_0(\psi)\xi$. $\bar{b}(\cdot)$ *and* $b_0(\cdot)$ *are bounded and Lipschitz continuous.*

Define

$$H_2(\alpha) = \log E \exp \alpha'\theta, \qquad \bar{R} = \sum_{-\infty}^{\infty} E\xi_j\xi_0'. \tag{2.15}$$

THEOREM 2.4. (RM case). *Under* (A1) *and* (A5)

$$H_1(\alpha, \psi, s) = h_0(s)H_2(h_1(s)g'b_0'(\psi)\alpha),$$

where $g = \sum_{i=0}^{\infty} g_i$.

Proof. By the corollary to Theorem 2.1, we can calculate H_1 by using $\tilde{a}_j$ instead of a_j, where $\tilde{a}_j = 0$ for $j \in [m_n(i\delta), m_n(i\delta + \delta'))$, $\delta' < \delta$, for each i, and using $\tilde{a}_j = a_j$ otherwise, and taking limits in the order $\lim_\delta \lim_{\delta'} \lim_n$. We now do an additional approximation. Let $N_0 < 0$ and define $\tilde{\xi}_j = \sum_{k=-\infty}^{j} \tilde{g}_{j-k}\theta_k$, where $\tilde{g}_j = g_j$ for $j \leqq N_0$ and $\tilde{g}_j = 0$ for $j \geqq N_0$. We show that we obtain the correct H_1 by taking limits in the order

$$\lim_{\delta} \lim_{\delta'} \lim_{N_0} \lim_{n}. \tag{2.16}$$

Write $\hat{a}_j = \tilde{a}_j\alpha_j'b_0(\psi_j)$, where $\alpha_j = \alpha(i\delta)$ and $\psi_j = \psi(i\delta)$ for $j \in [m_n(i\delta), m_n(i\delta + \delta))$. Define

$$\hat{A}_k^n = \sum_{j=n}^{m_n(T)-1} \hat{a}_j g_{j-k}/\lambda_n, \qquad \tilde{A}_k^n = \sum_{j=n}^{m_n(T)-1} \hat{a}_j \tilde{g}_{j-k}/\lambda_n.$$

Since

$$\sum_{j=n}^{m_n(T)-1} \hat{a}_j\xi_j/\lambda_n = \sum_k \hat{A}_k^n\theta_k,$$

in order to show (2.16), we need to show that

$$\lim_{\delta} \lim_{\delta'} \lim_{N_0} \lim_{n} \lambda_n \left[\log E \exp \sum_k \hat{A}_k^n\theta_k - \log E \exp \sum_k \tilde{A}_k^n\theta_k \right] = 0. \tag{2.17}$$

Using the independence of the θ_k, and the fact that $E \exp \sum A_k\theta_k = \exp \frac{1}{2}\sum_k E|A_k\theta_k|^2$, (2.17) follows by a direct evaluation in the Gaussian case.

Following the proof of Theorem 2.1, to obtain (2.17) for the bounded noise case, we need only show that for any constant K,

$$\lim_{N_0} \lim_{n} \lambda_n \log \exp K \sum_k |\tilde{A}_k^n - \hat{A}_k^n| = 0.$$

Equivalently, it is enough to show that

$$\lambda_n \sum_{j=n}^{m_n(T)-1} \hat{a}_j \sum_k |g_{j-k} - \tilde{g}_{j-k}|/\lambda_n \to 0, \qquad \text{as } n \to \infty \text{ and then } N_0 \to \infty.$$

But this follows from the fact that $\sum_k |g_k - \tilde{g}_k| \to 0$ as $N_0 \to \infty$. Thus (2.17) holds, and in the evaluation of H_1, we can use $\tilde{a}_j$ in lieu of a_j and assume that $g_k = 0$ for $k \geqq N_0$ for arbitrary $N_0 < \infty$, and take limits in the order (2.16).

Thus, the functional H_1 is given by (2.18).

$$\lim_{\delta} \lim_{\delta'} \lim_{N_0} \lim_{n} \lambda_n \log E \exp \left\{ \sum_{k=n-N_0}^{m_n(T)-1} \left(\sum_{j=n}^{m_n(T)-1} \tilde{a}_j\alpha_j'b_0(\psi_j)\tilde{g}_{j-k} \right) \theta_k/\lambda_n \right\}. \tag{2.18}$$

In fact, owing to the definition of $\tilde{a}_j$, $\tilde{g}_j$, we can replace $n - N_0$ by $m_n(\delta') - N_0$. Due to

the definitions of $\tilde{a}_j$ and $\tilde{g}_j$, for fixed k and large enough n, the coefficients $b_0(\psi_j)$ and α_j of θ_k will depend on k only via the index i of the set $[m_n(i\delta), m_n(i\delta+\delta))$ in which k lies. Thus using the independence of the $\{\theta_k\}$ and the definition of H_2 given in (2.15), we get that

$$(2.18) = \lim_{\delta} \lim_{\delta'} \lim_{N_0} \lim_{n} \lambda_n \sum_{i=0}^{T/\delta-1} \sum_{k=m_n(i\delta)}^{m_n(i\delta+\delta)-1} H_2\left(\sum_j \tilde{a}_j \tilde{g}'_{j-k} b'_0(\psi(i\delta))\alpha(i\delta)/\lambda_n\right).$$

The theorem follows from this and the continuity of H_2. Q.E.D.

The KW *case.* Define $r_j = a_j/c_j$.

$A6_{KW}$. *There is a continuous and positive function* $h_2(\cdot)$ *such that* $(\lambda_n = a_n/c_n^2)$

$$\lambda_{m_n(s+\delta)}/\lambda_n \to h_2(s)$$

as $n \to \infty$ *and then* $\delta \downarrow 0$.

THEOREM 2.5 (KW case). *Assume* (A1), (A6) *and the Gaussian case of* (A5). *Then* $(A2_{KW})$ *holds with*

$$H_1(\alpha, \psi, s) = \tfrac{1}{2} h_2(s) \alpha' b_0(\psi) \bar{R} b'_0(\psi) \alpha. \tag{2.19}$$

Also, the limit in $(A2_{KW})$ *is the same if the coefficients used there are replaced by* $\hat{a}_j^\delta$, $\hat{c}_j^\delta$ *such that* $\hat{a}_n^\delta/a_n \to 1$ *and* $\hat{c}_n^\delta/c_n \to 1$ *as* $n \to \infty$ *and then* $\delta \to 0$.

Proof. For ease of calculation, we set $\alpha = \text{constant}$, $\psi = \text{constant}$, and use the coefficients $r_n = a_n/c_n$. We have

$$\begin{aligned}
&\lambda_n \log E \exp \alpha' \sum_{j=n}^{m_n(T)-1} r_j b_0(\psi) \xi_j/\lambda_n \\
&\quad = \lambda_n \log \exp E\left[\alpha' \sum_{j=n}^{m_n(T)-1} r_j b_0(\psi)\xi_j/\lambda_n\right]^2 \\
&\quad = \frac{1}{2} \sum_{i,j=n}^{m_n(T)-1} r_i r_j \alpha' b_0(\psi) R(i-j) b'_0(\psi)\alpha/\lambda_n \\
&\quad \to \frac{1}{2}\int_0^T h_2(s)\, ds\, \alpha' b_0(\psi)\bar{R} b'_0(\psi)\alpha.
\end{aligned} \tag{2.20}$$

The last assertion of the theorem can readily be seen from the form of the sum above. Q.E.D.

THEOREM 2.6 (KW case). *Let* $\{\xi_n\}$ *be i.i.d. and bounded and assume* (A1) *and* $(A6_{KW})$. *Then the last sentence of Theorem* 2.5 *holds. Define* (*the* H_0 *below is the same as the* H_0 *of* $(A2'_{RM})$)

$$H_0(\alpha, \psi) = \log E \exp \alpha' b(\psi, \xi).$$

Then

$$\begin{aligned}
H_1(\alpha, \psi, s) &= h_2(s)\alpha' H_{0,\alpha\alpha}(0, \psi)\alpha/2 \\
&= h_2(s)\alpha' E b(\psi, \xi) b'(\psi, \xi)\alpha/2.
\end{aligned} \tag{2.21}$$

Proof. $H_0(\cdot, \psi)$ is infinitely differentiable, since the ξ_n are bounded. We have $H_0(0, \psi) = 0 = H_{0,\alpha}(0, \psi)$, since $Eb(\psi, \xi_n) = 0$, and in general

$$H_{0,\alpha}(\alpha, \psi) = \frac{Eb(\psi, \xi) \exp \alpha' b(\psi, \xi)}{E \exp \alpha' b(\psi, \xi)}. \tag{2.22a}$$

Also, using the definition $E_b(\psi, \alpha) = Eb(\psi, \xi) \exp \alpha' b(\psi, \xi)$,

(2.22b)
$$H_{0,\alpha\alpha}(\alpha, \psi) = \frac{E[b(\psi, \xi) \exp \alpha' b(\psi, \xi) - E_b(\psi, \alpha)][b(\psi, \xi) \exp \alpha' b(\psi, \xi) - E_b(\psi, \alpha)]'}{(E \exp \alpha' b(\psi, \xi))^2}.$$

We need to evaluate the limit of

$$\lambda_n \sum_{i=0}^{T/\delta - 1} \sum_{j=m_n(i\delta)}^{m_n(i\delta+\delta)-1} H_0(r_j \alpha(i\delta)/\lambda_n, \psi(i\delta)).$$

Owing to the smoothness of $H_0(\cdot, \psi)$, we can write

$$\lambda_n H_0(r_j \alpha(i\delta)/\lambda_n, \psi(i\delta)) = \frac{r_j^2}{2\lambda_n} \alpha'(i\delta) H_{0,\alpha\alpha}(0, \psi(i\delta))\alpha(i\delta) + \lambda_n O\left(\frac{r_j^3}{\lambda_n^3}\right), \tag{2.23}$$

where $O(\cdot)$ is uniform in α, ψ in each bounded set. Then the expression (2.23) converges to $\int_0^T H_1(\alpha(s), \psi(s), s)\, ds$, as $n \to \infty$ and $\delta \to 0$. The first assertion of the theorem readily follows from the special form of (2.23). Q.E.D.

Remark. Observe that the H-functional in (2.21) has the *Gaussian* form, although the ξ_j are not Gaussian. Let $a_n = n^{-1}$, $c_n = n^{-\gamma}$ with $2\gamma < 1$. Then $\lambda_n = n^{2\gamma - 1}$ and $h_2(s) = \exp(2\gamma - 1)s$. If $a_n = n^{-\rho}$, $c_n = n^{-\gamma}$ with $2\gamma < \rho$, then $h_2(s) = 1$.

The last two results can be combined and extended.

THEOREM 2.7 (KW case). *Let $a_n/c_n^2 \downarrow 0$ and assume* (A1), (A5) *and* ($A6_{KW}$). *Define the* KW *H-function by*

$$\int_0^T H_1(\alpha(s), \psi(s), s)\, ds$$
$$= \lim_\delta \lim_n \log E \exp\left\{\frac{1}{\lambda_n} \sum_{i=0}^{T/\delta-1} \alpha'(i\delta) \sum_{j=m_n(i\delta)}^{m_n(i\delta+\delta)-1} a_{m_n(i\delta)} b(\psi(i\delta), \xi_j)/c_{m_n(i\delta)}\right\}.$$

The limit is the same if $\hat{a}_j^\delta$, $\hat{c}_j^\delta$ replace $a_{m_n(i\delta)}$, $c_{m_n(i\delta)}$ in each interval $[m_n(i\delta), m_n(i\delta+\delta)-1)$ if $\hat{a}_j^\delta/a_j \to 1$, $\hat{c}_j^\delta/c_j \to 1$ as $j \to \infty$ then $\delta \to 0$. Furthermore, we have the Gaussian form

$$H_1(\alpha, \psi, s) = \tfrac{1}{2} h_2(s) \alpha' b_0(\psi) g H_{2,\alpha\alpha}(0) g' b_0'(\psi) \alpha.$$

The proof uses the partial summation idea of Theorem 2.4 and the expansion idea of Theorem 2.6, and is omitted.

3. The limit theorems. Until mentioned otherwise, attention is restricted to the RM case. Define the Legendre transform L and *action functional S* by

$$L(\beta, x, s) = \sup_\alpha [\alpha'\beta - H(\alpha, x, s)], \tag{3.1}$$

$$S(T, \phi) = \begin{cases} \int_0^T L(\dot{\phi}, \phi, s)\, ds & \text{if } \phi \text{ is absolutely continuous,} \\ \infty & \text{otherwise.} \end{cases} \tag{3.2}$$

Using the representation (2.10), and $H(\alpha, x, s) = \alpha' \bar{b}(x) + H_1(\alpha, x, s)$ and $h_0(s)h_1(s) = 1$, we have

$$\begin{aligned} L(\beta, x, s) &= \sup_\alpha [\alpha'(\beta - \bar{b}(x)) - h_0(s) H_0(h_1(s)\alpha, x)] \\ &= h_0(s) L_0(\beta - \bar{b}(x), x), \end{aligned} \tag{3.3}$$

where

$$L_0(\beta, x) = \sup_{\alpha} [\alpha'\beta - H_0(\alpha, s)]. \tag{3.4}$$

The key result is Theorem 3.2, which proves (1.6). Theorem 3.1 gives an estimate which is needed in the proof of Theorem 3.2. Owing to the assumptions on the noise, some modification of the classical results [5] is needed.

All the functions ϕ below, with or without affixes, are in $C_x[0, T]$.

THEOREM 3.1. *Assume* (A1), ($A2_{RM}$) *and* ($A3_{RM}$). *Given positive c, h and s and function* $\phi(\cdot)$, *there is an* $n_0 < \infty$ *such that for* $n \geqq n_0$ ($d(\cdot, \cdot)$ *is the* sup norm *distance*)

$$P\{d(x^n, \phi) \leqq c\} \geqq \exp -[S(T, \phi) + h]/\lambda_n \tag{3.5}$$

$$P\{d(x^n, \Phi_s) > c\} \leqq \exp -[s - h]/\lambda_n \tag{3.6}$$

where $\Phi_s = \{\phi \in C_x[0, T]: S(T, \phi) \leqq s\}$.

Before proving Theorem 3.1, we state the desired result, Theorem 3.2. Inequalities (3.5) and (3.6) imply (3.7). The result is well-known in large deviations theory. See [5], [6], and the remark following the theorem statement.

THEOREM 3.2. *Under* (A1), ($A2_{RM}$) *and* ($A3_{RM}$), *for any* $A \subset C_x[0, T]$,

$$\begin{aligned} -\inf_{\phi \in A^0} S(T, \phi) &\leqq \underline{\lim}_n \lambda_n \log P_x\{x^n(\cdot) \in A\} \\ &\leqq \overline{\lim}_n \lambda_n \log P_x\{x^n(\cdot) \in A\} \leqq -\inf_{\phi \in \bar{A}} S(T, \phi). \end{aligned} \tag{3.7}$$

Remark. For purposes of self containment, a brief outline of the proof that (3.5) and (3.6) imply (3.7) will be given. Let $\phi_0 \in A^0$ and let $N_\rho(\phi_0)$ be a ρ-neighborhood of ϕ_0 in A^0. Fix small $h > 0$. Then for large n, (3.5) implies

$$\lambda_n \log P\{x^n \in A\} \geqq \lambda_n \log P\{d(x^n, \phi_0) < \rho\} \geqq -[S(T, \phi_0) + h/2].$$

Choose ϕ_0 such that the r.h.s. is within $h/2$ of $\inf_{\phi \in A^0} S(T, \phi)$. This yields the left side of (3.7). For any $s > 0$ such that Φ_s is disjoint (distance $> \rho > 0$) from $\bar{A}$, we have for large n

$$\lambda_n \log P\{x^n \in \bar{A}\} \leqq \lambda_n \log P\{d(x^n, \Phi_s) > \rho\},$$

where 2ρ is the distance between $\bar{A}$ and the compact set Φ_s. Let $s = \inf_{\phi \in \bar{A}} S(T, \phi) - h/2$ and choose ρ appropriately to complete the details.

Proof of Theorem 3.1. Fix δ, and let $\psi(\cdot)$ be piecewise constant (Δ = width of intervals of constancy). If ϕ is piecewise constant, with value $\phi(i\Delta)$ on $[i\Delta, i\Delta + \Delta)$, define $\bar{\phi}$ to be its piecewise linear interpolation. Define for ϕ absolutely continuous (let $S^\psi(T, \phi) = \infty$ otherwise).

$$S^\psi(T, \phi) = \int_0^T L(\dot{\phi}(s), \psi(s), s)\, ds, \tag{3.8}$$

$$\Phi_s^\psi = \{\phi \in C_x[0, T]: S^\psi(T, \phi) \leqq s\}. \tag{3.9}$$

For each ψ, x and n, define the processes $\{X_j^{\psi,n}, j \geqq n\}$ and $x^{\psi,n}$ by $X_n^{\psi,n} = x$ and for $j \geqq n$ (writing ψ_j^n for $\psi(t_j - t_n)$)

$$X_{j+1}^{\psi,n} = X_j^{\psi,n} + a_j \bar{b}(\psi_j^n) + a_j b(\psi_j^n, \xi_j), \tag{3.10}$$

and $x^{\psi,n}(t) = X_j^{\psi,n}$ on $[t_j - t_n, t_{j+1} - t_n)$. Thus $S^\psi(T, \phi)$ is the action functional corresponding to $\{x^{\psi,n}, n = 1, 2, \cdots\}$.

It is proved in Appendix 1 that for each positive a, h and s and each $\phi \in C_x[0, T]$, there is an $n_0 < \infty$ such that for $n \geqq n_0$,

$$P\{d(x^{\psi,n}, \phi) < a\} \geqq \exp - \left[S^{\psi}(T, \phi) + \frac{h}{2} \right] \Big/ \lambda_n, \tag{3.11}$$

$$P\{d(x^{\psi,n}, \Phi_s^{\psi}) > a\} \leqq \exp - \left[s - \frac{h}{2} \right] \Big/ \lambda_n. \tag{3.12}$$

Let $\{\psi^k\}$ denote a sequence of step functions which converge uniformly on $[0, T]$ to a continuous function $\phi(\cdot)$. Then Friedlin's [5, Lemma 3.2] yields (with only a minor adjustment to take care of the fact that $L(\beta, x, s)$ is time dependent) that there is a sequence ϕ^k which converges to ϕ uniformly on $[0, T]$ and such that

$$S^{\psi^k}(T, \phi^k) \to S(T, \phi). \tag{3.13}$$

Fix ϕ and $h > 0$. For each $\varepsilon > 0$ choose ϕ^{ε} and a step function ψ^{ε} such that $d(\phi^{\varepsilon}, \phi) < \varepsilon$, $d(\psi^{\varepsilon}, \phi) < \varepsilon$ and

$$S^{\psi^{\varepsilon}}(T, \phi^{\varepsilon}) \leqq S(T, \phi) + h/2. \tag{3.14}$$

We proceed to use this approximation and (3.11), (3.12) to obtain (3.5) and (3.6), following the proof in [5] as closely as possible.

It follows from the estimates in Appendix 2 (Lemma A.4) that for each $c > 0$, $\delta > 0$ and $K < \infty$ there are $n_0 < \infty$, $c_0 > 0$ and $\varepsilon_0 > 0$ and a set B_n such that $n \geqq n_0$, $c' \leqq c_0$ and $\varepsilon \leqq \varepsilon_0$ imply that $P\{B_n\} \leqq \exp - K/\lambda_n$, and (compare with [5, bottom of p. 141])

$$\{\omega\colon d(x^n, \phi) < c\} \supset \{\omega\colon d(x^{\psi^{\varepsilon},n}, \phi^{\varepsilon}) < c'\} - B_n. \tag{3.15}$$

By this inclusion and (3.11) and (3.14), for small ε and large n

$$P\{d(x^n, \phi) < c\} \geqq \exp - \left[S^{\psi^{\varepsilon}}(T, \phi^{\varepsilon}) + \frac{h}{2} \right] \Big/ \lambda_n + (\exp - K/\lambda_n)$$

$$\geqq \exp - [S(T, \phi) + h]/\lambda_n$$

which is (3.5).

If $\{\xi_j\}$ is bounded, then the Lipschitz condition on $\bar{b}(\cdot)$ an $b(\cdot, \xi)$ imply that for each $B < \infty$ the set of piecewise linear interpolations

$$R_n = \{x^n(t), t \leqq T\}$$

are in a compact set R_0, not depending on n. As far as the estimates (3.5), (3.6) or (3.11), (3.12) are concerned, it does not matter whether we used x^n and $x^{n,\psi}$ or their piecewise linear interpolations. Henceforth we use the piecewise linear interpolations, without altering the notation. In the Gaussian case, due to the unboundedness of $\{\xi_n\}$, the R_n are not in a compact set since the linear interpolations of $\sum_{j=n}^{m_n(t)} a_j b_0(X_j^n)\xi_j$ are not in a compact set—since the ξ_j are unbounded. The interpolations of $\sum_{j=n}^{m_n(T)} a_j \bar{b}(X_j^n)$ are in a compact set because $\bar{b}(\cdot)$ is bounded. But the estimates in Lemma A.2 of Appendix 2 imply the following: For each K and $\rho > 0$, there are $N < \infty$ and $\phi_1, \cdots, \phi_N$ such that for each n the union of the ρ- neighborhoods of the ϕ_i covers R_n except for a set of paths B_n whose probability is $\leqq k \exp - K/\lambda_n$, where k does not depend on n.

Fix positive s, c and ρ. By the above remarks there are $N < \infty$ and ϕ_i, $i \leqq N$, such that $d(\phi_i, \Phi_s) > c/2$ (Φ_s is compact in $C_x[0, T]$ by the lower semicontinuity of $S(T, \phi)$) and the union of the ρ-neighborhoods of the ϕ_i covers

$$F_n = R_n - (c/2\text{-neighborhood of } \Phi_s) - B_n$$

for each n. If $\rho < c/2$, then

$$P\{d(x^n, \Phi_s) > c\} \leqq \sum_{i=1}^{N} P\{d(x^n, \phi_i) < \rho\} + k \exp - K/\lambda_n. \tag{3.16}$$

By the results of Appendix 2 (Lemma A.4) for each $K < \infty$, there are $\rho_0 > 0$ and $n_0 < \infty$ such that $\rho \leqq \rho_0$ and $n \geqq n_0$ and $d(\phi, \psi) \leqq \rho$ imply that there is a λ (which we can choose such that it goes to zero as $\rho \to 0$) such that (compare with [5, (3.15)])

$$\{\omega\colon d(x^n, \phi) < \rho\} \subset \{\omega\colon d(x^{\psi,n}, \phi) < \lambda\} \cup C_{n\rho}(\psi, \phi), \tag{3.17}$$

where $P\{C_{n\rho}(\psi, \phi)\} \leqq \exp - K/\lambda_n$. Choose ρ small enough such that $\lambda < c/8$ and $\rho < c/4$.

For small enough $d(\phi_i, \psi_i)$, $d(\phi_i, \Phi_s) > c/2$ implies that $d(\phi_i, \Phi_s^{\psi_i}) > c/4$. This is a consequence of the lower semicontinuity of $S^{\psi}(T, \phi)$ in (ψ, ϕ) and the compactness of Φ_s. If this assertion is not true, then there is a sequence ψ_i^{ε} such that $d(\phi_i, \psi_i^{\varepsilon}) \to 0$ and $d(\phi_i, \phi_i^{\varepsilon}) \to 0$ where $S^{\psi_i^{\varepsilon}}(T, \phi_i^{\varepsilon}) \leqq s$ for all $\varepsilon > 0$. But there is an $s_1 > 0$ such that for any $h_0 > 0$ and small enough $\varepsilon > 0$ (by the cited lower semicontinuity and compactness)

$$S^{\psi_i^{\varepsilon}}(T, \phi_i^{\varepsilon}) \geqq S(T, \phi_i) - h_0 \geqq s + s_1 - h_0,$$

a contradiction.

We now put these estimates together to get (3.6) as follows. Fix $h > 0$. Choose step functions ψ_i such that $d(\psi_i, \phi_i) < \rho$ and is small enough to guarantee $d(\phi_i, \Phi_s^{\psi_i}) > c/4$. Then, by (3.12) and $\lambda < c/8$

$$\begin{aligned} \lambda_n \log P\{d(x^{\psi_i,n}, \phi_i) < \lambda\} &\leqq \lambda_n \log P\, d(x^{\psi_i,n}, \Phi_s^{\psi_i}) > c/8\} \\ &\leqq -(s - h/2) \end{aligned} \tag{3.18}$$

for large n. Finally, (3.16) to (3.18) yield (N can depend on ρ and K)

$$P\{d(x^n, \Phi_s) > c\} \leqq (k + N) \exp - K/\lambda_n + N \exp - (s - h/2)/\lambda_n,$$

which is less than $\exp - (s - h)/\lambda_n$ for large enough n and (3.6) is proved. Q.E.D.

The KW *case.*

THEOREM 3.3. *Let* $a_n/c_n^2 \downarrow 0$ *and*[1] $\sum_n^{m_n(T)} |r_{i+1} - r_i| \to^n 0$ *and assume* (A1), (A5) *and* ($A6_{KW}$). *Let* $\{\xi_n\}$ *be i.i.d. and either bounded or Gaussian. Then* (3.7) *holds for the* KW *case for the general form* $b(x, \xi)$. *Assume*

$$\xi_{n+1} = A\xi_n + C\theta_n, \tag{3.19}$$

where $\{\theta_n\}$ *are i.i.d. mean zero and either bounded or Gaussian, and the roots of* A *are inside the unit circle. Let* $b(x, \xi) = b_0(x)\xi$, *where* $b_0(\cdot)$ *also has a Lipshitz continuous and bounded gradient. Then* (3.7) *holds for the* KW *case.*

The proof is the same as that of Theorem 3.2 with the exception that the "exponential estimates" of Appendix 2 for the KW case are used.

Appendix 1.

Proof of (3.11) *and* (3.12). The proof closely follows that of Freidlin [5, Lemma 3.1]. Let $\delta\alpha_i$, $i \geqq 0$, be vectors in R^r. Define the piecewise constant function $\alpha(\cdot)$ by $\alpha(t) = \sum_{k=i}^{N-1} \delta\alpha_k$ for $t \in [i\Delta, i\Delta + \Delta)$, where $N = T/\Delta =$ integer. Let $\psi(\cdot)$ be constant on each $[i\Delta, i\Delta + \Delta)$ interval. Define $X_{\Delta}^{\psi,n} = (x^{\psi,n}(i\Delta), i = 1, \cdots, N)$. Define

$$h_n^{\psi}(\delta\alpha_0, \cdots, \delta\alpha_{N-1}) = \lambda_n \log E \exp \sum_{i=0}^{N-1} \delta\alpha_i' x^{\psi,n}(i\Delta + \Delta)/\lambda_n. \tag{A1.1}$$

[1] This holds if $r_n \downarrow 0$, the "typical" case.

Recall that

$$x^{\psi,n}(i\Delta+\Delta)=x^{\psi}(t_{m_n(i\Delta+\Delta)}-t_n).$$

Rewrite the sum in (A1.1) in the more convenient form (A1.2) in order to bring (A1.1) closer to (2.3) in appearance.

$$\begin{aligned}&\sum_{i=0}^{N-1}\delta\alpha_i'\left[\sum_{j=0}^{i}(x^{\psi,n}(j\Delta+\Delta)-x^{\psi,n}(j\Delta))+x\right]\\&\quad=\alpha'(0)x+\sum_{i=0}^{N-1}\alpha'(i\Delta)\sum_{j=m_n(i\Delta)}^{m_n(i\Delta+\Delta)-1}a_j[\bar{b}(\psi(i\Delta))+b(\psi(i\Delta),\xi_j)].\end{aligned}\tag{A1.2}$$

We can write

$$\begin{aligned}\lim_{n\to\infty}h_n^{\psi}(\delta\alpha_0,\cdots,\delta\alpha_{N-1})&=\int_0^T H(\alpha(s),\psi(s),s)\,ds+\alpha'(0)x\\&\equiv h^{\psi}(\delta\alpha_0,\cdots,\delta\alpha_{N-1}).\end{aligned}$$

The function h^{ψ} is continuously differentiable and convex, since $H(\cdot,\psi,s)$ is. Let $l^{\psi}(\beta_0,\cdots,\beta_{N-1})$ denote the Legendre transform of h^{ψ}:

$$l^{\psi}(\beta_0,\cdots,\beta_{N-1})=\sup_{\delta\alpha}\left[\sum_i\delta\alpha_i'\beta_i-h^{\psi}(\delta\alpha_0,\cdots,\delta\alpha_{N-1})\right].$$

We also have

$$l^{\psi}(\beta_0,\cdots,\beta_{N-1})=\int_0^T L(\dot{\beta}(s),\psi(s),s)\,ds=S^{\psi}(T,\beta),\tag{A1.3}$$

where we define $\dot{\beta}(t)=[\beta_{i+1}-\beta_i]/\Delta$ in $[i\Delta,i\Delta+\Delta)$ and $\beta(0)=x$. The proof is the same as that of [5, (3.1)], with a slight modification for the s-dependence.

By Gartner's theorem [11, Lemmas 1 and 2], for any $c>0$, $h>0$ vector $B=(\beta_0,\cdots,\beta_{N-1})$, $\beta_i\in R^r$, there is an $n_0<\infty$ such that for $n\geqq n_0$

$$P\{d(X_{\Delta}^{\psi,n},B)<c\}\geqq\exp-\left[l^{\psi}(\beta_0,\cdots,\beta_{N-1})+\frac{h}{2}\right]\Big/\lambda_n,\tag{A1.4a}$$

$$P\{d(X_{\Delta}^{\psi,n},\Phi_{\Delta,s}^{\psi})>c\}\leqq\exp-\left[s-\frac{h}{2}\right]\Big/\lambda_n,\tag{A1.4b}$$

where $\Phi_{\Delta,s}^{\psi}=\{B\colon l^{\psi}(\beta_0,\cdots,\beta_{N-1})\leqq s\}$ and $d(\cdot,\cdot)=$ sup norm distance. For each $\phi\in C_x[0,T]$, define $\bar{\phi}_{\Delta}(\cdot)$ to be the piecewise linear interpolation (interval Δ) of the vector $\phi_{\Delta}=(\phi(0),\cdots,\phi(N\Delta-\Delta))$. By Appendix 2, Lemma A.2, for each $c>0$ and $K<\infty$, there are $\rho_0>0$, $\Delta_0>0$ and $n_0<\infty$ such that $\rho\leqq\rho_0$, $\Delta\leqq\Delta_0$ and $n\geqq n_0$ imply

$$\begin{aligned}P\{d(x^{\psi,n},\phi)<c\}&\geqq P\{d(X_{\Delta}^{\psi,n},\phi_{\Delta})<\rho\}\\&\quad-P\{\max_{i\leqq N-1}\max_{i\Delta\leqq t\leqq i\Delta+\Delta}|x^{\psi,n}(t)-x^{\psi,n}(i\Delta)|>\rho\}\\&\geqq P\{d(X_{\Delta}^{\psi,n},\phi_{\Delta})<\rho\}-\exp-K/\lambda_n.\end{aligned}\tag{A1.5}$$

In fact, if $\{\xi_n\}$ is bounded, then the $\exp-K/\lambda_n$ term can be deleted.

If $L(\beta, \psi(\cdot), \cdot)$ were constant on each $[i\Delta, i\Delta+\Delta)$ interval, then the convexity of $L(\cdot, \psi, s)$ and Jensen's inequality yields

$$\begin{aligned}\int_0^T L(\dot{\phi}(s), \psi(s), s) &= \sum_0^{N-1} L\left(\frac{1}{\Delta}\int_{i\Delta}^{i\Delta+\Delta} \dot{\phi}(s)\, ds, \psi(i\Delta), i\Delta\right)\Delta \\ &\leqq \sum_0^{N-1} \int_{i\Delta}^{i\Delta+\Delta} L(\dot{\phi}(s), \psi(i\Delta), i\Delta)\, ds \qquad \text{(A1.6)} \\ &= \int_0^T L(\dot{\phi}(s), \psi(s), s)\, ds \equiv S^{\psi}(T, \phi).\end{aligned}$$

But a simple approximation argument gives the same result, since $L(\beta, x, s) = h_0(s)L_0(\beta - b(x), x)$ and $L_0(\cdot, x)$ is convex. Thus, by (A1.3), (A1.4a) and (A1.6), for given $c>0$ and $h>0$, and for large n and small Δ,

$$\begin{aligned}\text{(A1.5)} &\geqq \exp -\frac{1}{\lambda_n}\left[\int_0^T L(\dot{\phi}_{\Delta}(s), \psi(s), s)\, ds + \frac{h}{2}\right] - (\exp - K/\lambda_n) \\ &\geqq \exp -\left[S^{\psi}(T, \phi) + \frac{h}{2}\right]\Big/\lambda_n - (\exp - K/\lambda_n) \qquad \text{(A1.7)}\end{aligned}$$

which is equivalent to (3.11) for large enough K.

In order to derive (3.12), interpret inequality (A1.4b) as follows: Let $\bar{x}_{\Delta}^{\psi,n}(\cdot)$ denote the piecewise linear interpolation of $\{x^{\psi,n}(i\Delta), 1 \leqq i \leqq N\}$. Then, by (A1.3),

$$\Phi_{\Delta,s}^{\psi} = \{\phi(\cdot) \in C_x[0, T]:$$
$$\phi(\cdot) \text{ piecewise linear (on the } [i\Delta, i\Delta+\Delta) \text{ intervals)}, S^{\psi}(T, \phi) \leqq s\}.$$

Thus $\Phi_{\Delta,s}^{\psi} \subset \Phi_s^{\psi}$.

By $\Phi_{\Delta,s}^{\psi} \subset \Phi_s^{\psi}$ and the estimate used for the $P(\max \max)$ term in the middle term of (A1.5), inequality (A1.4b) implies that for large n

$$\begin{aligned}P\{d(x^{\psi,n}, \Phi_s^{\psi}) > 2c\} &\leqq P\{d(\bar{X}_{\Delta}^{\psi,n}, \Phi_s^{\psi}) > c\} + \exp - K/\lambda_n \\ &\leqq P\{d(\bar{X}_{\Delta}^{\psi,n}, \Phi_{\Delta,s}^{\Psi}) > c\} + \exp - K/\lambda_n \qquad \text{(A1.8)} \\ &\leqq \exp -\left[s - \frac{h}{2}\right]\Big/\lambda_n + \exp - K/\lambda_n\end{aligned}$$

which is equivalent to (3.12), since K can be chosen to be arbitrarily large by making Δ small and n large.

Appendix 2.

RM *case.* (A1) and ($A3_{RM}$) are assumed.

If $\{\xi_n\}$ is bounded, then the various estimates of the form $\exp - K/\lambda_n$ required in § 3 are trivially satisfied, so we need only work with the Gaussian case, where $\xi_n = \sum_j g_{n-j}\theta_j$, $\{\theta_j\}$ i.i.d., $E\theta_j = 0$ and $\sum_n n|g_n| < \infty$, $g_n = 0$ for $n<0$ and $\lambda_n = a_n$. In fact, we do a slightly more general case where there is a $k<\infty$ such that

$$E \exp|\alpha'\theta| \leqq \exp k(|\alpha| + |\alpha|^2). \qquad \text{(A2.1)}$$

Define $B = \Sigma|g_i|$, $B_i = \sum_{j=i}^{\infty}|g_j|$ and $\bar{B} = \sum_i (i+1)|g_i| = \sum_{i=0}^{n} B_i$.

LEMMA A.1. *Let* $|a_j - \hat{a}_j|/\lambda_n \leqq \varepsilon$ *for* $j \in [n, m_n(T)]$, *and let* $T \geqq t_0 > t_1 \geqq 0$. *Then there is a constant* k_1 *such that (for* $c - k_1(t_2 - t_1) - k_1\lambda_n\bar{B} \geqq 0$)

$$P\left\{\sum_{j=m_n(t_1)}^{m_n(t_2)-1} |(a_j - \hat{a}_j)\xi_j| \geqq c\right\} \leqq \exp -\frac{1}{\lambda_n}\frac{(c - k_1(t_2-t_1) - k_1\lambda_n\bar{B})^2}{4\varepsilon^2[(t_2-t_1) + \lambda_n\bar{B}]k_1}. \qquad \text{(A2.2)}$$

Proof. We need only do the case of scalar valued θ. For any $\lambda > 0$, the exponential Chebyshev's inequality and (A2.1) and the estimate $\lambda_n[m_n(t_2) - m_n(t_1)] = O(t_2 - t_1)$ yield the existence of a $k_1 < \infty$ such that for any $\lambda > 0$

$$P\left\{\sum_{j=m_n(t_1)}^{m_n(t_2)-1} \frac{|a_j - \hat{a}_j|}{\lambda_n}|\xi_j| \geqq c/\lambda_n\right\}$$

$$\leqq P\left\{\varepsilon\lambda B \sum_{i=m_n(t_1)}^{m_n(t_2)-1} |\theta_i| + \varepsilon\lambda \sum_{i=-\infty}^{m_n(t_1)-1} B_{m_n(t_1)-i}|\theta_i| \geqq \lambda c/\lambda_n\right\}$$

$$\leqq (\exp(-\lambda c/\lambda_n))(\exp k(\varepsilon\lambda B + \varepsilon^2\lambda^2 B^2))^{m_n(t_2)-m_n(t_1)}$$

$$\cdot \exp k \sum_{i=0}^{\infty} (\varepsilon\lambda B_i + \varepsilon^2\lambda^2 B_i^2)$$

$$\leqq \exp[-\lambda c + k_1((t_2 - t_1) + \lambda_n \bar{B})(\varepsilon\lambda + \varepsilon^2\lambda^2)]/\lambda_n.$$

Minimizing the r.h.s. with respect to $\lambda > 0$ yields (A2.2). Q.E.D.

LEMMA A.2. *Let* $\sup_n \sup_{j\in[n,m_n(T)]} |\hat{a}_j/a_n| < \infty$. *Then for each* $K < \infty$, *and* $c > 0$, *there are* $n_0 < \infty$ *and* $\Delta_0 > 0$ *such that* $n \geqq n_0$ *and* $\Delta \leqq \Delta_0$ *imply that*

$$P\left\{\max_{i\leqq T/\Delta} \sum_{m_n(i\Delta)}^{m_n(i\Delta+\Delta)-1} \hat{a}_j|\xi_j| \geqq c\right\} \leqq \exp - K/\lambda_n. \tag{A2.3}$$

Proof. By Lemma A.1, there is a constant k_2 such that for large n and small enough Δ,

$$(\text{A2.3}) \leqq \frac{T}{\Delta} \exp - \frac{k_2}{\lambda_n \Delta},$$

which yields the result. Q.E.D.

Note that Lemma A.2 implies that for each K,

$$P\{\max_{i\Delta\leqq T} \max_{t\leqq\Delta} d(x^n(i\Delta), x^n(i\Delta + t)) \geqq c\} \leqq \exp - K/\lambda_n$$

for large n—and similarly for the $x^{\psi,n}(\cdot)$. This is what gives us the "almost compactness" of R_n—used in the proof of Theorem 3.1.

LEMMA A.3. *Define*

$$Y_{i+1} = x + \sum_{j=n}^{i} a_j[\bar{b}(Y_j) + b_0(Y_j)\xi_j] + \sigma_i,$$

$$\tilde{Y}_{i+1} = x + \sum_{j=n}^{i} a_j[\bar{b}(\tilde{Y}_j) + b_0(\tilde{Y}_j)\xi_j] + \rho_i,$$

where $\sup_i [|\rho_i| + |\sigma_i|] \leqq \bar{\rho}$. *Then for each* $c > 0$ *and* $K < \infty$, *there is a* $\rho_0 > 0$ *such that* $\bar{\rho} \leqq \rho_0$ *implies that*

$$P\{\sup_{n\leqq i<m_n(T)} |Y_i - \tilde{Y}_i| \geqq c\} \leqq \exp - K/\lambda_n.$$

Proof. Define $\delta_j = |Y_j - \tilde{Y}_j|$. There is a $k_0 < \infty$ such that

$$\delta_{i+1} \leqq |\rho_i - \sigma_i| + k_0 \sum_{j=n}^{i} a_j[\delta_j + \delta_j|\xi_j|].$$

By Gronwall's inequality

$$\delta_i \leqq \sup_i |\rho_i - \sigma_i| \exp k_0 \sum_{j=n}^{m_n(T)-1} a_j(1+|\xi_j|). \tag{A2.4}$$

By choosing the c and ε suitably in Lemma A.1 ($\varepsilon = 1$, $\hat{a}_j = 0$), we obtain that there are constants k_i such that for any large enough M and for large n,

$$P\left\{\sum_{n}^{m_n(T)-1} a_j|\xi_j| \geqq M\right\} \leq \exp - k_2[M - k_1]^2/\lambda_n.$$

Choosing M and ρ_0 appropriately yields the result. Q.E.D.

LEMMA A.4. *Equations* (3.15) *and* (3.17) *hold.*

Proof. The proofs for the two cases are quite similar and we only do (3.17). We need to show that if $d(x^n, \phi)$ and $d(\psi, \phi)$ are small then so is $d(x^{\psi,n}, \phi)$, modulo a set of sufficiently small probability. Writing ϕ_j^n and ψ_j^n for $\phi(t_j - t_n)$ and $\psi(t_j - t_n)$, respectively, we have

$$\begin{aligned} X_{i+1}^n &= x + \sum_{j=n}^{i} a_j[\bar{b}(X_j^n) + b_0(X_j^n)\xi_j], \\ X_{i+1}^{\psi,n} &= x + \sum_{j=n}^{i} a_j[\bar{b}(\psi_j^n) + b_0(\psi_j^n)\xi_j]. \end{aligned} \tag{A2.5}$$

We can write

$$\phi_{i+1}^n = x + \sum_{n}^{i} a_j[\bar{b}(\phi_j^n) + b_0(\phi_j^n)\xi_j] + \rho_i,$$

$$X_{i+1}^{\psi,n} = x + \sum_{n}^{i} a_j[\bar{b}(\phi_j^n) + b_0(\phi_j^n)\xi_j] + \sigma_j,$$

where

$$\begin{aligned} \rho_i &= [\phi_{i+1}^n - X_{i+1}^n] + \sum_{j=n}^{i} a_j[\bar{b}(X_j^n) - \bar{b}(\phi_j^n)] \\ &\quad + \sum_{j=n}^{i} a_j[b_0(X_j^n) - b_0(\phi_j^n)]\xi_j, \\ \sigma_i &= \sum_{j=n}^{i} a_j[\bar{b}(\psi_j^n) - \bar{b}(\phi_j^n)] + \sum_{j=n}^{i} a_j[b_0(\psi_j^n) - b_0(\phi_j^n)]\xi_j. \end{aligned}$$

By Lemma A.3, if $d(x^n, \phi)$ and $d(\phi, \psi)$ are small enough, then the $\sup_i [|\rho_i| + |\sigma_i|]$ is small, except on a set whose probability is bounded by $\exp - K/\lambda_n$ for large n. This yields (3.17). Q.E.D.

The KW case. We first prove the estimates needed for the first part of Theorem 3.3, and then do the more general case.

LEMMA A.5. *Assume* (A1), (A5) *and* ($A6_{KW}$), *but let* $\{\xi_n\}$ *be i.i.d. and either bounded or Gaussian. Let* $|(\tilde{a}_j - a_j)/a_j| \leqq \varepsilon$ *and* $|(\tilde{c}_j - c_j)/c_j| \leqq \varepsilon$. *Then for each* $c > 0$ *and* $K < \infty$, *there are* $\varepsilon_0 > 0$, $n_0 < \infty$ *and* $\Delta_0 > 0$ *such that for* $\varepsilon \leqq \varepsilon_0$, $n \geqq n_0$ *and* $\Delta \leqq \Delta_0$ *we have*

$$P\left\{\max_{i \leqq T/\Delta} \max_{k < m_n(i\Delta+\Delta)} \left| \sum_{m_n(i\Delta)}^{k-1} \frac{\tilde{a}_j}{\tilde{c}_j} b(X_j^n, \xi_j) \right| \geqq c \right\} \leqq \exp - K/\lambda_n. \tag{A2.6}$$

Similarly for ψ_j^n *replacing* X_j^n.

Proof. It is sufficient to prove (A2.6) without the $\max_{i \leqq T/\Delta}$ and for the scalar case where $b(x, \xi) = \xi$. The sum $\sum \tilde{a}_j \xi_j / \tilde{c}_j$ is a martingale. Hence for $\lambda > 0$

$$P\left\{ \max_{k < m_n(i\Delta+\Delta)} \sum_{m_n(i\Delta)}^{k-1} \frac{\tilde{a}_j}{\tilde{c}_j} \xi_j \geqq c \right\}$$

$$= P\left\{ \max_{k < m_n(i\Delta+\Delta)} \left(\exp \frac{\lambda}{\lambda_n} \sum_{m_n(i\Delta)}^{k-1} \tilde{a}_j \xi_j / \tilde{c}_j \right) \geqq \exp \lambda c / \lambda_n \right\}$$

$$\leqq (\exp - \lambda c / \lambda_n) E \exp \frac{\lambda}{\lambda_n} \sum_{m_n(i\Delta)}^{m_n(i\Delta+\Delta)-1} \tilde{a}_j \xi_j / \tilde{c}_j$$

$$= (\exp - \lambda c / \lambda_n) \exp \sum_{m_n(i\Delta)}^{m_n(i\Delta+\Delta)-1} H_2(\tilde{a}_j \lambda / \tilde{c}_j \lambda_n),$$

where $H_2(\alpha) = \log E \exp \alpha \xi$. Expanding H_2 as H_0 was expanded in Theorem 2.6, noting that the leading term in the expansion is

$$\tfrac{1}{2} H_{2,\alpha\alpha}(0) \sum_{m_n(i\Delta)}^{m_n(i\Delta+\Delta)-1} \lambda^2 \tilde{a}_j^2 / \lambda_n^2 \tilde{c}_n^2 \sim H_{2,\alpha\alpha}(0) h_2(i\Delta) \lambda^2 \Delta / 2\lambda_n,$$

and choosing λ and Δ appropriately (then repeating for $-\xi_j$ replacing ξ_j) yields (A2.6). Q.E.D.

We omit the proof of the following lemma.

LEMMA A.6. *Assume* (A1), (A6) *and let* $\bar{b}(\cdot)$ *be bounded and Lipschitz continuous. Let* $\{\xi_n\}$ *be i.i.d. and either bounded or Gaussian. In the Gaussian case, let* $b_0(x, \xi) = \xi$ *and in the bounded case, let* $b_0(\cdot, \xi)$ *be bounded and uniformly Lipschitz continuous. Then* (3.15) *and* (3.17) *hold for the* KW *case.*

We now turn to the more general KW case, under the conditions associated with (3.19). We need to show that (and also for ψ_j^n replacing X_j^n) for each $c > 0$ and $K < \infty$, there is an n_0 such that for $n \geqq n_0$ (recall that $r_j = a_j / c_j$)

$$\text{(A2.7)} \qquad P\left\{ \max_{i\Delta \leqq T} \max_{k < m_n(i\Delta+\Delta)} \left| \sum_{j=m_n(i\Delta)}^{k} r_j b_0(X_j^n) \xi_j \right| \geqq c \right\} \leqq \exp - K / \lambda_n.$$

It is enough to prove it without the $\max_{i\Delta \leqq T}$ term. Also, we need the KW analogue of Lemma A.4. Following the calculation in Lemma A.4, we have

$$\text{(A2.8)} \qquad |\phi_{i+1}^n - X_{i+1}^{\psi,n}| = |\sigma_i - \rho_i|,$$

where

$$\rho_i = [\phi_{i+1}^n - X_{i+1}^n] + \sum_{j=n}^{i} a_j [\bar{b}(X_j^n) - \bar{b}(\phi_j^n)] + \sum_{j=n}^{i} r_j [b_0(X_j^n) - b_0(\phi_j^n)] \xi_j,$$

$$\sigma_i = \sum_{j=n}^{i} a_j [\bar{b}(\psi_j^n) - \bar{b}(\phi_j^n)] + \sum_{j=n}^{i} r_j [b_0(\psi_j^n) - b_0(\phi_j^n)] \xi_j.$$

Since $\sum_{m_n(t_1)}^{m_n(t_2)} r_j \to \infty$ as $n \to \infty$ if $t_2 > t_1$, we need to be more careful in the KW case.

LEMMA A.7. *Under the conditions of Theorem* 3.3, (3.15), (3.17) *and* (A2.7) *hold for the* KW *case.*

Proof. Again, only the KW analogue of the case (3.17) dealt with in Lemma A.4 will be dealt with. The proof of (A2.7) is similar. The proof for (3.15) is a little harder but uses essentially the same estimates and we comment on it briefly after the proof. Let v_j denote either $b_0(X_j^n) - b_0(\phi_j^n)$ or $b_0(\psi_j^n) - b_0(\phi_j^n)$.

Part 1. Define

$$z_i = \sum_n^i r_j v_j \xi_j = \sum_n^i r_j v_j (A-I)^{-1}(A-I)\xi_j.$$

Absorbing $(A-I)^{-1}$ into v_j, we have

$$z_i = \sum_n^i r_j v_j (\xi_{j+1} - \xi_j) - \sum_n^i r_j v_j C\theta_j. \tag{A2.9}$$

By a partial summation, the first sum in (A2.9) is

$$r_i v_i \xi_{i+1} - r_n v_n \xi_n + \sum_n^{i-1} (r_{j+1}v_{j+1} - r_j v_j)\xi_{j+1}. \tag{A2.10}$$

We now get probability estimates for the terms in (A2.9) and (A2.10). To obtain (3.17), we will need to show that if $d(x^n, \phi)$ and $d(\phi, \psi)$ are small, then (modulo a set of appropriately small probability) $d(x^{\psi,n}, \phi)$ is small. Let τ_n denote the first escape time of x^n from the ρ-neighborhood of ϕ. Then we need only show that if $d(\phi, \psi)$ is small, so is[2] (modulo the appropriate set of small probability) $d(x^{\psi,n}(\cdot \cap \tau_n), \phi(\cdot \cap \tau_n))$.

Let $I_j = 1$ if $d(X_k^n, \phi) < \rho$ for $k \leqq j$, and set $I_j = 0$ otherwise. Since θ_j is mean zero and independent of the bounded v_j, $\sum_n^i r_j v_j I_j C\theta_j = Z_i^n$ is a martingale. Then (doing the scalar case and dropping C)

$$P\{\sup_{i \leqq m_n(T)} Z_i^n \geqq c\} \leqq (\exp - \lambda c/\lambda_n) D_n,$$

where (write $m = m_n(T) - 1$)

$$D_n = E \exp \frac{\lambda}{\lambda_n} \sum_n^m r_j v_j \theta_j I_j = E \exp \sum_n^{m-1} \frac{\lambda}{\lambda_n} r_j v_j \theta_j I_j \cdot \exp H_2\left(\frac{\lambda}{\lambda_n} r_m v_m I_m\right).$$

Expanding $H_2(\alpha)$ as $H_0(\alpha, x)$ was expanded in Theorem 2.6, bounding $|v_m I_m|$ (by a quantity which goes to zero as $\rho \to 0$ and $d(\phi, \psi) = \rho_1 \to 0$) and iterating, choosing λ appropriately, and repeating for $-\theta_j$ replacing θ_j yields that for each $c > 0$ and $K < \infty$ there are $\rho_0 > 0$ and $\rho_{10} > 0$ such that for $\rho \leqq \rho_0$, $\rho_1 \leqq \rho_{10}$,

$$P\{\sup_{n \leqq i \leqq m_n(T)} |Z_i^n| \geqq c\} \leqq \exp - K/\lambda_n \tag{A2.11}$$

for large n.

Part 2. We now show that for each $c > 0$ and $K < \infty$,

$$P\{\sup_{n \leqq i \leqq m_n(T)} |r_i \xi_i| \geqq c\} \leqq \exp - K/\lambda_n \tag{A2.12}$$

for large n. (Since $\{v_i\}$ is bounded, we need not include it in (A2.12).) We have $\xi_{i+1} = \sum_{j=-\infty}^i Q_{i-j}\theta_j$ where $|Q_n| \to 0$ geometrically as $n \to \infty$. Thus, it is sufficient to prove (A2.12) for a scalar $\{\xi_i, \theta_i\}$ case, where $\xi_{i+1} = \sum_{-\infty}^i f_{i-j}\theta_j$ and $|f_i| \leqq d^i$, $d < 1$, and some small c_1 replaces c. Now,

$$P\{r_i \xi_i \geqq c_1\} \leqq (\exp - \lambda c_1/\lambda_n) E \exp r_i \sum_j f_{i-j}\theta_j \lambda/\lambda_n$$

$$= (\exp - \lambda c_1/\lambda_n) \exp \sum_j H_2(r_i f_{i-j}\lambda/\lambda_n).$$

Expand $\sum_j H_2$ and note that the leading term in the expansion is

$$\frac{1}{2}\sum_j H_{2,\alpha\alpha}(0) r_i^2 g_{i-j}^2 \lambda^2/\lambda_n^2 \leqq (\text{constant}) a_i \lambda^2/\lambda_n.$$

[2] Define $a \cap b = \min(a, b)$. Thus $\phi(\cdot \cap \tau_n)$ is the function ϕ, stopped at τ_n: $\phi(s \cap \tau_n) = \phi(s)$ for $s \leqq \tau_n$ and $= \phi(s \cap \tau_n)$ for $s \geqq \tau_n$.

Choosing λ appropriately yields

$$P\{r_i\xi_i \geqq c_1\} \leqq \exp -2K/a_n\lambda_n \tag{A2.13}$$

for large n. The (scalar case) l.h.s. of (A2.12) is thus $\leqq [m_n(T)-n]\exp -2K/a_n\lambda_n$, which (together with a similar estimate for $-\xi_j$ replacing ξ_j and the use of $[m_n(T)-n] = O(1/a_n)$) yields (A2.12) for large n.

Part 3. We have

$$(r_{j+1}v_{j+1} - r_jv_j) = (r_{j+1}-r_j)v_j + (v_{j+1}-v_j)r_{j+1}.$$

For $v_j = [b_0(X_j^n) - b_0(\phi_j^n)]$, $(v_{j+1}-v_j) = O(a_j + r_j|\xi_j|) + [b_0(\phi_{j+1}^n) - b_0(\phi_j^n)]$. Thus we need exponential estimates for the maxima of the sums

$$\sum_{j=n}^{i} |r_{j+1}-r_j||\xi_j|, \quad \sum_{j=n}^{i} r_{j+1}(b_0(\phi_{j+1}^n) - b_0(\phi_j^n))\xi_j, \quad \sum_{j=n}^{i} r_jr_{j+1}|\xi_j|^2. \tag{A2.14}$$

We now do the third sum, replacing r_{j+1} by r_j and using (w.l.o.g.) the scalar model used in Part 2.

Since $r_j^2 = O(a_n\lambda_n)$ and $a_n[m_n(T)-n]$ is bounded and $\{\xi_j\}$ stationary, it is sufficient to show that for each $c_1 > 0$ and $K < \infty$,

$$P\left\{\frac{1}{N}\sum_{1}^{N}|\xi_j|^2 \geqq \frac{c_1}{\lambda_n}\right\} \leqq \exp - K/\lambda_n \tag{A2.15}$$

for large N and n. A straightforward calculation yields

$$\sum_{1}^{N}|\xi_i^2| \leqq \frac{2}{(1-d)^2}\left[\sum_{j=0}^{N-1}\theta_j^2 + \sum_{j=-\infty}^{-1}\theta_j^2 d^{-j}\right].$$

Thus by the usual exponential Chebychev's inequality, there is a constant k_2 such that the l.h.s. of (A2.15) is bounded above by

$$\left(\exp -\frac{\lambda c_1}{\lambda_n}\right) E \exp \frac{k_2\lambda}{N}\left[\sum_{j=0}^{N-1}\theta_j^2 + \sum_{j=-\infty}^{-1}\theta_j^2 d^{-j}\right].$$

In the bounded noise case $\exp k_2\lambda\theta_j^2/N \leqq \exp k_3\lambda/N$ for some constant k_3, and in the Gaussian case, for $2(E\theta_j^2)k_2\lambda/N < 1$, $E\exp k_2\lambda\theta_j^2/N \leqq (1-2k_2\lambda E\theta_j^2/N)^{-1}$. In either case for large N and n, we can choose λ to obtain (A2.15).

Appropriate exponential estimates of all the other terms can also be obtained by similar methods, and we omit the details. We note only that the condition $\sum_n^{m_n(T)-1}|r_{j+1}-r_j| \to^n 0$ is used to handle the first term of (A2.14). Upon assembling the estimates we obtain that for any $K < \infty$ and $c > 0$ and small enough ρ and $d(\phi, \psi)$

$$P\{\sup_{n \leqq i \leqq m_n(T)} |\phi_i^n - X_i^{\psi,n}| I_i \geqq c\} \leqq \exp - K/\lambda_n$$

for large n, which implies (3.17) for the KW case. Q.E.D.

Only a remark will be made on obtaining (3.15). For simplicity, set $\bar{b}(\cdot) = 0$. Write

$$X_{i+1}^{\psi,n} = X_i^{\psi,n} + r_ib_0(X_i^{\psi,n})\xi_i + \varepsilon_j^n,$$

$$X_{i+1}^n = X_i^n + r_ib_0(X_i^n)\xi_i, \qquad \varepsilon_i^n = r_i[b_0(\psi_i) - b_0(X_i^{\psi,n})]\xi_i,$$

$$B_{0i}^n(\xi_i) = \int_0^1 ds[b_0(X_i^n + s(X_i^{\psi,n} - X_i^n))\xi_i]_x,$$

$$\Delta_i^n = X_i^{\psi,n} - X_i^n.$$

Then

$$\Delta_{k+1}^n = \sum_{i=n}^{k} \prod_{j=i+1}^{k} (I + r_j B_{0j}^n(\xi_j))\varepsilon_i^n$$

$$= \prod_{n}^{k} (I + r_j B_{0j}^n(\xi_j)) \sum_{i=n}^{k} \left[\prod_{j=n}^{i} (I + r_j B_{0j}^n(\xi_j)) \right]^{-1} \varepsilon_i^n.$$

By the methods of Lemma A.7, we can show that for any $K < \infty$ and $c > 0$, there is an $M < \infty$ such that

$$P\left\{ \sup_{n \leqq k \leqq m_n(T)} \left| \prod_{n}^{k} (I + r_j B_{0j}^n(\xi)) \right| \geqq M \right\} \leqq \exp - K/\lambda_n,$$

$$P\{ \sup_{n \leqq k \leqq m_n(T)} |r_j B_{0j}^n(\xi_j)| > c \} \leqq \exp - K/\lambda_n.$$

Using these estimates, an argument similar to that of Lemma A.7 yields (3.15a).

REFERENCES

[1] J. SACKS, *Asymptotic distribution of stochastic approximation procedures*, Ann. Math. Statist., 29 (1953), pp. 373-405.

[2] H. J. KUSHNER AND H. HUANG, *Rates of convergence for stochastic approximation type algorithms*, this Journal, 19 (1981), pp. 87-105.

[3] A. D. VENTSEL AND M. I. FREIDLIN, *Large Deviations*, Springer-Verlag, Berlin, 1983.

[4] R. AZENCOTT, *Grandes deviations et applications*, Lecture Notes in Mathematics, 774, Springer-Verlag, Berlin, 1980.

[5] M. I. FRIEDLIN, *The averaging principle and theorems on large deviations*, Russian Math. Surveys 33, July-Dec., 1978, pp. 117-176.

[6] R. R. BAHADUR, *Some Limit Theorems in Statistics*, CBMS Regional Conference Series in Applied Mathematics 4, Society for Industrial and Applied Mathematics, Philadelphia, 1971.

[7] *Grandes deviations et applications statistiques*, Astérisque, 68, Société Mathématique de France, 1979.

[8] M. COTTRELL, J.-C. FORT AND G. MALGOUYRES, *Evénements rares pour l'étude de certains algorithmes stochastiques*, IEEE Trans. Automat. Control, AC-28 (1983), pp. 907-919.

[9] H. J. KUSHNER, *Asymptotic behavior of stochastic approximation and large deviations*, LCDS Report #83-1, Brown Univ., IEEE Trans. Automat. Control, (1984), to appear.

[10] ———, *Robustness and approximation of escape times and large deviations estimates for systems with small noise effects*, LCDS Report #82-5, Brown Univ., SIAM J. Appl. Math., 44 (1984), pp. 160-182.

[11] J. GARTNER, *On large deviations from the invariant measure*, Theory Prob. Appl., 22 (1977), pp. 24-39.

[12] A. P. KOROSTELEV, *Convergence of recursive stochastic algorithms under Gaussian disturbances*, Cybernetics, 15 (1979), pp. 537-544.

[13] ———, *Multistep procedures of stochastic optimization*, Theory Prob. Appl., 26 (1981), pp. 621-628.

[14] ———, *Stochastic Recurrent Processes*, Nauka, Moscow, 1984. (In Russian.)

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

004

STOCHASTIC OPTIMIZATION PROBLEMS WITH INCOMPLETE INFORMATION ON DISTRIBUTION FUNCTIONS*

YU. ERMOLIEV†, A. GAIVORONSKI† AND C. NEDEVA‡

Abstract. The main purpose of this paper is to discuss numerical optimization procedures, based on duality theory, for stochastic extremal problems in which the distribution function is only partially known. We formulate such problems as minimax problems in which the "inner" problem involves optimization with respect to probability measures. The latter problem is solved using generalized linear programming techniques. Then we state the dual problem to the initial stochastic optimization problem. Numerical procedures that avoid the difficulties associated with solving the "inner" problem are proposed.

Key words. stochastic optimization, moment problem, duality theory, minimax problems

1. Introduction. A conventional stochastic programming problem may be formulated with some generality as minimization of the function

$$T(x) = E_y v(x, y) = \int_Y v(x, y)\, dH(y) \tag{1}$$

subject to

$$x \in X \subset R^n, \tag{2}$$

where $y \in Y \subset R^m$ is a vector of random parameters, $H(y)$ is a given distribution function and $v(x, \cdot)$ is a random function possessing all the properties necessary for expression (1) to be meaningful [8].

In practice, we often do not have full information on $H(y)$; we sometimes only have some of its characteristics, in particular bounds for the mean value or other moments. Such information can often be written in terms of constraints

$$Q^k(H) = E_y q^k(y) = \int_Y q^k(y)\, dH(y) \leqq 0, \qquad k = \overline{1, l} \tag{3}$$

$$\int_Y dH(y) = 1, \tag{4}$$

where the $q^k(y)$, $k = \overline{1, l}$, are known functions. We could, for example, have the following constraints on joint moments:

$$c_{\tau_1, \tau_2, \cdots, \tau_m} \leqq E y_1^{\tau_1} \cdots y_m^{\tau_m} \leqq C_{\tau_1, \tau_2, \cdots, \tau_m}, \tag{5}$$

where $C_{\tau_1, \tau_2, \cdots, \tau_m}$, $c_{\tau_1, \tau_2, \cdots, \tau_m}$ are given constants.

Consider the following problem: find a vector x which minimizes

$$T(x) = \max_{H \in K} \int_Y v(x, y)\, dH(y), \tag{6}$$

subject to constraints (2), where K is the set of functions H satisfying constraints (3) and (4).

* Received by the editors December 1, 1983, and in revised form August 1, 1984.

† International Institute for Applied Systems Analysis, Laxenburg, Austria, and Glushkov Institute of Cybernetics, Kiev, USSR.

‡ Institute of Mathematics, Bulgarian Academy of Sciences, Sophia, Bulgaria.

Special cases of this problem have been studied in [4], [5], [24]. Under certain assumptions concerning the family K and the function $v(\cdot)$, the solution of the "inner" problem has a simple analytical form and hence (6) is reduced to a conventional nonlinear programming problem. The main purpose of this paper is to discuss numerical methods for the solution of problem (6) in the more general case. Sections 2, 3, and 4 deal with the reduction of this problem to minimax-type problems without randomized strategies and describe numerical methods based on some of the same ideas as generalized linear programming. A quite general method for solving the resulting minimax-type problems, in which the inner problem of maximization is not concave, is considered in § 5.

2. Optimization with respect to distribution functions. The possible methods of minimizing $T(x)$ depend on solution procedures for the following "inner" maximization problem: find a distribution function H that maximizes

$$Q^0(H) = Eq^0(y) = \int_Y q^0(y)\, dH(y) \tag{7}$$

subject to

$$Q^k(H) = Eq^k(y) = \int_Y q^k(y)\, dH(y) \leqq 0, \qquad k = \overline{1, l}, \tag{8}$$

$$\int_Y dH(y) = 1, \tag{9}$$

where q^k, $k = \overline{0, l}$, are given functions $R^m \to R^1$. This is a generalization of the known moments problem (see, for instance, [15], [16], [17]). It can also be regarded as a generalization of the nonlinear programming problem

$$\max \{q^0(y): q^k(y) \leqq 0, y \in Y, k = \overline{1, l}\}$$

to an optimization problem involving randomized strategies [7], [12], [14].

It appears possible to solve problem (7)–(9) by means of a modification of the revised simplex method [7], [13]. This modification is based on Krein's "geometrical approach" to the theory of moments [8], [15], [17]. Consider the set

$$Z = \{z: z = (q^0(y), q^1(y), \cdots, q^l(y)), y \in Y\}$$

and suppose that Z is compact. This will be true, for instance, if Y is compact and functions q_k, $k = \overline{0, l}$, are continuous. Consider also the convex hull of Z:

$$\operatorname{co} Z = \left\{ z: z = \sum_{t=1}^{N} p_t z^t,\ z^t \in Z,\ \sum_{r=1}^{N} p_t = 1,\ p_t \geqq 0,\ t = \overline{1, N} \right\},$$

where N is an arbitrary finite number. Then general results from convex analysis lead to

$$\operatorname{co} Z = \left\{ Q = (Q^0(H), Q^1(H), \cdots, Q^l(H)) | H \geqq 0,\ \int_Y dH = 1 \right\}. \tag{10}$$

Therefore problem (7)–(9) is equivalent to maximizing z_0 subject to

$$z = (z_0, z_1, \cdots, z_l) \in \operatorname{co} Z, \quad z_k \leqq 0, \quad k = \overline{1, l}.$$

According to the Caratheodory theorem, each point on the boundary of co Z can be

represented as a convex combination of at most $l+1$ points from Z:

$$\operatorname{co} Z = \left\{ z\colon z_k = \sum_{j=1}^{l+1} q^k(y^j)p_j,\ k=\overline{0,l},\ p_j \geqq 0,\ \sum_{j=1}^{l+1} p_j = 1,\ y^j \in Y \right\}.$$

Thus problem (7)-(9) is equivalent to the following generalized linear programming problem [2]: find points $y^j \in Y$, $j=\overline{1,t}$, $t \leqq l+1$ and real numbers p_j, $j=\overline{1,t}$, such that

$$\sum_{j=1}^{t} q^0(y^j)p_j = \max \tag{11}$$

subject to

$$\sum_{j=1}^{t} q^k(y^j)p_j \leqq 0, \quad k=\overline{1,l}, \tag{12}$$

$$\sum_{j=1}^{t} p_j = 1, \qquad p_j \geqq 0, \quad j=\overline{1,t}. \tag{13}$$

Consider arbitrary points $\bar{y}^j$, $j=\overline{1,l+1}$ (setting $t=l+1$), and for the fixed set $\{\bar{y}^1\bar{y}^2, \cdots, \bar{y}^{l+1}\}$ find a solution $\bar{p}=(\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_{l+1})$ of problem (11)-(13) with respect to p. Assume that $\bar{p}$ exists and that $(\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_{l+1})$ are the corresponding dual variables, i.e., solve the problem

$$\min u_{l+1} \tag{14}$$

subject to

$$q^0(\bar{y}^j) - \sum_{k=1}^{l} u_k q^k(\bar{y}^j) - u_{l+1} \leqq 0, \qquad j=\overline{1,l+1}, \tag{15}$$

$$u_k \geqq 0, \qquad k=\overline{1,l}. \tag{16}$$

Now let y be an arbitrary point of Y. Consider the following augmented problem of maximization with respect to $(p_1, p_2, \cdots, p_{l+1}, p)$: maximize

$$\sum_{j=1}^{l+1} q^0(\bar{y}^j)p_j + q^0(y)p \tag{17}$$

subject to

$$\sum_{j=1}^{l+1} q^k(\bar{y}^j)p_j + q^k(y)p \leqq 0, \qquad k=\overline{1,l}, \tag{18}$$

$$\sum_{j=1}^{l+1} p_j + p = 1. \tag{19}$$

It is clear that if there exists a point y^* such that

$$q^0(y^*) - \sum_{k=1}^{l} \bar{u}_k q^k(y^*) - \bar{u}_{l+1} > 0$$

then the solution $\bar{p}$ could be improved by dropping one of the columns $(q^0(\bar{y}^j), q^1(\bar{y}^j), \cdots, q^l(\bar{y}^j), 1)$, $j=\overline{1,l+1}$, from the basis and replacing it by the column $(q^0(y^*), q^1(y^*), \cdots, q^l(y^*), 1)$, following the revised simplex method. Point y^* could be defined as

$$y^* = \arg\max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} \bar{u}_k q^k(y) \right]. \tag{20}$$

Then a new solution p^* of (11)-(13) with fixed $y = y^*$ can be determined in the same way as $\bar{p}$, together with the dual variables u^*. This method gives us a conceptual framework for solving not only (6) but also some more general classes of problems.

If $y(x) = (y^1(x), y^2(x), \cdots, y^{l+1}(x))$, $p(x) = (p_1(x), p_2(x), \cdots, p_{l+1}(x))$ is a solution of the inner optimization problem for fixed x, then the function (6) may be nondifferentiable with subgradient

$$T_x(x) = \sum_{j=1}^{l+1} v_x(x, y^j(x)) p_j(x),$$

where v_x is a subgradient of function $v(\cdot, y)$. Nondifferentiable optimization techniques could therefore be used to minimize $T(x)$. The main difficulty of such an approach would be to obtain a solution of (20) and exact values of $y(x)$, $p(x)$ at each current point x^s for iterations $s = 0, 1, \cdots$. This last difficulty can sometimes be avoided by dealing with approximate solutions rather than precise values $y(x)$, $p(x)$, and using ε-subgradient methods (see [20], [21]). Generalized linear programming methods which do not require exact solutions of subproblem (20) are studied in § 4.

3. Duality relations. The duality relations for problem (7)-(9) enable us to find a more general approach to the solution of problem (6). Consider the following problem

$$\min_{u \in U^+} \max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} u_k q^k(y) \right], \tag{21}$$

where

$$U^+ = \{u: u = (u_1, u_2, \cdots, u_l), u_i \geqq 0, i = \overline{1, m}\}.$$

This problem can be regarded as dual to (7)-(9) or (11)-(13), but to explain this we must introduce some more definitions.

In what follows we shall use the same letter, say H, for both the distribution function and the underlying probabilistic measure, where this will not cause confusion. We shall denote by $Y^+(H)$ the collection of all closed subsets A of Y such that $H(A) = 1$, and by supp H the support set of distribution H, i.e.,

$$\text{supp } H = \bigcap_{A \in Y^+(H)} A.$$

Set

$$\psi(u) = \max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} u_k q^k(y) \right],$$

$$U^* = \{u^*: u^* \in U^+, \psi(u^*) = \min_{u \in U^+} \psi(u)\},$$

$$Y(u) = \left\{ y: y \in Y, \psi(u) = q^0(y) - \sum_{k=1}^{l} u_k q^k(y) \right\}.$$

Then the following generalization of the results given in [10] holds.

THEOREM 1. *Assume that*

1. *Y is compact and $q^k(y)$, $k = \overline{0, l}$, are continuous;*
2. *$0 \in \text{int co } Z$.*

Then

1. *Solutions to both problem* (7)-(9) *and problem* (21) *exist, and the optimal values of the objective functions of both problems are equal.*

2. *For any solution H^* of problem* (7)-(9) *there exists a* $u^* \in U^*$ *such that*

$$\operatorname{supp} H^* \subseteq Y(u^*).$$

In other words, the duality gap vanishes in nonlinear programs with randomized strategies. A proof of this theorem can be derived from general duality results [18], [21] and the theory of moments [15]. The proof given below is close to the ideas expressed in [21] and illustrates certain connections with results from conventional nonlinear programming.

Proof. From (10), problem (7)-(9) is equivalent to

$$\max \{z_0 \colon z = (z_1, z_2, \cdots, z_l) \in \operatorname{co} Z,\ z_k \leqq 0,\ k = \overline{1, l}\}, \tag{22}$$

where $Z = \{z \colon z = (q^0(y), q^1(y), \cdots, q^l(y)),\ y \in Y\}$. From assumption 1 of Theorem 1, co Z is a convex compact set and therefore a solution $z^* = (z_0^*, z_1^*, \cdots, z_l^*)$ to problem (22) exists. Let $L(u, z)$ be a Lagrange function for (22):

$$L(u, z) = z_0 - \sum_{k=1}^{l} u_k z_k.$$

From assumption 2,

$$z_0^* = \max_{z \in \operatorname{co} Z} \min_{u \in U^+} L(u, z) = \min_{u \in U^+} \max_{z \in \operatorname{co} Z} L(u, z).$$

Note that any other regularity assumption which secures the previous equality can be used instead of assumption 2. According to (10), there exists for any $z \in \operatorname{co} Z$ a distribution H such that

$$z_k = \int_Y q^k(y)\, dH(y), \qquad \int_Y dH(y) = 1, \qquad k = \overline{0, l}.$$

We therefore have

$$L(u, z) = \bar{L}(u, H) = \int_Y \left[q^0(y) - \sum_{k=1}^{l} u_k q^k(y) \right] dH(y)$$

and

$$\max_{z \in \operatorname{co} Z} L(u, z) = \max \left\{ \bar{L}(u, H) \middle| H \geqq 0, \int_Y dH(y) = 1 \right\}.$$

Obviously

$$\max \left\{ \int_Y \left[q^0(y) - \sum_{k=1}^{l} u_k q^k(y) \right] dH(y) \middle| H \geqq 0, \int_Y dH(y) = 1 \right\}$$

$$= \max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} u_k q^k(y) \right],$$

which proves the first part of the theorem.

Under the assumptions of the theorem we know that for any solution $(z_0^*, \cdots, z_l^*)$ there exists a $u \in U^*$ such that

$$z_0^* = \max_{z \in \operatorname{co} Z} L(u^*, z).$$

Thus, for any optimal distribution H^* we have

$$\int_Y q^0(y)\, dH^*(y) = \max\left\{\int_Y \left[q^0(y) - \sum_{k=1}^{l} u_k^* q^k(y)\right] dH(y) \mid H \geqq 0, \int_Y dH(y) = 1\right\}$$

$$= \max_{y \in Y}\left[q^0(y) - \sum_{k=1}^{l} u_k^* q^k(y)\right]$$

which proves the second part of the theorem.

Remark 1. From the duality theorem above we have

$$\max_{H \in K} \int v(x, y)\, dH(y) = \min_{u \in U^+} \max_{y \in Y}\left[v(x, y) - \sum_{k=1}^{m} u_k q^k(y)\right]$$

for each fixed $x \in X$, where $v(x, \cdot)$ is a continuous function. Problem (6) can then be reduced to a minimax-type problem as follows: minimize the function

$$\gamma(x, u) = \max_{y \in Y}\left[v(x, y) - \sum_{k=1}^{m} u_k q^k(y)\right]$$

with respect to $x \in X$, $u \geqq 0$.

Remark 2. Theorem 1 can be used to characterize optimal distributions for a variety of nonlinear optimization problems with distribution functions. The approach is, first, to state necessary optimality conditions through linearization and then to apply Theorem 1. This is illustrated in the following example.

Consider the optimization problem

(7a) $$\max g^0(H),$$

(8a) $$g^i(H) \leqq 0, \qquad i = \overline{1, m},$$

(9a) $$\int_Y dH(y) = 1,$$

where $g^i(H)$, $i = \overline{1, m}$, are nonlinear functionals depending on distribution functions H with support set Y.

THEOREM 1a. *Assume that the following statements are true*:

1. *Set Y is a compact subset of Euclidean space R^n.*
2. *For any distibutions H_1, H_2 such that* supp $H_1 \subseteq Y$, supp $H_2 \subseteq Y$ *we have*

$$g^i(H_1 + \alpha(H_2 - H_1)) = g^i(H_1) + \alpha \int_Y q^i(y, H_1)\, d(H_2 - H_1) + \varepsilon(\alpha, H_1, H_2),$$

where $i = \overline{1, m}$, $\alpha \in [0, 1]$ and $\varepsilon(\alpha, H_1, H_2)/\alpha \to 0$ as $\alpha \to 0$.

3. *Functions $q^i(y, H)$ are continuous in y for every H such that* supp $H \subseteq Y$; *for any H_1, H_2 such that* supp $H_1 \subseteq Y$, supp $H_2 \subseteq Y$ *we have*

$$|q^i(y, H_1) - q^i(y, H_2)| \leqq \left|\int_Y \lambda_i(y, H_1, H_2)\, d(H_1 - H_2)\right|,$$

where $|\lambda_i(y, H_1, H_2)| \leqq K_1 < \infty$ for some K_1 which does not depend on H_1, H_2.

4. *Functions $g^i(H)$, $i = \overline{1, m}$, are convex, i.e.,*

$$g^i(\alpha_1 H_1 + \alpha_2 H_2) \leqq \alpha_1 g^i(H_1) + \alpha_2 g^i(H_2),$$

$$\alpha_1 \geqq 0, \quad \alpha_2 \geqq 0, \quad \alpha_1 + \alpha_2 = 1.$$

5. *There exists an $\bar{H}$ such that* supp $\bar{H} \subseteq Y$ *and $g^i(\bar{H}) < 0$ for $i = \overline{1, m}$.*

Then:

1. *A solution of problem* (7a)-(9a) *exists.*
2. *For any such solution* H^* *we have*

$$\int_Y q^0(y, H^*)\, dH^* = \min_{u \in U^+} \varphi(u, H^*),$$

where

$$\varphi(u, H^*) = \max_{y \in Y} \left[q_0(y, H^*) - \sum_{i \in I_0} u_i q^i(y, H^*) \right] + \sum_{i \in I_0} u_i c_i,$$

$$I_0 = \{i: g^i(H^*) = 0\}, \qquad c_i = \int_Y q^i(y, H^*)\, dH^*(y).$$

3. *If* H^* *is a solution of* (7a)-(9a), *then for some* $u^* \in U^+$ *we have* $\operatorname{supp} H^* \subseteq Y(u^*, H^*)$, *where*

$$\varphi(u^*, H^*) = \min_{u \in U^+} \varphi(u, H^*),$$

$$Y(u, H) = \left\{ y: y \in Y, \varphi(u, H) = q^0(y, H) - \sum_{i \in I_0} u_i q^i(y, H) + \sum_{i \in I_0} u_i c_i \right\}.$$

Thus, the main assumptions of this theorem are the existence, continuity (in some sense) and boundedness of the directional derivatives of functions $g^i(H)$.

The following theorem is analogous to known results in linear programming and provides a useful stopping rule for methods of the type described in § 2 (see also § 4).

THEOREM 2. (optimality condition). *Let the assumptions of Theorem* 1 *hold and let* $\bar{p}$ *be a solution of problem* (11)-(13) *for fixed* $\bar{y} = (\bar{y}^1, \bar{y}^2, \cdots, \bar{y}^t)$, $\bar{y} \in R^{t \times m}$. *Then the pair* $\bar{y}, \bar{p}$ *is an optimal solution of problem* (11)-(13) *if and only if for given* $\bar{y}$ *there exists a solution* $(\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_{l+1})$ *of problem* (14)-(16) *such that*

$$q^0(y) - \sum_{k=1}^{l} \bar{u}_k q^k(y) - \bar{u}_{l+1} \leqq 0 \quad \text{for all } y \in Y.$$

Proof.

1. Suppose that $\bar{y}, \bar{p}$ is an optimal solution of problem (11)-(13), that $(u_1^*, u_2^*, \cdots, u_{l+1}^*)$ is a solution of problem (14)-(16) for given $\bar{y}$, and that $(\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_l)$ is a solution of the problem (21). Take

$$\bar{u}_{l+1} - \min_{u \in U^+} \max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} u_k q^k(y) \right] = \max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} \bar{u}_k q^k(y) \right].$$

We shall show that $\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_{l+1}$ is a solution of problem (14)-(16). Consider the two functions

$$\psi(u) = \max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} u_k q^k(y) \right],$$

$$\psi_1(u) = \max_{1 \leqq j \leqq t} \left[q^0(\bar{y}^j) - \sum_{k=1}^{l} u_k q^k(\bar{y}^j) \right].$$

According to Theorem 1

$$\sum_{j=1}^{t} q^0(\bar{y}^j) \bar{p}_j = \min_{u \in U^+} \psi(u).$$

Since problem (11)-(13) is dual to problem (14)-(16) for given $\bar{y}$, then

$$\sum_{j=1}^{t} q^0(\bar{y}^j)\bar{p}_j = \min_{u \in U^+} \psi_1(u).$$

Therefore

$$\psi(\bar{u}) = \min_{u \in U^+} \psi(u) = \min_{u \in U^+} \psi_1(u) = \psi_1(u^*) = u^*_{l+1}, \tag{23}$$

where

$$u^* = (u^*_1, u^*_2, \cdots, u^*_l).$$

Since $\bar{y}^j \in Y$, $j = \overline{1, t}$, then $\psi_1(u) \leqq \psi(u)$ for $u \in U^+$. In particular, $\psi_1(\bar{u}) \leqq \psi(\bar{u})$. But (23) implies

$$\psi_1(\bar{u}) \geqq \min_{u \in U^+} \psi_1(u) = \psi(\bar{u}),$$

and this gives $\psi_1(\bar{u}) = \psi(\bar{u}) = \psi_1(u^*)$. Hence $(\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_{l+1})$ is a solution of problem (14)-(16).

2. Suppose now that for given $\bar{y}$ there exists a solution $(\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_{l+1})$ of problem (14)-(16) such that

$$q^0(y) - \sum_{k=1}^{l} \bar{u}_k q^k(y) - \bar{u}_{l+1} \leqq 0, \qquad y \in Y.$$

From the duality between problems (11)-(13) and (14)-(16) we have

$$\sum_{j=1}^{t} q^0(\bar{y}^t)\bar{p}_j \geqq \psi(\bar{u}),$$

where $\bar{p} = (\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_t)$ is a solution of problem (11)-(13) for given $\bar{y}$. On the other hand, the duality between problems (21) and (11)-(13) leads to the inequality

$$\sum_{j=1}^{t} q^0(y^j)p_j \leqq \psi(\bar{u}),$$

for any $\{y^1, y^2, \cdots, y^t\}$, p satisfying (12)-(13). In other words,

$$\sum_{j=1}^{t} q^0(\bar{y}^j)\bar{p}_j \geqq \sum_{j=1}^{t} q^0(y^j)p_j$$

and this completes the proof.

The next theorem provides a means of deriving a solution to the initial problem (7)-(9) from a solution of problem (21), and is complementary to Theorem 1.

THEOREM 3. *Assume that the assumptions of Theorem* 1 *are satisfied and that* $\psi(\bar{u}) = \min\{\psi(u) | u \in U^+\}$. *Let* $\bar{y} = (\bar{y}^1, \bar{y}^2, \cdots, \bar{y}^t)$, *where* $\bar{y} \in R^{t \times m}$ *and* $\bar{y}^i \in Y(\bar{u})$, *and let* $\bar{p}$ *be a solution of problem* (11)-(13) *for given* $\bar{y}$. *Suppose also that there is a solution* p^* *to the inequalities* (12)-(13) *for* $y^j = \bar{y}^j$ *such that*

$$\sum_{j=1}^{t} q^k(\bar{y}^j)p^*_j \leqq 0, \qquad k \in I_0, \tag{24}$$

$$\sum_{j=1}^{t} q^k(\bar{y}^j)p^*_j = 0, \qquad k \in I_+, \tag{25}$$

where $I_+ = \{k | \bar{u}_k > 0\}$, $I_0 = \{k | \bar{u}_k = 0\}$.

Then the pair $\bar{y}, \bar{p}$ *is an optimal solution of problem* (11)-(13).

Proof. The vectors

$$q(\bar{y}^j) = (-q^1(\bar{y}^j), -q^2(\bar{y}^j), \cdots, -q^l(\bar{y}^j)), \qquad j = \overline{1, t}$$

are subgradients of the convex function

$$\psi_1(u) = \max_{1 \leqq j \leqq t} \left[q^0(\bar{y}^j) - \sum_{k=1}^{l} u_k q^k(\bar{y}^j) \right]$$

at a point $\bar{u}$. Therefore conditions (24)-(25) are necessary and sufficient for point $\bar{u}$ to be an optimal solution of the problem

$$\min \{\psi_1(u) | u \in U^+\}.$$

From the definition of the set $Y(\bar{u})$ we obtain $\psi(\bar{u}) = q^0(\bar{y}^j) - \sum_{k=1}^{l} \bar{u}_k q^k(\bar{y}^j)$, $j = \overline{1, t}$ and therefore $\psi(u) = \min \{\psi_1(u) | u \in U^+\}$ which gives

$$\min \{\psi_1(u) | u \in U^+\} = \min \{\psi(u) | u \in U^+\}. \tag{26}$$

The minimization of $\psi_1(u)$, $u \in U^+$, is equivalent to problem (14)-(16). Hence $\bar{u} = (\bar{u}_1, \bar{u}_2, \cdots, \bar{u}_l)$ together with $\bar{u}_{l+1} = \psi_1(\bar{u})$ give a solution of problem (14)-(16). Since problem (14)-(16) is dual to problem (11)-(13), then problem (11)-(13) has a solution, say $\bar{p} = (\bar{p}_1, \bar{p}_2, \cdots, \bar{p}_t)$, and

$$\bar{u}_{l+1} = \sum_{j=1}^{t} q^0(\bar{y}^j) \bar{p}_j.$$

This together with (26) yields

$$\sum_{j=1}^{t} q^0(\bar{y}^j) \bar{p}_j = \min \{\psi(u) | u \in U^+\},$$

and this completes the proof.

4. Algorithms. Theorems 2 and 3 justify a dual approach to problem (7)-(9) which may involve simultaneous approximation of both primal and dual variables subject to (24)-(25). In this section we consider several versions of generalized-linear-programming-based method discussed briefly in § 2. In all cases the current estimate of optimal solution satisfies (24)-(25) at each iteration. The convergence of such algorithms has been investigated in a number of papers [2], [23], under the assumption that the initial column entries for all previous iterations of subproblem (24) and the exact solutions at each iteration are stored in the memory. There are various ways of avoiding this expansion of the memory, mainly through selective deletion of these columns [1], [6], [19], [22]. The aim of this section is to discuss a way of avoiding not only the expansion of the memory, but also the need to have a precise solution of (20). The last is important in connection with initial problem (6), as mentioned in § 2.

Description of Algorithm 1. Fix points $y^{0,1}, y^{0,2}, \cdots, y^{0,l+1}$ and solve problem (11)-(13) with respect to p for $y^j = y^{0,j}$, $j = \overline{1, l+1}$ and $t = l+1$. Suppose that a solution $p^0 = (p_1^0, p_2^0, \cdots, p_{l+1}^0)$ to this problem exists. Let $u^0 = (u_1^0, u_2^0, \cdots, u_{l+1}^0)$ be a solution of the dual problem (14)-(16) with respect to u. The vector u^0 satisfies the following constraints for $y \in \{y^{0,1}, y^{0,2}, \cdots, y^{0,l+1}\}$:

$$q^0(y) - \sum_{k=1}^{l} u_k^0 q^k(y) - u^{l+1} \leqq 0, \qquad u_k^0 \geqq 0, \quad k = \overline{1, l}. \tag{27}$$

If u^0 satisfies condition (27) for all $y \in Y$, then the pair $\{y^{0,1}, t^{0,2}, \cdots, y^{0,l+1}\}$, p^0 is a solution of the original problem (11)-(13). If this is not the case, consider a new point

y^0 such that

$$\Delta(y^0, u^0) = q^0(y^0) - \sum_{k=1}^{l} u_k^0 q^k(y^0) - u_{l+1}^0 > 0$$

and

$$q^0(y^0) - \sum_{k=1}^{l} u_k^0 q^k(y^0) \geqq \max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} u_k^0 q^k(y) \right] - \varepsilon_0$$

for some $\varepsilon_0 > 0$.

Denote by $p^1 = (p_1^1, p_2^1, \cdots, p_{l+1}^1)$ a solution of the augmented problem (17)-(19) with respect to p for fixed $\bar{y}^j = y^{0,j}$, $y = y^0$. We shall use $y^{1,1}, y^{1,2}, \cdots, y^{1,l+1}$ to denote those points $y^{0,1}, \cdots, y^{0,l+1}, y^0$ that correspond to the basic variables of solution p^1.

Thus, the first step of the algorithm is terminated and we pass to the next step: determination of u^1, y^1, etc. In general, after the sth iteration we have points $y^{s,1}, y^{s,2}, \cdots, y^{s,l+1}$, a solution $p^s = (p_1^s, p_2^s, \cdots, p_{l+1}^s)$ and the corresponding solution $u^s = (u_1^s, u_2^s, \cdots, u_{l+1}^s)$ to the dual problem (14)-(16). For an $\varepsilon_s > 0$, find y^s such that

$$\Delta(y^s, u^s) = q^0(y^s) - \sum_{k=1}^{l} u_k^s q^k(y^s) - u_{l+1}^s > 0$$

and

$$q^0(y^s) - \sum_{k=1}^{l} u_k^s q^k(y^s) \geqq \max_{y \in Y} \left[q^0(y) - \sum_{k=1}^{l} u_k^s q^k(y) \right] - \varepsilon_s.$$

If we do not obtain $\Delta(y^s, u^s) > 0$ for decreasing values of ε_s we arrive at an optimal solution; otherwise we have to solve the augmented problem (17)-(19) for $\bar{y}^j = y^{s,j}$, $y = y^s$.

Denote by $y^{s+1,1}, y^{s+1,2}, \cdots, y^{s+1,l+1}$ those points from $\{y^{s,1}, y^{s,2}, \cdots, y^{s,l+1}\} \cup y^0$ that correspond to the basic variables of the solution p^{s+1} of the problem (17)-(19). The pair $\{y^{s+1,1}, y^{s+1,2}, \cdots, y^{s+1,l+1}\}$, p^{s+1} is the new approximate solution to the original problem.

Define

$$I_s^0 = \{k: u_k^s = 0\}, \; I_s^1 = \{k: u_k^s > 0\}$$

and

$$A_s = \{e: e = (e_1, e_2, \cdots, e_l), \|e\| = 1, e_k \geqq 0 \text{ for } k \in I_s^0 \text{ and arbitrary } e_k \text{ for } k \in I_s^1\}.$$

A_s is actually a set of feasible directions for set U^+ at point u^s. Let

$$\gamma_s = \max_{e \in A_s} \min_{j: p_j^s > 0} \sum_{k=1}^{l} q^k(y^{s,j}) e_k.$$

Note that $\gamma_s \leqq 0$ if the solution of the problem (17)-(19) exists because

$$\text{co}\,\{(q^1(y^{s,j}), q^2(y^{s,j}), \cdots, q^l(y^{s,j})), \forall j: p_j^s > 0\}$$

is a set of subgradients of the function

$$\max_{j: p_j^s > 0} \left\{ q^0(y^{s,j}) - \sum_{k=1}^{l} u_k q^k(y^{s,j}) \right\}$$

at point u^s, and this function has a minimum at u^s.

In order to prove that this method is convergent we require, broadly speaking, that $\gamma_s < 0$ and tends to zero only as we approach the optimal solution.

Consider the functions

$$\psi^s(u) = \max_{1 \leqq j \leqq l+1} \left[q^0(y^{s,j}) - \sum_{k=1}^{l} u_k q^k(y^{s,j}) \right].$$

THEOREM 4. *Let the conditions of Theorem* 1 *be satisfied, and the following additional conditions hold*:

1. *There exists a nondecreasing function* $\tau(\alpha)$, $\alpha \in [0, \infty)$, $\tau(0) = 0$, $\tau(\alpha) > 0$ *for* $\alpha > 0$, *and*

$$\gamma_s \leqq -\tau(\psi(u^s) - \psi^s(u^s)). \tag{28}$$

2. $\varepsilon_s > 0$, $\varepsilon_s \to 0$ *for* $s \to \infty$.

Then any convergent subsequence of sequence $\{y^{s,1}, y^{s,2}, \cdots, y^{s,l+1}\}$, p^s *converges to a solution of problem* (7)-(9).

Note that it is not necessary to know ε_s exactly; all we need is $\varepsilon_s \to 0$ as $s \to \infty$.

Proof.

1. First let us prove that the sequence $\{u^s\}$ is bounded. Suppose, arguing by contradiction, that there exists a subsequence $\{u^{s_r}\}$ such that $\|u^{s_r}\| \to \infty$ as $r \to \infty$. Assumption 2 of Theorem 1 implies that $\psi(u^{s_r}) \to \infty$ and therefore that $\psi(u^{s_r}) - \psi^{s_r}(u^{s_r}) \to \infty$, since $\psi^{s_r}(u^{s_r}) \leqq \min_{u \in U^+} \psi(u)$. Hence, there exist $\bar{r}$ and $\delta > 0$ such that for $r > \bar{r}$,

$$\gamma_{s_r} \leqq -\delta.$$

Now let us fix an arbitrary point $\bar{u} \in U^+$ and estimate $\psi(\bar{u})$. We obtain

$$\psi(\bar{u}) \geqq \psi^{s_r}(\bar{u}) \geqq \psi^{s_r}(u^{s_r}) = \sup_{h \in G^{s_r}} (g, \bar{u} - u^{s_r}),$$

where G^{s_r} is a set of subgradients of function ψ^{s_r} at point u^{s_r}. The definition of ψ^s implies that

$$G^{s_r} \supseteq \operatorname{co}\{-q^1(y^{s_r,i}), -q^2(y^{s_r,i}), \cdots, -q^l(y^{s_r,i}), \forall i: p_i^{s_r} > 0\} = D^{s_r}$$

and therefore (28) leads to

$$\begin{aligned}
\psi(\bar{u}) &\geqq \psi^{s_r}(u^{s_r}) + \sup_{g \in D^{s_r}} (g, \bar{u} - u^{s_r}) \\
&\geqq \psi^{s_r}(u^{s_r}) + \|\bar{u} - u^{s_r}\| \min_{e \in A_{s_r}} \max_{i: p_i^{s_r} > 0} \left(-\sum_{k=1}^{l} e_k q^k(y^{s_r,i}) \right) \\
&= \psi^{s_r}(u^{s_r}) - \|\bar{u} - u^{s_r}\| \max_{e \in A_{s_r}} \min_{i: p_i^{s_r} > 0} \sum_{k=1}^{l} e_k q^k(y^{s_r,i}) \\
&\geqq \psi^{s_r}(u^{s_r}) + \delta \|\bar{u} - u^{s_r}\|.
\end{aligned}$$

This last inequality yields $\psi(\bar{u}) \to \infty$ if $\|u^{s_r}\| \to \infty$, and therefore sequence $\{u^s\}$ is bounded.

2. We shall now estimate the evolution of the quantity $w_s = \psi^s(u^s)$, where

$$u_s = \arg \min_{u \in U^+} \psi^s(u).$$

Using the same argument as in part 1 of the proof we obtain:

$$\begin{aligned} w_{s+1} &= \psi^{s+1}(u^{s+1}) = \psi^{s}(u^{s+1}) \\ &\geqq \psi^{s}(u^{s}) + \sup_{g \in D^s} (g, u^{s+1} - u^{s}) \\ &\geqq \psi^{s}(u^{s}) + \|u^{s+1} - u^{s}\| \min_{e \in A_s} \max_{i: p_i^s > 0} \left(- \sum_{k=1}^{l} e_k q^k(y^{s,i}) \right) \\ &= \psi^{s}(u^{s}) - \|u^{s+1} - u^{s}\| \max_{e \in A_s} \min_{i: p_i^s > 0} \sum_{l=1}^{l} e_k q^k(y^{s,i}) \\ &\geqq \psi^{s}(u^{s}) + \tau(\psi(u^{s}) - \psi^{s}(u^{s}))\|u^{s+1} - u^{s}\|. \end{aligned}$$

Sequence $\{u^s\}_{s=0}^{\infty}$ is bounded and so $\psi(u^s) = \sup_{y \in Y} (q^0(y) - \sum_{i=1}^{l} u_i^s q^i(y))$ must also be bounded; thus $\psi^s(u^s)$ is bounded since $\psi^s(u^s) \leqq \psi(u^s)$. This together with the previous inequality immediately gives

$$\tau(\psi(u^{s}) - \psi^{s}(u^{s}))\|u^{s+1} - u^{s}\| \to 0,$$

which implies

$$\min \{\psi(u^{s}) - \psi^{s}(u^{s}), \|u^{s+1} - u^{s}\|\} \to 0. \tag{29}$$

Now consider any convergent subsequence $\{u^{s_r}\}$ of sequence $\{u^s\}$. We can assume from (29) without loss of generality that either $\|u^{s_r} - u^{s_r+1}\| \to 0$ or $\psi(u^{s_r}) - \psi^{s_r}(u^{s_r}) \to 0$. In the latter case we get $\psi^{s_r}(u^{s_r}) \to \min_{u \in U^+} \psi(u) = \psi^*$ because $\psi(u^{s_r}) \geqq \psi^*$ and $\psi^{s_r}(u^{s_r}) \leqq \psi^*$. In the case $\|u^{s_r} - u^{s_r+1}\| \to 0$ we get the following:

$$\begin{aligned} \psi(u^{s_r}) - \psi^{s_r}(u^{s_r}) &= \psi(u^{s_r}) - \psi^{s_r+1}(u^{s_r}) + \psi^{s_r+1}(u^{s_r}) \\ &\quad - \psi^{s_r+1}(u^{s_r+1}) + \psi^{s_r}(u^{s_r+1}) - \psi^{s_r}(u^{s_r}) \\ &\leqq \varepsilon_{s_r} + \|u^{s_r} - u^{s_r+1}\| \left(\sup_{g \in G^{s_r+1}} \|g\| + \sup_{g \in G^{s_r}} \|g\| \right) \end{aligned}$$

so that once again $\psi(u^{s_r}) - \psi^{s_r}(u^{s_r}) \to 0$ and we obtain $\psi^{s_r}(u^{s_r}) \to \min_{u \in U^+} \psi(u)$. However, according to Theorem 1 $\min_{u \in U^+} \psi(u)$ is the optimal solution of the initial problem; $\min_{u \in U^+} \psi^s(u)$ is the optimal solution of problem (11)-(13). Therefore the solution of (11)-(13) tends to the solution of the initial problem, and any convergent subsequence of sequence $\{y^{s,1}, y^{s,2}, \cdots, y^{s,l+1}\}$, p^s, where $s = 0, 1, \cdots$ converges to the optimal solution of the initial problem.

This method can be viewed as the generalized linear programming method applied to problem (7)-(9) [3], [21], [23]. It drops all points $y^{s,i}$ which do not correspond to basic variables. Theorem 4 shows that in some (rare) cases this method does not converge; however, this is not surprising because in certain cases the simplex method does not converge either. It may be possible to modify the algorithm in different ways to ensure convergence.

If we keep all previous points $y^{0,1}, y^{0,2}, \cdots, y^{0,l+1}, y^0, y^1, \cdots$ and solve problem (14)-(16) with an increasing number of corresponding columns, then the method appears to be a form of Kelley's method for minimizing function $\psi(u)$, which converges under the assumptions of Theorem 1. However, it is impossible to allow the set of points to increase *ad infinitum* in practical computations.

In the following modification of the algorithm presented above some nonbasic columns are dropped when an additional inequality is satisfied.

Description of Algorithm 2.

1. We first choose a sequence of positive real numbers $\{\mu_s\}_{s=0}^{\infty}$, take $r_0 = 0$ and select initial points $\{y^{0,1}, y^{0,2}, \cdots, y^{0,l+1}\}$ such that problem (11)-(13) has a solution with respect to p for $y^j = y^{0,j}$, $j = \overline{1, l+1}$. Let p^0 be a solution of this problem and u^0 be the corresponding dual variables. We then have to find y^0 such that

$$q^0(y^0) - \sum_{k=1}^{l} u_k^0 q^k(y^0) \geqq \psi(u^0) - \varepsilon_0,$$

where ε_0 is a positive number. If for any ε_0 and the corresponding y_0 we have

$$\Delta(y^0, u^0) = q^0(y^0) - \sum_{k=1}^{l} u_k^0 q^k(y^0) - u_{l+1}^0 \leqq 0,$$

then the pair $\{y^{0,1}, y^{0,2}, \cdots, y^{0,l+1}\}, p^0$ is an optimal solution of problem (7)-(9). Otherwise it is necessary to select ε_0, y^0 such that $\Delta(y^0, u^0) > 0$ and take $\Delta_0 = \Delta(y^0, u^0)$.

Suppose that after the sth iteration we have points $y^{s,j}, j = \overline{1, l_s}$, a solution $p^s = (p_1^s, p_2^s, \cdots, p_{l_s}^s)$ of problem (11)-(13) for $y^j = y^{s,j}, j = \overline{1, l_s}$, $t = l_s$, a corresponding solution $u^s = (u_1^s, u_2^s, \cdots, u_{l+1}^s)$ of the dual problem (14)-(16), a positive integer number r_s and a positive number Δ_s.

2. Find an approximate solution y^s such that

$$\left(q^0(y^s) - \sum_{k=1}^{l} u_k^s q^k(y^s)\right) > \psi(u^s) - \varepsilon_s$$

and

$$q^0(y^s) - \sum_{k=1}^{l} u_k^s q^k(y^s) - u_{l+1}^s = \Delta(y^s, u^s) > 0.$$

If this is not possible for all $\varepsilon_s > 0$ then we have arrived at a solution. Otherwise consider the following two cases:

(a) $\Delta(y^s, u^s) \leqq (1 - \mu_{r_s})\Delta_s$. In this case take $\Delta_{s+1} = \Delta(y^s, u^s)$, $l_{s+1} = l+1$, $r_{s+1} = r_s + 1$ and denote by $y^{s+1}, y^{s+1,2}, \cdots, y^{s+1,l+1}$ those points from $\{y^{s,1}, y^{s,2}, \cdots, y^{s,l+1}\} \cup y^s$ that correspond to the basic variables of the solution p^{s+1}.

(b) $\Delta(y^s, u^s) > (1 - \mu_{r_s})\Delta_s$. In this case take

$$\Delta_{s+1} = \Delta_s, \quad l_{s+1} = l_s + 1, \quad r_{s+1} = r_s, \quad y^{s+1,j} = y^{s,j}, \quad j = \overline{1, l_s}, \quad y^{s+1,l_{s+1}} = y^s.$$

Find a solution of problem (11)-(13) for $t = l_{s+1}$, $y^j = y^{s+1,j}, j = \overline{1, l_{s+1}}$ and the corresponding dual variables u^{s+1}, and proceed to the next iteration.

THEOREM 5. *Suppose that the conditions of Theorem* 1 *are satisfied and the following additional conditions hold*:

1. $\varepsilon_s > 0$, $\varepsilon_s \to 0$, $\sum_{s=0}^{\infty} \mu_s = \infty$, $\mu_s \geqq 0$.

2. $\varepsilon_s / \mu_s \to 0$.

Then $\sum_{i=1}^{l_s} p_i^s q_0(y^{s,i})$ *tends to the optimal value of problem* (7)-(9).

Proof.

1. Suppose that the inequality $\Delta(y^s, u^s) \leqq (1 - \mu_{r_s})\Delta_s$ is satisfied only a finite number of times. This implies the existence of a number s_0 such that for $s \geqq s_0$ the method turns into Kelley's cutting-plane algorithm for minimization of the convex function $\psi(u)$, where the values of $\psi(u)$ are calculated with an error that tends to zero. From assumption 2 of Theorem 1, $\psi(u)$ has a bounded set of minimum points and the initial approximating function $\psi^{s_0}(u)$ has a minimum. Thus such a method

would converge to the minimum of function $\psi(u)$ and $\Delta(y^s, u^s) \leqq \psi(u^s) - \psi^s(u^s) \to 0$, which contradicts the assumption that for $s \geqq s_0$ the inequality $\Delta(y^s, u^s) \leqq (1-\mu_{r_s})\Delta_s$ is not satisfied.

2. There exists a subsequence s_k such that

$$\Delta(y^{s_k}, u^{s_k}) \leqq (1-\mu_k)\Delta_{s_k} = (1-\mu_k)\Delta(y^{s_{k-1}}, u^{s_{k-1}}).$$

From the definition of the algorithm we have

$$\psi(u^{s_k}) - \psi^{s_k}(u^{s_k}) - \varepsilon_{s_k} \leqq \Delta(y^{s_k}, u^{s_k}) \leqq \psi(u^{s_k}) - \psi^{s_k}(u^{s_k})$$

and therefore

$$\psi(u^{s_k}) - \psi^{s_k}(u^{s_k}) - \varepsilon_{s_k} \leqq (1-\mu_k)(\psi(u^{s_{k-1}}) - \psi^{s_{k-1}}(u^{s_{k-1}})).$$

Making the substitution $\psi(u^s) - \psi^s(u^s) = w^s$ we obtain

$$w^{s_k} \leqq (1-\mu_k) w^{s_{k-1}} + \varepsilon_{s_k}.$$

Suppose now that there is $\alpha > 0$ and m such that $w^{s_k} \geqq \alpha$ for $k \geqq m$. From the assumption 2 we have that $\varepsilon_{s_k}/\mu_k < \alpha/2$ for sufficiently large k, hence

$$w^{s_k} \leqq w^{s_{k-1}} - \mu_k \frac{\alpha}{2}, \qquad k > m$$

and

$$w^{s_n} \leqq w^{s_m} - \frac{\alpha}{2} \sum_{i=m+1}^{n} \mu_i$$

for $n > m$. This gives contradiction with boundedness of w^{s_k} because $\sum_{i=1}^{\infty} \mu_i = \infty$. Therefore sequence t_k exists such that $w^{t_k} \to 0$, which implies $\psi^{t_k}(u^{t_k}) \to \min_{u \in U^+} \psi(u)$ because $\psi^{t_k}(u^{t_k}) \leqq \psi(u)\ \forall u \in U^+$. But for any s we have $\psi^{s+1}(u^{s+1}) \geqq \psi^s(u^s)$, which together with $\psi^{t_k}(u^{t_k}) \to \min_{u \in U^+} \psi(u)$ leads to $\psi^s(u^s) \to \min_{u \in U^+} \psi(u)$. However, $\psi^s(u^s) = \sum_{i=1}^{l_s} p_i^s q^0(y^{s,i})$ and $\min_{u \in U^+} \psi(u)$ is an optimal value of problem (7)-(9) due to Theorem 1. This completes the proof.

Remark. It is not necessary to know the ε_s precisely; we need only that $\varepsilon_s/\mu_s \to 0$. For example, μ_s may be a small positive constant and the precision with which problem (20) is solved may gradually increase. This will give automatic fulfillment of the assumption 2.

Various ways of dropping the cuts in cutting-plane methods have been suggested in [2], [3]. The following method, which keeps only $l+1$ points at each iteration, is close to some methods in [3].

Instead of problem (14)-(16), solve the following problem at each iteration:

$$\min\ (u_{l+1} + \varepsilon \|u^s - u\|^2),$$

$$q^0(\bar{y}^j) - \sum_{k=1}^{l} q^k(\bar{y}^j) u_k - u_{l+1} \leqq 0, \qquad j = \overline{1, l+1},$$

$$u_k \geqq 0, \qquad k = \overline{1, l},$$

where u^s is the solution at the previous iteration. That this modified version converges can be proved in a similar way to Theorem 5.

5. Stochastic procedure. By a corollary of Theorem 1, Problem 6 is reduced to a minimax problem with a possibly nonconcave inner problem of maximization and a convex final problem of minimization. A vast amount of work has been done on minimax problems but virtually all of the existing numerical methods fail if the inner

problem is nonconcave. To overcome this difficulty, we adopt an approach based on stochastic optimization techniques.

Consider the faily general minimax problem

$$\min_{x\in X}\max_{y\in Y} f(x, y), \tag{30}$$

where $f(x, y)$ is a continuous function of (x, y) and a convex function of x for each $y\in Y$, $X\subset R^n$, $Y\subset R^m$. Although

$$F(x)=\max_{y\in Y} f(x, y) \tag{31}$$

is a convex function, to compute a subgradient

$$F_x(x)=f_x(x, y)|_{y=y(x)},$$

$$y(x)=\arg\max_{y\in Y} f(x, y), \tag{32}$$

$$f_x(x, y)\in\partial_x f(x, y)=\{g|f(z, y)-f(x, y)\geqq\langle g, z-x\rangle, \forall z\in X\}$$

requires a solution $y(x)$ of nonconcave problem (32). In order to avoid the difficulties involved in computing $y(x)$ one could try to approximate Y by an ε-set Y_ε and consider

$$y^\varepsilon(x)=\arg\max_{y\in Y_\varepsilon} f(x, y)$$

instead of $y(x)$. But, in general, this would require a set Y_ε containing a very large number of elements. An alternative is to use the following ideas. Consider a sequence of sets Y_s, $s=0, 1, \cdots$ and the sequence of functions

$$F^s(x)=\max_{y\in Y_s} f(x, y).$$

It can be proved (see, for instance, [9]) that, under certain assumptions concerning the behavior of sequence F^s, the sequence of points generated by the rule

$$x^{s+1}=x^s-\rho_s F_x^s(x^s), \qquad s=0, 1, \cdots, \tag{33}$$

$$F_x^s(x^s)\in\partial F^s(x^s)=\{g|F^s(x)-F^s(x^s)\geqq\langle g, x-x^s\rangle, \forall x\}$$

(where the step size ρ_s satisfies assumptions such as $\rho_s\geqq 0$, $\rho_s\to 0$, $\sum_{s=0}^{\infty}\rho_s=\infty$) tends, in some sense, to follow the time-path of optimal solutions: for $s\to\infty$

$$\lim[F^s(x^s)=\min F^s(x)]=0.$$

We will show below how Y_s (which depends on x^s) can be chosen so that we obtain the convergence

$$\min F^s(x)\to\min F(x),$$

where Y_s contains only a finite number $N_s\geqq 2$ of random elements.

The principal peculiarity of procedure (33) is its nonmonotonicity. Even for differentiable functions $F^s(x)$, there is no guarantee that x^{s+1} will belong to the domain

$$\{x|F^t(x)<F^t(x^s)\}, \qquad t\geqq s+1$$

of smaller values of functions F^{s+1}, F^{s+2}, $\cdots$. The procedure adopted here is the following (see [11]).

We start by choosing initial points x^0, y^0, a probability measure P on set Y and an integer $N_0\geqq 1$. Suppose that after the sth iteration we have arrived at points x^s, y^s.

The next approximations x^{s+1}, y^{s+1} are then constructed in the following way. Choose $N_s \geqq 1$ points

$$y^{s,1}, y^{s,2}, \cdots, y^{s,N_s}$$

which are sampled from the distribution corresponding to the measure P, and determine the set

$$Y_s = \{y^{s,1}, y^{s,2}, \cdots, y^{s,N_s}\} \cup y^{s,0},$$

where $y^{s,0} = y^s$. Take

$$y^{s+1} = \operatorname{Arg} \max_{y \in Y_s} f(x^s, y)$$

and compute

$$x^{s+1} = \pi_X[x^s - \rho_s f_x(x^s, y^{s+1})], \qquad s = 0, 1, \cdots$$

where ρ_s is the step size and π_X is the result of a projection operation on X.

Before studying the convergence of this algorithm, we should first clarify the notation used:

$P(A)$ is a probability measure of set $A \supseteq Y$,

$X^* = \operatorname{Arg} \min_{x \in X} F(x)$,

$Y_\varepsilon^*(x) = \{y | y \in Y, f(x, y) \geqq F(x) - \varepsilon\}, \qquad \varepsilon > 0$,

$p(\varepsilon, x) = P\{Y_\varepsilon^*(x)\}$,

$\gamma(\varepsilon) = \inf_{x \in X} p(\varepsilon, x)$,

$\tau(k, \varepsilon) = \max \{\tau | \sum_{s=k-\tau}^{k-1} \rho_s \leqq \varepsilon, \tau \leqq k\}$,

i.e., $\tau(k, \varepsilon)$ is the largest number of steps preceding step k for which the sum of step sizes does not exceed ε.

THEOREM 6. *Assume that*

1. *X is a convex compact set in R^n and Y is a compact set in R^m,*
2. *$f(x, y)$ is a continuous function of (x, y) and a convex function of x for any $y \in Y$,*

$$\sup_{\substack{x \in X \\ y \in Y}} \|f_x(x, y)\| = K < \infty,$$

3. *Measure P is such that $\gamma(\varepsilon) > 0$ for $\varepsilon > 0$,*
4. *$\rho_s \to +0$, $\sum_{s=0}^{\infty} \rho_s = \infty$.*

Then for $s \to \infty$

$$E \min_{z \in X^*} \|x^s - z\| \to 0.$$

If, in addition, there exists an $\varepsilon_0 > 0$ such that for all $\varepsilon \leqq \varepsilon_0$ and each $0 < q < 1$

$$\sum_{s=0}^{\infty} q^{\tau(s,\varepsilon)} < \infty, \tag{34}$$

then, as $s \to \infty$,

$$\min \{\|x^s - z\| | z \in X^*\} \to 0$$

with probability 1.

Proof.

1. First of all let us prove that

$$E[F(x^s) - f(x^s, y^s)] \to 0.$$

To simplify the notation we shall assume that $N_s = N \geqq 1$. According to the algorithm

$$f(x^s, y^{s+1}) \geqq f(x^s, y^{s,v}), \qquad v = \overline{0, N}$$

and therefore

$$f(x^{s+1}, y^{s+1}) - f(x^{s+1}, y^{s,v}) \geqq [f(x^{s+1}, y^{s+1}) - f(x^s, y^{s+1})] + [f(x^s, y^{s,v}) - f(x^{s+1}, y^{s,v})].$$

According to the assumption 2 of the theorem there is a constant K such that

$$|f(x^{s+1}, y) - f(x^s, y)| \leqq K\|x^{s+1} - x^s\| \leqq K^2\rho_s,$$

hence

$$f(x^{s+1}, y^{s+1}) \geqq f(x^{s+1}, y^{s,v}) - 2K^2\rho_s.$$

We also have

$$f(x^{s+1}, y^{s+2}) \geqq f(x^{s+1}, y^{s+1,v}), \qquad v = \overline{0, N},$$

or, in particular, for $v = 0$

$$f(x^{s+1}, y^{s+2}) \geqq f(x^{s+1}, y^{s+1}).$$

Therefore

$$f(x^{s+1}, y^{s+2}) \geqq f(x^{s+1}, y^{k,v}) - 2K^2\rho_s, \qquad k = s, s+1, v = \overline{0, N},$$

and in the same way

$$f(x^{s+2}, y^{s+2}) \geqq f(x^{s+2}, y^{k,v}) - 2K^2(\rho_s + \rho_{s+1}), \qquad k = s, s+1, v = \overline{0, N}$$

etc.

Continuing this chain of inequalities, we arrive at the following conclusion:

$$f(x^s, y^s) \geqq f(x^s, y^{k,v}) - 2K^2 \sum_{l=s-\tau(s,\varepsilon)}^{s-1} \rho_l$$

$$k = \overline{s - \tau(s, \varepsilon), s-1}, \qquad v = \overline{0, N}.$$

Thus, if

$$Y_{s,\varepsilon} = \{y^{k,v}, v = \overline{0, N}, k = \overline{s - \tau(s, \varepsilon), s-1}\}$$

then

$$f(x^s, y^s) \geqq \max_{y \in Y_{s,\varepsilon}} f(x^s, y) - 2K^2\varepsilon.$$

It is easy to see from this that

$$P\{F(x^s) - f(x^s, y^s) > (1+2K^2)\varepsilon\} \leqq P\{F(x^s) - \max_{y \in Y_{s,\varepsilon}} f(x^s, y) > \varepsilon\} \leqq [1 - \gamma(\varepsilon)]^{N\tau(s,\varepsilon)}.$$

Since $\rho_s \to 0$, then $\tau(s, \varepsilon) \to \infty$ as $s \to \infty$. Hence

$$[1 - \gamma(\varepsilon)]^{N\tau(s,\varepsilon)} \to 0$$

as $s \to \infty$, and therefore

$$\tau_s = P\{F(x^s) - f(x^s, y^s) > \varepsilon\} \to 0$$

for arbitrary $\varepsilon > 0$. Compactness of X and Y together with the continuity of $f(x, y)$ gives now $\max_{x \in X, y \in Y} |f(x, y)| < C < \infty$ which implies

$$E[F(x^s) - f(x^s, y^s)] \leqq \varepsilon + C\tau_s \to \varepsilon$$

for arbitrary $\varepsilon > 0$. This gives finally

$$E[F(x^s) - f(x^s, y^s)] \to 0.$$

2. We shall now show that, under assumption (34), $F(x^s) - f(x^s, y^s) \to 0$ with probability 1. It is sufficient to verify that

$$P\{\sup_{k \geqq s} [F(x^k) - f(x^k, y^k)] > (1 + 2K^2)\varepsilon\} \to 0.$$

We have

$$P\{\sup_{k \geqq s} [F(x^k) - f(x^k, y^k)] > (1 + 2K^2)\varepsilon\}$$

$$\leqq P\{\sup_{k \geqq s} [F(x^k) - \max_{y \in Y_{k,\varepsilon}} f(x^k, y)] > \varepsilon\}$$

$$\leqq \sum_{k=s}^{\infty} P\{F(x^k) - \max_{y \in Y_{k,\varepsilon}} f(x^k, y) > \varepsilon\} \leqq \sum_{k=s}^{\infty} [1 - \gamma(\varepsilon)]^{N\tau(k,\varepsilon)} \to 0,$$

since from assumption (34) the series

$$\sum_{k=s}^{\infty} [1 - \gamma(\varepsilon)]^{N\tau(k,\varepsilon)} \to 0$$

as $s \to \infty$.

3. Let us now prove that $Ew(x^s) \to 0$ as $s \to \infty$, where

$$w(x) = \min_{z \in X^*} \|x - z\|^2.$$

We have

$$w(x^{s+1}) = \|x^{s+1} - x_s^*\|^2 \leqq w(x^s) - 2\rho_s \langle f_x(x^s, y^s), x^s - x_s^* \rangle + \rho_s^2 \|f_x(x^s, y^s)\|^2$$

$$\leqq w(x^s) - 2\rho_s [f(x^s, y^s) - f(x_s^*, y^s)] + K^2 \rho_s^2$$

$$\leqq w(x^s) - 2\rho_s [f(x^s, y^s) - \min_{x \in X} F(x)] + K^2 \rho_s^2$$

$$\leqq w(x^s) - 2\rho_s [F(x^s) - \min_{x \in X} F(x)] + 2\rho_s [F(x^s) - f(x^s, y^s)] + K^2 \rho_s^2.$$

Taking the mathematical expectation of both sides of this inequality leads to

$$Ew(x^{s+1}) \leqq Ew(x^s) - 2\rho_s E[F(x^s) - \min_{x \in X} F(x)] + 2\rho_s \beta_s + K^2 \rho_s^2, \tag{35}$$

where $\beta_s \to 0$ as $s \to \infty$ since it has already been proved that

$$E[F(x^s) - f(x^s, y^s)] \to 0 \quad \text{for } s \to \infty.$$

Now let us suppose, contrary to our original assumption, that

$$Ew(x^s) > \alpha > 0, \qquad s \geqq s_0.$$

It is easy to see that because of boundedness and continuity we also have

$$E[F(x^s) - \min_{x \in X} F(x)] > \delta > 0,$$

where $\delta = \delta(\alpha)$ is a constant which depends on α. Then for sufficiently large $s \geqq s_1$

$$Ew(x^{s+1}) \leqq Ew(x^s) - 2\rho_s [\delta - 2\beta_s - K^2 \rho_s] \leqq Ew(x^s) - \delta\rho_s \tag{36}$$

since $\rho_s \to 0$, $\beta_s \to 0$ and therefore we can suppose that

$$\delta - 2\beta_s - K^2\rho_s > \delta/2, \qquad s \geqq s_1.$$

Summing the inequality (36) from s_1 to k, $k \to \infty$, we obtain from assumption (4) a contradiction to the nonnegativeness of $Ew(x^s)$. Hence, a subsequence $\{x^{s_k}\}$ exists such that

$$Ew(x^{s_k}) \to 0$$

as $k \to \infty$. Therefore for a given $\alpha > 0$ a number $k(\alpha)$ exists such that

$$Ew(x^{s_k}) < \alpha,$$

where $s_k > s_{k(\alpha)}$. Let r be such that $s_k \leqq r \leqq s_{k+1}$ and $Ew(x^r) > \alpha$. Take l such that

$$l = \min_{s_k < i \leqq r} \{i: Ew(x^j) > \alpha \quad \text{for } i \leqq j \leqq r\}.$$

Since $\rho_s \to 0$ and $\beta_s \to 0$, we may assume that $2\beta_s + K^2\rho_s < \delta(\alpha)$ for $s > s_{k(\alpha)}$. This and (36) together imply that $Ew(x^r) \leqq Ew(x^l)$. Now from (35) and the definition of l we get

$$Ew(x^l) \leqq Ew(x^{l-1}) + 2\rho_l\beta_l + K^2\rho_l^2 \leqq \alpha + 2\rho_l\beta_l + K^2\rho_l^2.$$

Thus $Ew(x^s) \to 0$, because α was chosen arbitrarily and $\rho_l \to 0$.

4. It can be proved that $w(x^s)$ converges to 0 with probability 1 in the same way that we have already proved mean convergence. We have the inequality

$$w(x^{s+1}) \leqq w(x^s) - 2\rho_s[F(x^s) - \min_{x \in X} F(x)] + 2\rho_s\gamma_s + K^2\rho_s^2,$$

where $\gamma_s \to 0$ with probability 1 because it has already been shown that under assumption (34)

$$F(x^s) - f(x^s, y^s) \to 0 \quad \text{as } s \to \infty$$

with probability 1. If we now assume that

$$w(x^s) > \alpha, \qquad s \geqq s_0$$

for some element of probabilistic space we will also have

$$F(x^s) - \min_{x \in X} F(x) > \delta > 0$$

etc.

We shall now give a special case in which condition (34) is satisfied.

Example. Assume that $\rho_s = 1/s^b$, $a > 0$, $0 < b \leqq 1$. Then Raab's test for series convergence shows that condition (34) is satisfied.

REFERENCES

[1] V. P. BULATOV, *Imbedding Methods for Optimization Problems*, Nauka, Novosibirsk, 1977. (In Russian.)

[2] G. B. DANTZIG, *Linear Programming and Extensions*, Princeton Univ. Press, Princeton, NJ, 1963.

[3] V. F. DEMIANOV, *Nondifferentiable Optimization*, Nauka, Moscow, 1982. (In Russian.)

[4] J. DUPACOVA, *Minimax approach to stochastic linear programming and the moment problem*, Economicko-matematicky obzor, 13 (1977), pp. 279-307. (In Czech., extended abstract in English in Z. Angew. Math. Mech., 58 (1978), pp. T466-T467.)

[5] ———, *Minimax stochastic programs with nonseparable penalties*, Lecture Notes in Control and Information Sciences 22, Springer-Verlag, Berlin, 1980, pp. 157-163.

[6] B. C. EAVES AND W. ZANGWILL, *Generalized cutting plane algorithms*, this Journal, 9 (1971), pp. 529-542.

[7] YU. ERMOLIEV, *Method for stochastic programming in randomized strategies*, Kibernetica, 1 (1970), pp. 3-9. (In Russian.)
[8] ———, *Methods of Stochastic Programming*, Nauka, Moscow, 1976. (In Russian.)
[9] YU. ERMOLIEV AND A. GAIVORONSKI, *Simultaneous nonstationary optimization, estimation and approximation procedures*, Collaborative paper CP-82-16, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1982.
[10] YU. ERMOLIEV AND C. NEDEVA, *Stochastic optimization problems with partially known distribution functions*, Collaborative Paper CP-82-60, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1982.
[11] YU. ERMOLIEV AND A. GAIVORONSKY, *A stochastic algorithm for minimax problems*, Collaborative paper CP-82-88, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1982.
[12] S. FROMOVITZ, *Nonlinear programming with randomization*, Management Sci., 9 (1965), pp. 831-846.
[13] A. N. GOLODNIKOV, *Finding optimal distribution functions in stochastic programming problems*, Dissertation abstract, Institute of Cybernetics Press, Kiev, 1979. (In Russian.)
[14] A. I. KAPLINSKI AND A. I. PROPOI, *Stochastic approach to nonlinear programming*, Automatika i Telemekhanika, 3 (1970), pp. 49-55. (In Russian.)
[15] J. H. B. KEMPERMAN, *The general moment problem: A geometric approach*, Ann. Math. Statist., 39 (1968), pp. 93-122.
[16] ———, *On a class of moment problems*, Proc. Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 2, Univ. California Press, Berkeley and Los Angeles, 1972, pp. 101-126.
[17] M. KREIN AND A. NUDELMAN, *The Markov Moment Problem and Extremal Problems*, Trans. Math. Monographs 50, American Mathematical Society, Providence, RI, 1977.
[18] KY FAN, *Convex sets and their applications*, Argonne National Laboratory, Argonne, IL, 1959.
[19] L. NAZARETH AND R. WETS, *Algorithms for stochastic programs: The case of nonstochastic tenders*, Working Paper WP-83-5, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1983.
[20] E. NURMINSKI AND A. ZELIKHOVSKI, *ε-quasigradient method for solving nonsmooth extremal problems*, Kibernetika, 1 (1977), pp. 109-113. (In Russian.)
[21] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton Univ. Press, Princeton, NJ, 1970.
[22] D. TOPKIS, *Cutting plane methods without nested constraint sets*, Oper. Res. 18 (1970), pp. 404-413.
[23] R. WETS, *Grundlagen Konvexer Optimierung*, Lecture Notes in Economics and Mathematical Systems 137, Springer-Verlag, Berlin, 1976.
[24] J. ŽAČKOVA, *On minimax solutions of stochastic linear programming problems*, Čas. Pěst. Mat., 91 (1966), pp. 423-430.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

005

ON THE PHASE PORTRAIT OF THE MATRIX RICCATI EQUATION ARISING FROM THE PERIODIC CONTROL PROBLEM*

MARK A. SHAYMAN†

Abstract. A comprehensive description is provided of the properties of the matrix Riccati differential equation in which the coefficient matrices are T-periodic for some $T>0$. The main results include: (1) a classification of all the periodic equilibria (T-periodic solutions); (2) necessary and sufficient conditions for convergence to the uniformly asymptotically stable (completely unstable) periodic equilibrium as $t\to\infty$ ($t\to-\infty$); (3) necessary and sufficient conditions for a solution to escape in finite forward or backward time; (4) a description of every almost periodic solution; (5) a proof that every solution which does not escape in finite time is asymptotically almost periodic and an explicit formula for the limiting almost periodic solution.

Key words. Riccati differential equation, phase portrait, periodic system, linear quadratic optimal control, Lagrange-Grassmann manifold

1. Introduction. By the Riccati differential equation (RDE) we mean the time-varying quadratic differential equation

$$\dot{K}(t)=-A'(t)K(t)-K(t)A(t)+K(t)B(t)B'(t)K(t)-C'(t)C(t)$$

defined on the vector space $S(n)$ of real symmetric $n\times n$ matrices. $A(t)$, $B(t)$, and $C(t)$ are real matrices of dimensions $n\times n$, $n\times m$, and $p\times n$ respectively. By the periodic Riccati differential equation (PRDE) we mean the special case of the RDE in which $A(t)$, $B(t)$, and $C(t)$ are each periodic with period $T>0$ and are integrable over $[0, T]$. Due to its central role in the solution of the least squares control problem for a periodic linear system, the PRDE has been the subject of several investigations in recent years, including [16], [18], [7], [12], [1], [4], [13], [17], [3], [19]. However, the theory of the PRDE has remained underdeveloped in comparison to the theory of the time-invariant RDE.

The purpose of this paper is to generalize to the PRDE a number of key results in the theory of the time-invariant RDE. We briefly describe the organization of this paper. In § 2, we describe how the RDE is extended to a differential equation on the so-called Lagrange-Grassmann manifold $\mathscr{L}(n)$, which may be viewed as a compactification of the space of $n\times n$ symmetric matrices. This viewpoint is extremely helpful for the proofs of many of our results. In § 3, we classify all of the periodic equilibria of the PRDE by generalizing a well-known result of J. C. Willems [24] which classifies the equilibria for the time-invariant RDE. The signatures of the periodic equilibria are described provided $(C(\cdot), A(\cdot))$ is observable. In § 4, the stability properties of the periodic equilibria are determined. In particular, we give necessary and sufficient conditions for convergence to the uniformly asymptotically stable periodic equilibrium as $t\to\infty$ and for convergence to the completely unstable periodic equilibrium as $t\to-\infty$. Section 5 gives necessary and sufficient conditions for a solution to have a finite escape in either forward or backward time. In § 6, we determine all of the almost periodic solutions of the PRDE. These solutions are contained in oscillating manifolds, each of which is isomorphic to a product of Grassmann manifolds. In § 7, we prove that under mild assumptions, every solution either escapes in finite forward (backward)

* Received by the editors February 14, 1984. This research was partially supported by the National Science Foundation under grant ECS-8301015. An abridged version of §§ 1-3 of this paper was presented at the Allerton Conference on Communication, Control, and Computing, Monticello, Illinois, October 1983.
† Department of Systems Science and Mathematics, Washington University, St. Louis, Missouri 63130.

time or converges to an almost periodic solution as $t \to \infty$ ($t \to -\infty$). If a given solution has no finite escape time, we give an explicit formula for the limiting almost periodic solution. Combined with the necessary and sufficient conditions for finite escape times given in § 5, these results furnish a description of the asymptotic behavior of *every* solution of the PRDE, and hence constitute a "complete global phase portrait". The results in §§ 6 and 7 extend to the PRDE recent results of Shayman [22], [23], for the time-invariant RDE.

Also of interest is the PRDE which arises from a periodic control problem in which there are conflicting objectives. In this case the nonnegative definite matrix $C'(t)C(t)$ is replaced by an arbitrary symmetric T-periodic matrix $Q(t)$. All of our results, except for those concerning the signature of solutions (which require an observability assumption) may be extended to this more general situation simply by adding to the hypothesis of each theorem the assumption that the set of periodic equilibria is nonempty. The justification for this claim is that the proofs of our theorems depend on $Q(t)$ being of the form $C'(t)C(t)$ only so that the existence of the periodic equilibrium $K^+(t)$ can be concluded from the known result Theorem 2. (The proof of Theorem 2 in the literature [13] depends on $Q(t)$ being of the form $C'(t)C(t)$.) If it is assumed that there exists a periodic equilibrium $K_e(t)$, the change of variables $\tilde{K}(t) \equiv K(t) - K_e(t)$ yields a PRDE in which $Q(t) \equiv 0$ (which is trivially of the form $C'(t)C(t)$). Then the existence of $K^+(t)$ for the original PRDE is established by applying Theorem 2 to the transformed PRDE and then transforming back.

The following notation will be used: If X is any matrix, Sp X denotes the subspace spanned by the columns of X. If X is a square matrix, Spec (X) denotes the spectrum of X. Also $L^+(X)$ ($L^-(X)$) denotes the invariant subspace associated with the eigenvalues of X inside (outside) the unit circle. If S is any subspace of $\mathbb{R}^{2n}$, $S^\perp$ denotes the orthogonal complement of S using the standard inner product on $\mathbb{R}^{2n}$. $\psi_A(t, t_0)$ is the transition matrix for the linear differential equation $\dot{x}(t) = A(t)x(t)$.

2. Background. Associated with the RDE is the $2n \times 2n$ Hamiltonian matrix

$$H(t) = \begin{bmatrix} A(t) & -B(t)B'(t) \\ -C'(t)C(t) & -A'(t) \end{bmatrix}.$$

Let $\Phi(t, t_0)$ denote the transition matrix corresponding to $H(t)$. Partition $\Phi(t, t_0)$ into $n \times n$ blocks as

$$\Phi(t, t_0) = \begin{bmatrix} \phi_{11}(t, t_0) & \phi_{12}(t, t_0) \\ \phi_{21}(t, t_0) & \phi_{22}(t, t_0) \end{bmatrix}.$$

Let $K(t, K_0, t_0)$ denote the solution of the RDE which goes through K_0 at time t_0. It is well known [5, p. 156] that

$$K(t, K_0, t_0) = [\phi_{21}(t, t_0) + \phi_{22}(t, t_0)K_0][\phi_{11}(t, t_0) + \phi_{12}(t, t_0)K_0]^{-1}. \tag{1}$$

This formula is easily verified by differentiation. It is valid as long as the inverse exists. The matrix $\phi_{11}(t, t_0) + \phi_{12}(t, t_0)K_0$ becoming singular at $t = t_1$ is equivalent to the solution $K(t, K_0, t_0)$ having a finite escape time at $t = t_1$.

In order to study the RDE, it is useful to compactify the phase space. The natural compactification of the phase space $S(n)$ for the RDE has been described by R. Hermann and C. Martin [11], [15], and is the so-called Lagrange-Grassmann manifold $\mathscr{L}(n)$. It is defined as follows: Let J denote the $2n \times 2n$ matrix $\left[\begin{smallmatrix} 0 & I \\ -I & 0 \end{smallmatrix}\right]$ and define a skew-symmetric bilinear form ω on $\mathbb{R}^{2n}$ by $\omega(x, y) = x'Jy$. Let $G^n(\mathbb{R}^{2n})$ denote the Grassmann manifold of all n-dimensional subspaces of $\mathbb{R}^{2n}$. Then $\mathscr{L}(n) =$

$\{S \in G^n(\mathbb{R}^{2n})\colon \omega(x, y) = 0, \forall x, y \in S\}$. Thus, $\mathscr{L}(n)$ consists of those n-dimensional subspaces of $\mathbb{R}^{2n}$ on which ω vanishes identically.

It is natural to view $\mathscr{L}(n)$ as a compactification of $S(n)$. To see this, we note that the n-dimensional subspace $\operatorname{Sp}\begin{bmatrix} I \\ K \end{bmatrix}$ (the column space of the $2n \times n$ matrix $\begin{bmatrix} I \\ K \end{bmatrix}$) belongs to $\mathscr{L}(n)$ if and only if the $n \times n$ matrix K is symmetric. Thus, we can define a mapping $\gamma\colon S(n) \to \mathscr{L}(n)$ by $\gamma(K) = \operatorname{Sp}\begin{bmatrix} I \\ K \end{bmatrix}$. Let $\mathscr{L}_0(n)$ consist of those subspaces in $\mathscr{L}(n)$ which are complementary to the n-dimensional subspace $\operatorname{Sp}\begin{bmatrix} 0 \\ I \end{bmatrix}$. In other words, $\mathscr{L}_0(n) = \{S \in \mathscr{L}(n)\colon S \cap \operatorname{Sp}\begin{bmatrix} 0 \\ I \end{bmatrix} = 0\}$. It is easy to show that γ is an embedding of $S(n)$ into $\mathscr{L}(n)$ with image $\mathscr{L}_0(n)$. Since $\mathscr{L}(n)$ is compact and $\mathscr{L}_0(n)$ is an open and dense subset of $\mathscr{L}(n)$, we can use the embedding γ to identify $S(n)$ with $\mathscr{L}_0(n)$ and thereby regard $\mathscr{L}(n)$ as a compactification of $S(n)$.

Define a time-varying flow on $\mathscr{L}(n)$ by $S(t, S_0, t_0) = \Phi(t, t_0)(S_0)$, the image of the subspace S_0 under the nonsingular linear mapping $\Phi(t, t_0)$. Note that since $H(t)$ is a Hamiltonian matrix $(JH(t) + H'(t)J = 0)$, it follows that $\Phi(t, t_0)$ is symplectic $(\Phi'(t, t_0)J\Phi(t, t_0) = J)$. This implies that if $S_0 \in \mathscr{L}(n)$, then $\Phi(t, t_0)(S_0) \in \mathscr{L}(n)$, so $S(t, S_0, t_0)$ is indeed a flow on $\mathscr{L}(n)$. If $K_0 \in S(n)$, then

$$S(t, \gamma(K_0), t_0) = \operatorname{Sp} \Phi(t, t_0)\begin{bmatrix} I \\ K_0 \end{bmatrix} = \operatorname{Sp}\begin{bmatrix} \phi_{11}(t, t_0) + \phi_{12}(t, t_0)K_0 \\ \phi_{21}(t, t_0) + \phi_{22}(t, t_0)K_0 \end{bmatrix}.$$

From (1), we have

$$\gamma(K(t, K_0, t_0)) = \operatorname{Sp}\begin{bmatrix} I \\ [\phi_{21}(t, t_0) + \phi_{22}(t, t_0)K_0][\phi_{11}(t, t_0) + \phi_{12}(t, t_0)K_0]^{-1} \end{bmatrix}$$

$$= \operatorname{Sp}\begin{bmatrix} \phi_{11}(t, t_0) + \phi_{12}(t, t_0)K_0 \\ \phi_{21}(t, t_0) + \phi_{22}(t, t_0)K_0 \end{bmatrix}$$

which is valid as long as the solution $K(t, K_0, t_0)$ continues to exist. Thus, the flow of the RDE on $S(n)$ is related to the flow which was defined on $\mathscr{L}(n)$ by the equation

$$\gamma(K(t, K_0, t_0)) = S(t, \gamma(K_0), t_0) \tag{2}$$

as long as $K(t, K_0, t_0)$ exists, or equivalently, as long as $S(t, \gamma(K_0), t_0)$ remains in the subset $\mathscr{L}_0(n)$ of $\mathscr{L}(n)$. In other words, (2) says that if we use the embedding γ to identify $S(n)$ with $\mathscr{L}_0(n)$, then the restriction to $\mathscr{L}_0(n)$ of the flow $S(t, S_0, t_0)$ is identified with the flow of the RDE. Thus, the embedding γ permits us to view $\mathscr{L}(n)$ as $S(n)$ together with some points added at infinity to obtain a compact space, and to view $S(t, S_0, t_0)$ as the extension of the flow of the RDE to the enlarged space. We will refer to the PRDE extended to $\mathscr{L}(n)$ as the *extended periodic Riccati differential equation* (EPRDE).

For our purposes, there are two main advantages to regarding the RDE as a differential equation on $\mathscr{L}(n)$. The first advantage is that it permits us to explicitly consider solutions of the differential equation "at infinity." The second advantage is that the flow on $\mathscr{L}(n)$ is particularly simple. It is given by the natural action of nonsingular linear transformations on subspaces.

3. Periodic equilibria for the periodic Riccati differential equation. Throughout this section, it is assumed that the matrices $A(\cdot)$, $B(\cdot)$, and $C(\cdot)$ in the RDE are each periodic with period $T > 0$, and are integrable on $[0, T]$. The first result relates periodic solutions to invariant subspaces of $\Phi(t_0 + T, t_0)$. By a periodic equilibrium, we mean a T-periodic solution of the PRDE.

LEMMA 1. (a) *If* $K(t, K_0, t_0)$ *is a periodic equilibrium of the* PRDE, *then* $\gamma(K_0)$ *is* $\Phi(t_0+T, t_0)$*-invariant and* $S(t, \gamma(K_0), t_0) \in \mathscr{L}_0(n), \forall t$. (b) *If* S_0 *is* $\Phi(t_0+T, t_0)$*-invariant and* $S(t, S_0, t_0) \in \mathscr{L}_0(n), \forall t$, *then* $K(t, \gamma^{-1}(S_0), t_0)$ $(= \gamma^{-1}(S(t, S_0, t_0)))$ *is a periodic equilibrium of the* PRDE.

Proof. (a) Suppose that $K(t, K_0, t_0)$ is a periodic equilibrium of the PRDE. Then $K(t, K_0, t_0)$ exists for all t, which implies (see (2)) that $S(t, \gamma(K_0), t_0) \in \mathscr{L}_0(n), \forall t$. Since $K(t_0+T, K_0, t_0) = K_0$, (2) implies that $\gamma(K_0) = \gamma(K(t_0+T, K_0, t_0)) = S(t_0+T, \gamma(K_0), t_0) = \Phi(t_0+T, t_0)(\gamma(K_0))$, so $\gamma(K_0)$ is $\Phi(t_0+T, t_0)$-invariant. (b) Suppose that S_0 is $\Phi(t_0+T, t_0)$-invariant and $S(t, S_0, t_0) \in \mathscr{L}_0(n), \forall t$. Then (2) gives the equation

$$\gamma(K(t, \gamma^{-1}(S_0), t_0)) = S(t, S_0, t_0)$$

which is valid for all t since $S(t, S_0, t_0)$ remains in $\mathscr{L}_0(n)$. ($\gamma^{-1}(S_0)$ is defined since $S_0 \in \mathscr{L}_0(n)$.) Hence, $K(t, \gamma^{-1}(S_0), t_0)$ exists for all t and

$$K(t, \gamma^{-1}(S_0), t_0) = \gamma^{-1}(S(t, S_0, t_0)).$$

Thus,

$$\begin{aligned} K(t+T, \gamma^{-1}(S_0), t_0) &= \gamma^{-1}(S(t+T, S_0, t_0)) = \gamma^{-1}(\Phi(t+T, t_0)(S_0)) \\ &= \gamma^{-1}(\Phi(t+T, t_0+T)\Phi(t_0+T, t_0)(S_0)) = \gamma^{-1}(\Phi(t, t_0)(S_0)) \\ &= \gamma^{-1}(S(t, S_0, t_0)) = K(t, \gamma^{-1}(S_0), t_0). \end{aligned}$$

Thus, $K(t, \gamma^{-1}(S_0), t_0)$ is a periodic equilibrium. □

If $K(t)$ is a periodic solution of the PRDE, then $\gamma(K(t_0))$ is a $\Phi(t_0+T, t_0)$-invariant subspace. The next result describes the relationship between the restriction $\Phi(t_0+T, t_0)|\gamma(K(t_0))$ and the transition matrix $\psi_{A-BB'K}(t, \tau)$ for the periodic closed-loop system $A(t) - B(t)B'(t)K(t)$.

LEMMA 2. *Suppose that* $K(t)$ *is a periodic equilibrium of the* PRDE. *Then* $\psi_{A-BB'K}(t_0+T, t_0)$ *is the matrix for* $\Phi(t_0+T, t_0)|\gamma(K(t_0))$ *with respect to the basis consisting of the columns of* $\left[\begin{smallmatrix} I \\ K(t_0) \end{smallmatrix}\right]$.

Proof. Let

$$\begin{bmatrix} X(t, t_0) \\ Y(t, t_0) \end{bmatrix} = \Phi(t, t_0) \begin{bmatrix} I \\ K(t_0) \end{bmatrix}.$$

Then $X(t_0, t_0) = I$. Also

$$\frac{d}{dt}\begin{bmatrix} X(t, t_0) \\ Y(t, t_0) \end{bmatrix} = \begin{bmatrix} A(t) & -B(t)B'(t) \\ -C'(t)C(t) & -A'(t) \end{bmatrix} \begin{bmatrix} X(t, t_0) \\ Y(t, t_0) \end{bmatrix}.$$

Since $K(t)$ exists for all t, $X(t, t_0)$ is never singular, and $K(t) = Y(t, t_0)X^{-1}(t, t_0)$. Hence,

$$(d/dt)X(t, t_0) = A(t)X(t, t_0) - B(t)B'(t)Y(t, t_0) = (A(t) - B(t)B'(t)K(t))X(t, t_0),$$

which shows that $X(t, t_0) = \psi_{A-BB'K}(t, t_0)$.

Let R denote the matrix for $\Phi(t_0+T, t_0)|\gamma(K(t_0))$ with respect to the basis $\left[\begin{smallmatrix} I \\ K(t_0) \end{smallmatrix}\right]$. Then

$$\begin{bmatrix} X(t_0+T, t_0) \\ Y(t_0+T, t_0) \end{bmatrix} = \Phi(t_0+T, t_0) \begin{bmatrix} I \\ K(t_0) \end{bmatrix} = \begin{bmatrix} I \\ K(t_0) \end{bmatrix} R,$$

which shows that $X(t_0+T, t_0) = R$. Hence $R = \psi_{A-BB'K}(t_0+T, t_0)$ as asserted. □

COROLLARY. *Suppose that $K(t)$ is a periodic equilibrium of the* PRDE. *Then the characteristic multipliers of the periodic closed-loop system $A(t)-B(t)B'(t)K(t)$ are the same as the eigenvalues of $\Phi(t_0+T, t_0)|\gamma(K(t_0))$.*

Lemma 1 shows that if S_0 is a $\Phi(t_0+T, t_0)$-invariant subspace which belongs to $\mathscr{L}(n)$, then $S(t, S_0, t_0)$ corresponds to a periodic equilibrium $K(t)=\gamma^{-1}(S(t, S_0, t_0))$ provided that $S(t, S_0, t_0)\in\mathscr{L}_0(n)$ for all t. Thus, the existence of periodic equilibria for the PRDE is related to the existence of $\Phi(t_0+T, t_0)$-invariant subspaces which belong to $\mathscr{L}(n)$. The next lemma will be useful in establishing the existence of such subspaces.

LEMMA 3. *Let M be a symplectic matrix. Suppose that S_1 and S_2 are each sums of primary components of M such that if $\lambda\in\operatorname{Spec}(M|S_1)$, then $\lambda^{-1}\notin\operatorname{Spec}(M|S_2)$. Then $x'Jy=0$, $\forall x\in S_1$, $\forall y\in S_2$.*

Proof. Let X_1 and X_2 be basis matrices for S_1 and S_2 respectively. Since S_1 and S_2 are M-invariant, there exist square matrices R_1, R_2 (of the appropriate dimensions) such that $MX_1=X_1R_1$, $MX_2=X_2R_2$. Since M is symplectic, we have $M'JM=J$. Thus, $X_2'JX_1=X_2'M'JMX_1=R_2'X_2'JX_1R_1$. Let $Z=X_2'JX_1$. Then $Z=R_2'ZR_1$ or $R_2'Z=ZR_1^{-1}$. The hypotheses imply that R_1^{-1} and R_2' have no eigenvalues in common. Hence, the only solution is $Z=0$. □

As a preliminary to obtaining further results regarding the periodic equilibria of the PRDE, we consider a special case. By the *stable homogeneous periodic Riccati differential equation* (SHPRDE), we mean the special case of the PRDE where $C(t)\equiv 0$ and all of the characteristic multipliers of $A(\cdot)$ are inside the unit circle. We use $\hat{H}(t)$ to denote the Hamiltonian matrix

$$\begin{bmatrix} A(t) & -B(t)B'(t) \\ 0 & -A'(t) \end{bmatrix}$$

for the SHPRDE, and we use $\hat{\Phi}(t, t_0)$ to denote the transition matrix corresponding to $\hat{H}(t)$. Let $W_A(t_0, t)$ denote the controllability Gramian for the pair $(A(\cdot), B(\cdot))$. In other words,

$$W_A(t_0, t)=\int_{t_0}^{t}\psi_A(t_0, \sigma)B(\sigma)B'(\sigma)\psi_A'(t_0, \sigma)\,d\sigma.$$

It may be verified directly that

$$\hat{\Phi}(t, t_0)=\begin{bmatrix} \psi_A(t, t_0) & -\psi_A(t, t_0)W_A(t_0, t) \\ 0 & \psi_A'(t_0, t) \end{bmatrix}. \tag{3}$$

We will need the following standard result concerning the discrete version of the (algebraic) Lyapunov equation.

LEMMA 4. *Suppose that the symmetric matrix Z satisfies the equation*

$$Z-R'ZR=-L$$

where $L\geqq 0$ and $|\lambda|<1$, $\forall\lambda\in\operatorname{Spec}(R)$. Then $Z\leqq 0$.

LEMMA 5. *Suppose that $S\in\mathscr{L}(n)$ and that S is $\hat{\Phi}(t_0+T, t_0)$-invariant. Let $\begin{bmatrix} X \\ Y \end{bmatrix}$ be any basis for S. Then $X'Y\leqq 0$.*

Proof. Note that since $S\in\mathscr{L}(n)$, $X'Y$ is symmetric. We begin by considering the special case where $S=L^-(\hat{\Phi}(t_0+T, t_0))$. In other words, S is the subspace corresponding to the eigenvalues of $\hat{\Phi}(t_0+T, t_0)$ which are *outside* the unit circle. It is clear from (3) that $\hat{\Phi}(t_0+T, t_0)$ has no eigenvalues on the unit circle, so S is an n-dimensional $\hat{\Phi}(t_0+T, t_0)$-invariant subspace. Applying Lemma 3 with $S_1=S_2=S$, we conclude that $S\in\mathscr{L}(n)$.

Let $\begin{bmatrix}\hat{X}\\ \hat{Y}\end{bmatrix}$ be a basis for S. Let $N=\psi_A(t_0+T, t_0)$ and let $W=W_A(t_0, t_0+T)$. From (3), we have

$$\hat{\Phi}(t_0+T, t_0)=\begin{bmatrix}N & -NW\\ 0 & (N^{-1})'\end{bmatrix}.$$

Since S is $\hat{\Phi}(t_0+T, t_0)$-invariant, there exists an $n\times n$ matrix R with $|\lambda|>1$, $\forall\lambda\in$ Spec (R) such that

$$\begin{bmatrix}N & -NW\\ 0 & (N^{-1})'\end{bmatrix}\begin{bmatrix}\hat{X}\\ \hat{Y}\end{bmatrix}=\begin{bmatrix}\hat{X}\\ \hat{Y}\end{bmatrix}R.$$

Thus,

$$N\hat{X}-NW\hat{Y}=\hat{X}R, \tag{4}$$

$$(N^{-1})'\hat{Y}=\hat{Y}R. \tag{5}$$

Premultiplying (4) by $\hat{Y}'$ gives

$$\hat{Y}'N\hat{X}-\hat{Y}'NW\hat{Y}=\hat{Y}'\hat{X}R. \tag{6}$$

Using (5) to substitute for $\hat{Y}'N$ in (6) gives

$$(R^{-1})'\hat{Y}'\hat{X}-(R^{-1})'\hat{Y}'W\hat{Y}=\hat{Y}'\hat{X}R. \tag{7}$$

Letting $P=\hat{Y}'\hat{X}$ gives

$$(R^{-1})'P-(R^{-1})'\hat{Y}'W\hat{Y}=PR,$$

or

$$P-(R^{-1})'PR^{-1}=-(R^{-1})'\hat{Y}'W\hat{Y}R^{-1}.$$

Since $W\geqq 0$, we conclude from Lemma 4 that $P\leqq 0$. Thus, $\hat{X}'\hat{Y}=\hat{Y}'\hat{X}\leqq 0$.

We continue to use $\begin{bmatrix}\hat{X}\\ \hat{Y}\end{bmatrix}$ to denote a basis for $L^-(\hat{\Phi}(t_0+T, t_0))$, but now we let S denote an arbitrary subspace in $\mathscr{L}(n)$ which is $\hat{\Phi}(t_0+T, t_0)$-invariant. Then

$$S=[S\cap L^+(\hat{\Phi}(t_0+T, t_0))]\oplus[S\cap L^-(\hat{\Phi}(t_0+T, t_0))]$$

$$=\left[S\cap \operatorname{Sp}\begin{bmatrix}I\\ 0\end{bmatrix}\right]\oplus\left[S\cap \operatorname{Sp}\begin{bmatrix}\hat{X}\\ \hat{Y}\end{bmatrix}\right].$$

Let

$$l=\dim S\cap \operatorname{Sp}\begin{bmatrix}I\\ 0\end{bmatrix}.$$

There exist $n\times l$ and $n\times(n-l)$ full rank matrices C, D such that

$$S\cap \operatorname{Sp}\begin{bmatrix}I\\ 0\end{bmatrix}=\operatorname{Sp}\begin{bmatrix}I\\ 0\end{bmatrix}C \quad\text{and}\quad S\cap \operatorname{Sp}\begin{bmatrix}\hat{X}\\ \hat{Y}\end{bmatrix}=\operatorname{Sp}\begin{bmatrix}\hat{X}\\ \hat{Y}\end{bmatrix}D.$$

So

$$S=\operatorname{Sp}\begin{bmatrix}C & \hat{X}D\\ 0 & \hat{Y}D\end{bmatrix}.$$

Let $X=[C\ \hat{X}D]$ and let $Y=[0\ \hat{Y}D]$. Since $S\in\mathscr{L}(n)$, $X'Y$ must be symmetric, which

implies that $C'\hat{Y}D = 0$. Then we obtain

$$X'Y = \begin{bmatrix} 0 & 0 \\ 0 & D'\hat{X}'\hat{Y}D \end{bmatrix}.$$

Since $\hat{X}'\hat{Y} \leqq 0$, we conclude that $X'Y \leqq 0$. □

COROLLARY. *Let $K(t)$ be a periodic equilibrium of the* SHPRDE. *Then $K(t) \leqq 0$.*

Proof. Given any t, we have $K(t) = K(t+T)$. Thus, $\mathrm{Sp}\left[\begin{smallmatrix} I \\ K(t) \end{smallmatrix}\right]$ is $\hat{\Phi}(t+T, t)$-invariant. By Lemma 5, $K(t) \leqq 0$. □

LEMMA 6. *Suppose that $(A(\cdot), B(\cdot))$ is controllable. Then every subspace belonging to $\mathscr{L}(n)$ which is $\hat{\Phi}(t_0+T, t_0)$-invariant belongs to $\mathscr{L}_0(n)$.*

Proof. Let $S \in \mathscr{L}(n)$ and suppose that S is $\hat{\Phi}(t_0+T, t_0)$-invariant. Let $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$ be a basis matrix for S. Since $\hat{\Phi}(t_0-nT, t_0) = [\hat{\Phi}(t_0+T, t_0)]^{-n}$, S is $\hat{\Phi}(t_0-nT, t_0)$-invariant. Hence, there exists an $n \times n$ matrix R such that

$$\hat{\Phi}(t_0-nT, t_0)\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} R.$$

From (3) we have

$$\psi_A(t_0-nT, t_0)X - \psi_A(t_0-nT, t_0)\, W_A(t_0, t_0-nT)\, Y = XR, \tag{8}$$

$$\psi_A'(t_0, t_0-nT)\, Y = YR. \tag{9}$$

Using (9) to substitute for Y in (8) gives

$$\psi_A(t_0-nT, t_0)X - \psi_A(t_0-nT, t_0)\, W_A(t_0, t_0-nT)\psi_A'(t_0-nT, t_0)\, YR = XR. \tag{10}$$

A straightforward calculation shows that

$$W_A(t_0, t_0-nT) = -\psi_A(t_0, t_0-nT)\, W_A(t_0-nT, t_0)\psi_A'(t_0, t_0-nT).$$

Substituting in (10) yields

$$\psi_A(t_0-nT, t_0)X + W_A(t_0-nT, t_0)\, YR = XR. \tag{11}$$

Suppose that there exists $v \neq 0$ with $Xv = 0$. Premultiplying (11) by $v'R'Y'$ and postmultiplying by v gives

$$v'R'Y'W_A(t_0-nT, t_0)\, YRv = v'R'Y'XRv. \tag{12}$$

Since $(A(\cdot), B(\cdot))$ is controllable, it follows from [6, Prop. 3] that $W_A(t_0-nT, t_0)$ is positive definite. By Lemma 5, $Y'X$ is negative semidefinite. We conclude from (12) that $YRv = 0$. Then (9) implies that $Yv = 0$. But this is impossible since $Xv = 0$ and $\left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$ has rank n. Hence, X must be nonsingular. □

We are now prepared to describe all of the periodic equilibria for the PRDE using the above results for the SHPRDE as tools. Recall that $L^+(\Phi(T, 0))$ is the $\Phi(T, 0)$-invariant subspace associated with the eigenvalues of $\Phi(T, 0)$ inside the unit circle. If $\Phi(T, 0)$ has no eigenvalues on the unit circle, $L^+(\Phi(T, 0))$ is n-dimensional. In this case, it follows from Lemma 3 that $L^+(\Phi(T, 0)) \in \mathscr{L}(n)$. If $\Phi(t, 0)(L^+(\Phi(T, 0))) \in \mathscr{L}_0(n)$, $\forall t$, then it follows from Lemma 1 that $K(t, \gamma^{-1}(L^+(\Phi(T, 0))), 0)$ is a periodic equilibrium. If this solution exists, we will denote it by $K^+(t)$. It follows from the Corollary to Lemma 2 that if it exists, $K^+(t)$ is the unique periodic equilibrium with the property that every characteristic multiplier of the associated closed-loop system is inside the unit circle.

If $\Phi(T, 0)$ has no eigenvalues on the unit circle and $\Phi(t, 0)(L^-(\Phi(T, 0))) \in \mathscr{L}_0(n)$, $\forall t$, then another periodic equilibrium is given by $K(t, \gamma^{-1}(L^-(\Phi(T, 0))), 0)$. We denote

this solution by $K^-(t)$. If it exists, $K^-(t)$ is the unique periodic equilibrium with the property that every characteristic multiplier of the associated closed-loop system is outside the unit circle.

The next result shows that if it exists $K^+(t)$ is the maximal periodic equilibrium (in the partial ordering on symmetric matrices). This is true even if $(A(\cdot), B(\cdot))$ is not controllable.

THEOREM 1. *Suppose that $\Phi(T, 0)$ has no eigenvalues on the unit circle and that $K^+(t)$ exists. If $K(t)$ is any other periodic equilibrium, then $K(t) \leqq K^+(t)$.*

Proof. Let $A^+(t) = A(t) - B(t)B'(t)K^+(t)$ and consider the SHPRDE

$$\dot{\hat{K}}(t) = -A^{+\prime}(t)\hat{K}(t) - \hat{K}(t)A^+(t) + \hat{K}(t)B(t)B'(t)\hat{K}(t).$$

The associated Hamiltonian matrix is

$$\hat{H}(t) = \begin{bmatrix} A^+(t) & -B(t)B'(t) \\ 0 & -A^{+\prime}(t) \end{bmatrix}.$$

Let $\hat{\Phi}(t, t_0)$ denote the transition matrix corresponding to $\hat{H}(t)$. Let $K(t)$ be a periodic equilibrium of the PRDE. It is straightforward to check that $K(t) - K^+(t)$ is a periodic equilibrium of the SHPRDE. By Lemma 1, the subspace $\mathrm{Sp}\begin{bmatrix} I \\ K(t)-K^+(t) \end{bmatrix}$ is $\hat{\Phi}(t+T, t)$-invariant. By Lemma 5, $K(t) - K^+(t) \leqq 0$ which completes the proof. □

We will need to use the following known result which is an immediate consequence of [13, Theorem 4].

THEOREM 2. *Suppose that $(A(\cdot), B(\cdot))$ is controllable and that $\Phi(T, 0)$ has no eigenvalues on the unit circle. Then the periodic equilibrium $K^+(t)$ of the* PRDE *exists.*

Remark 1. It is assumed in [13] that $\Phi(T, 0)$ has distinct eigenvalues. However, this assumption is not essential to the proof of the result we have quoted. It should be noted that the results in [13] are stated for the filtering version of the Riccati equation and must be "dualized" to obtain the corresponding results for the control version of the Riccati equation which we are considering. When this is done, the assumptions used in [13] to prove the existence of $K^+(t)$ are that $\Phi(T, 0)$ have no eigenvalues on the unit circle and that $(A(\cdot), B(\cdot))$ satisfy a special stabilizability condition for periodic systems which was introduced in [12]. It is not obvious that the controllability of $(A(\cdot), B(\cdot))$ in the usual sense implies $(A(\cdot), B(\cdot))$ is stabilizable in the special sense. However, it was recently proven [2] that this is indeed the case. A different proof for the existence of $K^+(t)$ is given in [3].

The next result is crucial in order to establish the existence of periodic equilibria in addition to $K^+(t)$. It generalizes to the PRDE the result in Lemma 6 for the SHPRDE.

LEMMA 7. *Suppose that $(A(\cdot), B(\cdot))$ is controllable and that $\Phi(T, 0)$ has no eigenvalues on the unit circle. Then every subspace belonging to $\mathscr{L}(n)$ which is $\Phi(t_0+T, t_0)$-invariant belongs to $\mathscr{L}_0(n)$.*

Proof. Let $S \in \mathscr{L}(n)$ and suppose that S is $\Phi(t_0+T, t_0)$-invariant. $K^+(t)$ exists by Theorem 2. Let

$$U(t) = \begin{bmatrix} I & 0 \\ -K^+(t) & I \end{bmatrix}$$

and let $\hat{H}(t)$ and $\hat{\Phi}(t, t_0)$ be defined as in the proof of Theorem 1. It is straightforward to check that $\hat{H}(t) = \dot{U}(t)U^{-1}(t) + U(t)H(t)U^{-1}(t)$ and $\hat{\Phi}(t, t_0) = U(t)\Phi(t, t_0)U^{-1}(t_0)$. Since $U(t)$ is periodic, we have

$$\hat{\Phi}(t_0+T, t_0) = U(t_0)\Phi(t_0+T, t_0)U^{-1}(t_0). \tag{13}$$

Since $U(t_0)$ is symplectic, $U(t_0)(S) \in \mathscr{L}(n)$. From (13) it follows that $U(t_0)(S)$ is

$\hat{\Phi}(t_0+T, t_0)$-invariant. By Lemma 6, $U(t_0)(S) \in \mathscr{L}_0(n)$. Let $\begin{bmatrix} X \\ Y \end{bmatrix}$ be a basis for S. Then

$$U(t_0)(S) = \operatorname{Sp}\begin{bmatrix} X \\ -K^+(t_0)X + Y \end{bmatrix}.$$

Since $U(t_0)(S) \in \mathscr{L}_0(n)$, X is nonsingular. But this implies that $S \in \mathscr{L}_0(n)$. □

Lemma 7 is a key result. It shows that if $(A(\cdot), B(\cdot))$ is controllable and $\Phi(T, 0)$ has no eigenvalues on the unit circle, then every T-periodic solution of the EPRDE on $\mathscr{L}(n)$ is completely contained in $\mathscr{L}_0(n)$ and hence corresponds to a T-periodic solution of the PRDE on $S(n)$. On the other hand, there is a one-to-one correspondence between T-periodic solutions of the EPRDE on $\mathscr{L}(n)$ and the set of those elements of $\mathscr{L}(n)$ which are $\Phi(T, 0)$-invariant. Hence, there is a one-to-one correspondence between the set of $\Phi(T, 0)$-invariant elements of $\mathscr{L}(n)$ and the set of periodic equilibria of the PRDE. The explicit form of this correspondence is given by the next result.

THEOREM 3. *Suppose that $(A(\cdot), B(\cdot))$ is controllable and $\Phi(T, 0)$ has no eigenvalues on the unit circle. Each $\Phi(T, 0)$-invariant subspace $S_0 \in \mathscr{L}(n)$ determines a unique periodic equilibrium of the* PRDE *which is given by $\gamma^{-1}(\Phi(t, 0)(S_0))$. Furthermore, every periodic equilibrium is obtained in this way.*

Proof. Let $S_0 \in \mathscr{L}(n)$ and suppose that S_0 is $\Phi(T, 0)$-invariant. Then $\Phi(t+T, t)(S(t, S_0, 0)) = \Phi(t+T, t)\Phi(t, 0)(S_0) = \Phi(t+T, T)\Phi(T, 0)(S_0) = \Phi(t, 0)(S_0) = S(t, S_0, 0)$, so $S(t, S_0, 0)$ is $\Phi(t+T, t)$-invariant. By Lemma 7, $S(t, S_0, 0) \in \mathscr{L}_0(n)$, $\forall t$. By Lemma 1(b), $\gamma^{-1}(\Phi(t, 0)(S_0))$ is a periodic equilibrium of the PRDE.

Now suppose that $K(t)$ is a periodic equilibrium of the PRDE. Let $S_0 = \gamma(K(0))$. By Lemma 1(a), $\gamma(K(0))$ is $\Phi(T, 0)$-invariant and $S(t, S_0, 0) \in \mathscr{L}_0(n)$, $\forall t$. By (2), we have $K(t) = K(t, K(0), 0) = \gamma^{-1}(S(t, S_0, 0)) = \gamma^{-1}(\Phi(t, 0)(S_0))$. □

This result describes every periodic equilibrium of the PRDE. However, the description is given in terms of the set of $\Phi(T, 0)$-invariant subspaces which belong to $\mathscr{L}(n)$. Consequently, the next task is to describe this set.

LEMMA 8. *Suppose that $\Phi(T, 0)$ has no eigenvalues on the unit circle. Then the mapping $S \overset{\delta}{\mapsto} S \cap L^+(\Phi(T, 0))$ is a bijection of the set of $\Phi(T, 0)$-invariant subspaces which belong to $\mathscr{L}(n)$ onto the set of $\Phi(T, 0)$-invariant subspaces of $L^+(\Phi(T, 0))$. If S_1 is a $\Phi(T, 0)$-invariant subspace of $L^+(\Phi(T, 0))$, then $\delta^{-1}(S_1) = S_1 \oplus ([J(S_1)]^\perp \cap L^-(\Phi(T, 0)))$.*

Proof. Suppose that $S \in \mathscr{L}(n)$ and that S is $\Phi(T, 0)$-invariant. Since S is $\Phi(T, 0)$-invariant, we have

$$S = [S \cap L^+(\Phi(T, 0))] \oplus [S \cap L^-(\Phi(T, 0))].$$

Let $S_1 = S \cap L^+(\Phi(T, 0))$ and let $S_2 = S \cap L^-(\Phi(T, 0))$. Since $S \in \mathscr{L}(n)$, $[J(S)]^\perp = S$, so $S_2 \subseteq [J(S_1)]^\perp \cap L^-(\Phi(T, 0))$. Since S is $\Phi(T, 0)$-invariant, so is S_1. Since $\Phi(T, 0)$ is symplectic, $[J(S_1)]^\perp$ is also $\Phi(T, 0)$-invariant. Hence,

$$[J(S_1)]^\perp = ([J(S_1)]^\perp \cap L^+(\Phi(T, 0))) \oplus ([J(S_1)]^\perp \cap L^-(\Phi(T, 0))).$$

Since $L^+(\Phi(T, 0)) \in \mathscr{L}(n)$ and $S_1 \subseteq L^+(\Phi(T, 0))$, it follows that $[J(S_1)]^\perp \cap L^+(\Phi(T, 0)) = L^+(\Phi(T, 0))$. Hence,

$$[J(S_1)]^\perp = L^+(\Phi(T, 0)) \oplus ([J(S_1)]^\perp \cap L^-(\Phi(T, 0))).$$

Thus, $\dim [J(S_1)]^\perp \cap L^-(\Phi(T, 0)) = (2n - \dim S_1) - n = n - \dim S_1 = \dim S_2$. We conclude that $S_2 = [J(S_1)]^\perp \cap L^-(\Phi(T, 0))$. Since S_2 is uniquely determined by S_1, it follows that δ is injective. It also shows that if S_1 belongs to the image of δ, then $\delta^{-1}(S_1) = S_1 \oplus ([J(S_1)]^\perp \cap L^-(\Phi(T, 0)))$.

It remains only to show that every $\Phi(T,0)$-invariant subspace of $L^+(\Phi(T,0))$ belongs to the image of δ. Let S_1 be a $\Phi(T,0)$-invariant subspace of $L^+(\Phi(T,0))$. Let $S = S_1 \oplus ([J(S_1)]^\perp \cap L^-(\Phi(T,0)))$. We claim that $S \in \mathscr{L}(n)$, S is $\Phi(T,0)$-invariant, and that $\delta(S) = S_1$. By the same argument as above, we have $\dim [J(S_1)]^\perp \cap L^-(\Phi(T,0)) = n - \dim S_1$. Hence, $\dim S = n$. The $\Phi(T,0)$-invariance of S_1 implies the $\Phi(T,0)$-invariance of $[J(S_1)]^\perp \cap L^-(\Phi(T,0))$ and consequently the $\Phi(T,0)$-invariance of S. Since $S_1 \subseteq L^+(\Phi(T,0))$ and $[J(S_1)]^\perp \cap L^-(\Phi(T,0)) \subseteq L^-(\Phi(T,0))$, the fact that $L^+(\Phi(T,0)) \in \mathscr{L}(n)$ and $L^-(\Phi(T,0)) \in \mathscr{L}(n)$ implies that $J(S_1) \perp S_1$ and $J([J(S_1)]^\perp \cap L^-(\Phi(T,0))) \perp [J(S_1)]^\perp \cap L^-(\Phi(T,0))$. This together with the trivial fact that $J(S_1) \perp [J(S_1)]^\perp \cap L^-(\Phi(T,0))$ implies that $J(S) \perp S$, so $J \in \mathscr{L}(n)$. Finally, it follows from the definition of S that $\delta(S) = S_1$. □

Using this result, we immediately obtain a refinement of Theorem 3.

THEOREM 4. *Suppose that $(A(\cdot), B(\cdot))$ is controllable and $\Phi(T,0)$ has no eigenvalues on the unit circle. Each $\Phi(T,0)$-invariant subspace of $L^+(\Phi(T,0))$ determines a unique periodic equilibrium of the* PRDE. *If S_1 is such a subspace, the corresponding periodic equilibrium is given by $\gamma^{-1}(\Phi(t,0)(S_1 \oplus ([J(S_1)]^\perp \cap L^-(\Phi(T,0)))))$. Furthermore, every periodic equilibrium of the* PRDE *is obtained in this way.*

This result is an improvement over Theorem 3 because it describes the periodic equilibria in terms of the invariant subspaces of $\Phi(T,0)|L^+(\Phi(T,0))$, which is a set of known structure. On the other hand, Theorem 3 describes the periodic equilibria in terms of the set of n-dimensional $\Phi(T,0)$-invariant subspaces which satisfy an additional condition, namely that $[J(S)]^\perp = S$.

Theorem 4 gives a one-to-one correspondence between the set of periodic equilibria and the set of $\Phi(T,0)$-invariant subspaces of $L^+(\Phi(T,0))$. By Lemma 2, $\psi_{A^+}(T,0)$ is the matrix for $\Phi(T,0)|L^+(\Phi(T,0))$ with respect to the basis given by the columns of $\begin{bmatrix} I \\ K^+(0) \end{bmatrix}$. Thus, there is a one-to-one correspondence between the set of periodic equilibria and the set of $\psi_{A^+}(T,0)$-invariant subspaces of $\mathbb{R}^n$. Using this fact, we will prove a very useful theorem which shows how every periodic equilibrium can be constructed from the pair $K^+(t)$ and $K^-(t)$.

We need two preliminary results.

LEMMA 9. *Suppose that $(A(\cdot), B(\cdot))$ is controllable and $\Phi(T,0)$ has no eigenvalues on the unit circle. Let $\Delta(t) = K^+(t) - K^-(t)$. Then $\Delta(t) > 0$, $\forall t$.*

Proof. Since $(A(\cdot), B(\cdot))$ is controllable and $\Phi(T,0)$ has no eigenvalues on the unit circle, it follows from Theorem 3 that both $K^+(t)$ and $K^-(t)$ exist. By Theorem 1, $\Delta(t) \geqq 0$. It is trivial that $\Delta(t)$ is singular if and only if $\gamma(K^+(t)) \cap \gamma(K^-(t)) \neq 0$. However, it follows easily from the definitions of $K^+(t)$ and $K^-(t)$ that $\gamma(K^+(t)) = L^+(\Phi(t+T,t))$ and $\gamma(K^-(t)) = L^-(\Phi(t+T,t))$, which implies that $\gamma(K^+(t)) \cap \gamma(K^-(t)) = 0$. Hence, $\Delta(t) > 0$. □

LEMMA 10. *Suppose that $(A(\cdot), B(\cdot))$ is controllable and that $\Phi(T,0)$ has no eigenvalues on the unit circle. Let $A^-(t) = A(t) - B(t)B'(t)K^-(t)$. Then*

$$\psi'_{A^-}(t,t_0)\Delta(t)\psi_{A^+}(t,t_0) = \Delta(t_0) \quad \forall t, t_0.$$

Proof. The equation holds trivially for $t = t_0$, so it suffices to show that the left-hand side is constant as a function of t. We have

$$\frac{d}{dt}[\psi'_{A^-}(t,t_0)\Delta(t)\psi_{A^+}(t,t_0)] = \psi'_{A^-}(t,t_0)[A^{-\prime}(t)\Delta(t) + \Delta(t)A^+(t) + \dot{\Delta}(t)]\psi_{A^+}(t,t_0).$$

It is straightforward to check that the quantity in brackets on the right-hand side is identically zero. □

COROLLARY. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and* $\Phi(T, 0)$ *has no eigenvalues on the unit circle. Then*

$$\psi'_{A^-}(t_0+T, t_0)\Delta(t_0)\psi_{A^+}(t_0+T, t_0)=\Delta(t_0) \quad \forall t_0.$$

Proof. The proof follows immediately from Lemma 10 and the periodicity of $\Delta(t)$. □

The following theorem shows how every periodic equilibrium can be constructed from $K^+(t)$ and $K^-(t)$. It generalizes to the periodic RDE a well-known result of J. C. Willems [24] for the time-invariant RDE.

THEOREM 5. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and* $\Phi(T, 0)$ *has no eigenvalues on the unit circle. There is a one-to-one correspondence between the set of periodic equilibria of the* PRDE *and the set of* $\psi_{A^+}(T, 0)$*-invariant subspaces of* $\mathbb{R}^n$. *Let* M_0 *be a* $\psi_{A^+}(T, 0)$*-invariant subspace, and let* $M(t)=\psi_{A^+}(t, 0)(M_0)$, *and let* $\Pi(t)$ *be the projection onto* $M(t)$ *along* $\Delta^{-1}(t)(M(t)^\perp)$. *The periodic equilibrium corresponding to* M_0 *is*

$$K(t)=K^+(t)\Pi(t)+K^-(t)[I-\Pi(t)].$$

Furthermore, $L^+(\psi_{A-BB'K}(t+T, t))=M(t)$ *and* $L^-(\psi_{A-BB'K}(t+T, t))=\Delta^{-1}(t)(M(t)^\perp)$.

Proof. From the comments following the proof of Theorem 4, we already know there is a one-to-one correspondence between the set of periodic equilibria and the set of $\psi_{A^+}(T, 0)$-invariant subspaces of $\mathbb{R}^n$. What remains is to show that the correspondence has the asserted form.

The one-to-one correspondence between periodic equilibria and $\psi_{A^+}(T, 0)$-invariant subspaces results from the composition of three one-to-one correspondences:

(1) There is a one-to-one correspondence between the set of periodic equilibria and the set of $\Phi(T, 0)$-invariant subspaces which belong to $\mathscr{L}(n)$, which associates the subspace S with the periodic equilibrium $K(t)=\gamma^{-1}(\Phi(t, 0)(S))$.

(2) There is a one-to-one correspondence between the set of $\Phi(T, 0)$-invariant subspaces which belong to $\mathscr{L}(n)$ and the set of $\Phi(T, 0)$-invariant subspaces of $L^+(\Phi(T, 0))$. If S_1 is a $\Phi(T, 0)$-invariant subspace of $L^+(\Phi(T, 0))$ then the corresponding $\Phi(T, 0)$-invariant subspace which belongs to $\mathscr{L}(n)$ is $S=S_1\oplus([J(S_1)]^\perp\cap L^-(\Phi(T, 0)))$.

(3) There is a one-to-one correspondence between the set of $\Phi(T, 0)$-invariant subspaces of $L^+(\Phi(T, 0))$ and the set of $\psi_{A^+}(T, 0)$-invariant subspaces of $\mathbb{R}^n$. If M is a $\psi_{A^+}(T, 0)$-invariant subspace of $\mathbb{R}^n$, then the corresponding $\Phi(T, 0)$-invariant subspace of $L^+(\Phi(T, 0))$ is

$$S_1=\begin{bmatrix} I \\ K^+(0) \end{bmatrix}(M).$$

We are given M_0, a $\psi_{A^+}(T, 0)$-invariant subspace of $\mathbb{R}^n$. For each t, $M(t)$ is defined to be the subspace $\psi_{A^+}(t, 0)(M_0)$. Since $\Delta(t)>0$ by Lemma 9, $\Delta^{-1}(t)(M(t)^\perp)$ is complementary to $M(t)$, so the projection $\Pi(t)$ is defined. Let $\Pi_0=\Pi(0)$. In other words, Π_0 is the projection onto M_0 along $\Delta^{-1}(0)(M_0^\perp)$. The $\Phi(T, 0)$-invariant subspace of $L^+(\Phi(T, 0))$ which corresponds to M_0 is

$$S_1=\begin{bmatrix} I \\ K^+(0) \end{bmatrix}(M_0)=\mathrm{Sp}\begin{bmatrix} I \\ K^+(0) \end{bmatrix}\Pi_0.$$

The $\Phi(T, 0)$-invariant subspace belonging to $\mathscr{L}(n)$ which corresponds to S_1 is $S=S_1\oplus([J(S_1)]^\perp\cap L^-(\Phi(T, 0)))$.

We claim that

$$[J(S_1)]^\perp\cap L^-(\Phi(T, 0))=\mathrm{Sp}\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I-\Pi_0).$$

First, we note that

$$\dim \operatorname{Sp}\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I-\Pi_0) = \dim \operatorname{Sp}(I-\Pi_0) = n - \dim M_0 = n - \dim S_1$$

$$= \dim([J(S_1)]^\perp \cap L^-(\Phi(T,0))).$$

Consequently, to prove the claim, it suffices to show that

$$\operatorname{Sp}\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I-\Pi_0) \subseteq [J(S_1)]^\perp \cap L^-(\Phi(T,0)).$$

Since

$$\operatorname{Sp}\begin{bmatrix} I \\ K^-(0) \end{bmatrix} = L^-(\Phi(T,0)),$$

we have

$$\operatorname{Sp}\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I-\Pi_0) \subseteq L^-(\Phi(T,0)).$$

So we need only show that

$$\operatorname{Sp}\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I-\Pi_0) \subseteq [J(S_1)]^\perp.$$

To show this, we must prove that

$$(I-\Pi_0)'[I \;\; K^-(0)]J\begin{bmatrix} I \\ K^+(0) \end{bmatrix}\Pi_0 = 0,$$

which is equivalent to $(I-\Pi_0)'\Delta(0)\Pi_0 = 0$. Since Π_0 is the projection onto M_0 along $\Delta^{-1}(0)(M_0^\perp)$, it follows immediately that this equation holds, which proves the claim. Thus, the $\Phi(T,0)$-invariant subspace belonging to $\mathscr{L}(n)$ which corresponds to S_1 is given by

$$S = \operatorname{Sp}\begin{bmatrix} I \\ K^+(0) \end{bmatrix}\Pi_0 \oplus \operatorname{Sp}\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I-\Pi_0). \tag{14}$$

The periodic equilibrium which corresponds to S is $K(t) = \gamma^{-1}(\Phi(t,0)(S))$. We claim that

$$\Phi(t,0)(S) = \operatorname{Sp}\begin{bmatrix} I \\ K^+(t)\Pi(t) + K^-(t)(I-\Pi(t)) \end{bmatrix}. \tag{15}$$

Note that it follows immediately from this claim that $\gamma^{-1}(\Phi(t,0)(S)) = K^+(t)\Pi(t) + K^-(t)(I-\Pi(t))$ which would prove the first assertion of the theorem. From (14), we have

$$\Phi(t,0)(S) = \operatorname{Sp}\Phi(t,0)\begin{bmatrix} I \\ K^+(0) \end{bmatrix}\Pi_0 \oplus \operatorname{Sp}\Phi(t,0)\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I-\Pi_0). \tag{16}$$

From the first paragraph of the proof of Lemma 2, it follows that if $K_1(t)$ is any periodic equilibrium of the PRDE, then

$$\Phi(t,t_0)\begin{bmatrix} I \\ K_1(t_0) \end{bmatrix} = \begin{bmatrix} I \\ K_1(t) \end{bmatrix}\psi_{A-BB'K_1}(t,t_0). \tag{17}$$

Applying (17) with $K_1(t) = K^+(t)$ gives

$$\operatorname{Sp} \Phi(t, 0)\begin{bmatrix} I \\ K^+(0) \end{bmatrix}\Pi_0 = \operatorname{Sp}\begin{bmatrix} I \\ K^+(t) \end{bmatrix}\psi_{A^+}(t, 0)\Pi_0.$$

Since $M(t) = \psi_{A^+}(t, 0)(M_0)$, we obtain

$$\operatorname{Sp} \Phi(t, 0)\begin{bmatrix} I \\ K^+(0) \end{bmatrix}\Pi_0 = \operatorname{Sp}\begin{bmatrix} I \\ K^+(t) \end{bmatrix}\Pi(t). \tag{18}$$

Applying (17) with $K_1(t) = K^-(t)$ gives

$$\operatorname{Sp} \Phi(t, 0)\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I - \Pi_0) = \operatorname{Sp}\begin{bmatrix} I \\ K^-(t) \end{bmatrix}\psi_{A^-}(t, 0)(I - \Pi_0).$$

Using Lemma 10, we have

$$\begin{aligned} \operatorname{Sp} \psi_{A^-}(t, 0)(I - \Pi_0) &= \operatorname{Sp} \Delta^{-1}(t)\psi_{A^+}(0, t)\Delta(0)(I - \Pi_0) \\ &= \Delta^{-1}(t)\psi_{A^+}(0, t)(M_0^\perp) \\ &= \Delta^{-1}(t)(M(t)^\perp) = \operatorname{Sp}(I - \Pi(t)). \end{aligned} \tag{19}$$

Thus,

$$\operatorname{Sp} \Phi(t, 0)\begin{bmatrix} I \\ K^-(0) \end{bmatrix}(I - \Pi_0) = \operatorname{Sp}\begin{bmatrix} I \\ K^-(t) \end{bmatrix}(I - \Pi(t)). \tag{20}$$

Using (18) and (20) in (16), we obtain

$$\begin{aligned} \Phi(t, 0)(S) &= \operatorname{Sp}\begin{bmatrix} \Pi(t) & I - \Pi(t) \\ K^+(t)\Pi(t) & K^-(t)(I - \Pi(t)) \end{bmatrix} \\ &= \operatorname{Sp}\begin{bmatrix} I \\ K^+(t)\Pi(t) + K^-(t)(I - \Pi(t)) \end{bmatrix}. \end{aligned}$$

This establishes the validity of (15) and shows that the periodic equilibrium which corresponds to M_0 is in fact $K(t) = K^+(t)\Pi(t) + K^-(t)(I - \Pi(t))$ as asserted.

Now we prove that $L^+(\psi_{A-BB'K}(t+T, t)) = M(t)$ and $L^-(\psi_{A-BB'K}(t+T, t)) = \Delta^{-1}(t)(M(t)^\perp)$. Since M_0 is $\psi_{A^+}(T, 0)$-invariant, we have $\psi_{A^+}(t+T, t)(M(t)) = \psi_{A^+}(t+T, t)\psi_{A^+}(t, 0)(M_0) = \psi_{A^+}(t+T, T)\psi_{A^+}(T, 0)(M_0) = \psi_{A^+}(t, 0)(M_0) = M(t)$. Hence, $M(t)$ is $\psi_{A^+}(t+T, t)$-invariant. It follows from this together with the Corollary to Lemma 10 that $\Delta^{-1}(t)(M(t)^\perp)$ is $\psi_{A^-}(t+T, t)$-invariant.

We claim that $\psi_{A-BB'K}(t+T, t)$ agrees with $\psi_{A^+}(t+T, t)$ on $M(t)$ and agrees with $\psi_{A^-}(t+T, t)$ on $\Delta^{-1}(t)(M(t)^\perp)$. Let $z_0 \in M_0$, and let $z(\tau) = \psi_{A^+}(\tau, 0)z_0$. From the part of the theorem which has already been proven, we know that $K(\tau)$ agrees with $K^+(\tau)$ on $M(\tau)$. Since $z(\tau) \in M(\tau)$, this implies that $(d/d\tau)z(\tau) = A^+(\tau)z(\tau) = (A(\tau) - B(\tau)B'(\tau)K(\tau))z(\tau)$. Consequently, $z(\tau) = \psi_{A-BB'K}(\tau, 0)z_0$. Hence,

$$\psi_{A-BB'K}(\tau, 0)z_0 = \psi_{A^+}(\tau, 0)z_0 \quad \forall \tau.$$

In particular, we have $\psi_{A-BB'K}(t+T, t)z(t) = \psi_{A-BB'K}(t+T, t)\psi_{A-BB'K}(t, 0)z_0 = \psi_{A-BB'K}(t+T, 0)z_0 = \psi_{A^+}(t+T, 0)z_0 = \psi_{A^+}(t+T, t)\psi_{A^+}(t, 0)z_0 = \psi_{A^+}(t+T, t)z(t)$. Since $z(t)$ is an arbitrary element of $M(t)$, this shows that $\psi_{A-BB'K}(t+T, t)$ agrees with $\psi_{A^+}(t+T, t)$ on $M(t)$.

Let $y_0 \in \Delta^{-1}(0)(M_0^\perp)$, and let $y(\tau) = \psi_{A^-}(\tau, 0)y_0$. From (19), we know that $y(\tau) \in \Delta^{-1}(\tau)(M(\tau)^\perp)$, and every element of $\Delta^{-1}(\tau)(M(\tau)^\perp)$ can be expressed in the

form $\psi_{A^-}(\tau,0)y_0$ for some choice of $y_0 \in \Delta^{-1}(0)(M_0^{\perp})$. From the part of the theorem which has already been proven, we know that $K(\tau)$ agrees with $K^-(\tau)$ on $\Delta^{-1}(\tau)(M(\tau)^{\perp})$. Since $y(\tau) \in \Delta^{-1}(\tau)(M(\tau)^{\perp})$, this implies that $(d/d\tau)y(\tau) = A^-(\tau)y(\tau) = (A(\tau) - B(\tau)B'(\tau)K(\tau))y(\tau)$. Hence, $y(\tau) = \psi_{A-BB'K}(\tau,0)y_0$. Thus,

$$\psi_{A-BB'K}(\tau,0)y_0 = \psi_{A^-}(\tau,0)y_0 \quad \forall \tau.$$

Consequently,

$$\begin{aligned}\psi_{A-BB'K}(t+T,t)y(t) &= \psi_{A-BB'K}(t+T,t)\psi_{A-BB'K}(t,0)y_0 \\ &= \psi_{A-BB'K}(t+T,0)y_0 = \psi_{A^-}(t+T,0)y_0.\end{aligned}$$

$$\psi_{A^-}(t+T,t)\psi_{A^-}(t,0)y_0 = \psi_{A^-}(t+T,t)y(t).$$

Since $y(t)$ is an arbitrary element of $\Delta^{-1}(t)(M(t)^{\perp})$, this shows that $\psi_{A-BB'K}(t+T,t)$ agrees with $\psi_{A^-}(t+T,t)$ on $\Delta^{-1}(t)(M(t)^{\perp})$.

At this point, we have proven that $M(t)$ and $\Delta^{-1}(t)(M(t)^{\perp})$ are complementary subspaces, that $M(t)$ is $\psi_{A^+}(t+T,t)$-invariant and $\Delta^{-1}(t)(M(t)^{\perp})$ is $\psi_{A^-}(t+T,t)$-invariant, that $\psi_{A-BB'K}(t+T,t)$ agrees with $\psi_{A^+}(t+T,t)$ on $M(t)$ and with $\psi_{A^-}(t+T,t)$ on $\Delta^{-1}(t)(M(t)^{\perp})$. In particular, $M(t)$ and $\Delta^{-1}(t)(M(t)^{\perp})$ are both $\psi_{A-BB'K}(t+T,t)$-invariant. Since every eigenvalue of $\psi_{A^+}(t+T,t)$ is inside the unit circle and every eigenvalue of $\psi_{A^-}(t+T,t)$ is outside the unit circle, it follows immediately that $L^+(\psi_{A-BB'K}(t+T,t)) = M(t)$ and $L^-(\psi_{A-BB'K}(t+T,t)) = \Delta^{-1}(t)(M(t)^{\perp})$, which completes the proof. □

Remark 2. Since $M(t)$ is completely determined by the choice of the $\psi_{A^+}(T,0)$-invariant subspace M_0, it is possible to give a formula which expresses the projection $\Pi(t)$ in terms of the projection Π_0. The formula is given by

$$\Pi(t) = \psi_{A^+}(t,0)\Pi_0[\psi_{A^+}(t,0)\Pi_0 + \psi_{A^-}(t,0)(I-\Pi_0)]^{-1}. \tag{21}$$

To prove this formula, let $\tilde{\Pi}(t)$ denote the right-hand side. First we show that the indicated inverse exists. Suppose not. Then there exists $x \in \mathbb{R}^n$ such that $\psi_{A^+}(t,0)\Pi_0 x = -\psi_{A^-}(t,0)(I-\Pi_0)x$. From the definition of $M(t)$, it is clear that $\psi_{A^+}(t,0)\Pi_0 x \in M(t)$. From (19) it follows that $-\psi_{A^-}(t,0)(I-\Pi_0)x \in \Delta^{-1}(t)(M(t)^{\perp})$. Since $M(t)$ and $\Delta^{-1}(t)(M(t)^{\perp})$ are complementary, we obtain $\Pi_0 x = 0$ and $(I-\Pi_0)x = 0$, which implies that $x = 0$. Thus, the indicated inverse exists.

Next we show that $\tilde{\Pi}(t)$ is the identity on $M(t)$. Let $x(t) \in M(t)$. Then there exists $x_0 \in M_0$ such that $x(t) = \psi_{A^+}(t,0)x_0$. Now, $[\psi_{A^+}(t,0)\Pi_0 + \psi_{A^-}(t,0)(I-\Pi_0)]x_0 = x(t)$. Hence, $\tilde{\Pi}(t)x(t) = x(t)$.

Finally, we show that $\tilde{\Pi}(t)$ is zero on $\Delta^{-1}(t)(M(t)^{\perp})$. Let $y(t) \in \Delta^{-1}(t)(M(t)^{\perp})$. By (19), there exists $y_0 \in \Delta^{-1}(0)(M_0^{\perp})$ such that $y(t) = \psi_{A^-}(t,0)y_0$. Now, $[\psi_{A^+}(t,0)\Pi_0 + \psi_{A^-}(t,0)(I-\Pi_0)]y_0 = y(t)$, from which it follows that $\tilde{\Pi}(t)y(t) = 0$. We conclude that $\tilde{\Pi}(t)$ is the projection onto $M(t)$ along $\Delta^{-1}(t)(M(t)^{\perp})$, which establishes (21).

Remark 3. It is interesting to consider the specialization of Theorem 5 to the case where $A(\cdot)$, $B(\cdot)$, and $C(\cdot)$ are constant matrices A, B, and C (and hence T-periodic for any $T > 0$). In this case, $K^+(t) = \gamma^{-1}(\Phi(t,0)(L^+(\Phi(T,0)))) = \gamma^{-1}(e^{Ht}(L^+(e^{HT})))$. It is clear that $L^+(e^{HT})$ is the sum of the primary components of H corresponding to its n eigenvalues in the left half-plane. Thus, $L^+(e^{HT})$ is H-invariant (and hence e^{Ht}-invariant as well). Consequently, the solution $K^+(t)$ is constant. Let K^+ denote this constant solution. Similarly, the solution $K^-(t)$ is constant which we denote by K^-. Let $A^+ \equiv A - BB'K^+$ and let $\Delta \equiv K^+ - K^-$. Then Theorem 5 specializes to give the next Corollary.

COROLLARY. *Suppose that A, B, C are constant with (A, B) controllable, and suppose that H has no eigenvalues on the imaginary axis. There is a one-to-one correspondence*

between the set of T-periodic solutions of the RDE *and the set of* e^{A^+T}*-invariant subspaces of* $\mathbb{R}^n$. *Let* M_0 *be an* e^{A^+T}*-invariant subspace, let* $M(t) \equiv e^{A^+t}(M_0)$, *and let* $\Pi(t)$ *be the projection onto* $M(t)$ *along* $\Delta^{-1}(M(t)^\perp)$. *The T-periodic solution corresponding to* M_0 *is* $K(t) = K^+\Pi(t) + K^-(I - \Pi(t))$. *Furthermore,* $L^+(\psi_{A-BB'K}(t+T, t)) = M(t)$ *and* $L^-(\psi_{A-BB'K}(t+T, t)) = \Delta^{-1}(M(t)^\perp)$.

This result generalizes Willems' classification of the equilibria of the time-invariant RDE [24, Thm. 6] to include the periodic solutions as well. A direct proof of a similar generalization appears in our paper [20].

The next result shows that the bijection described by Theorem 5 has the additional property of being order-preserving.

THEOREM 6. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and that* $\Phi(T, 0)$ *has no eigenvalues on the unit circle. Let* M_0 *and* $\tilde{M}_0$ *be* $\psi_{A^+}(T, 0)$*-invariant subspaces, and let* $K(t)$ *and* $\tilde{K}(t)$ *be the periodic equilibria which correspond to* M_0 *and* $\tilde{M}_0$ *respectively. Then* $K(t) \leqq \tilde{K}(t)$ *if and only if* $M_0 \subseteq \tilde{M}_0$.

Proof. Let $M(t) = \psi_{A^+}(t, 0)(M_0)$ and let $\tilde{M}(t) = \psi_{A^+}(t, 0)(\tilde{M}_0)$. Let $\Pi(t)$ be the projection onto $M(t)$ along $\Delta^{-1}(t)(M(t)^\perp)$, and let $\tilde{\Pi}(t)$ be the projection onto $\tilde{M}(t)$ along $\Delta^{-1}(t)(\tilde{M}(t)^\perp)$. It follows from Theorem 5 that

$$\tilde{K}(t) - K(t) = \Delta(t)[\tilde{\Pi}(t) - \Pi(t)]. \tag{22}$$

Suppose that $M_0 \subseteq \tilde{M}_0$. Then $M(t) \subseteq \tilde{M}(t)$, and $\Delta^{-1}(t)(\tilde{M}(t)^\perp) \subseteq \Delta^{-1}(t)(M(t)^\perp)$. Using the fact that $M(t) \oplus \Delta^{-1}(t)(M(t)^\perp) = \mathbb{R}^n$ and $\tilde{M}(t) \oplus \Delta^{-1}(t)(\tilde{M}(t)^\perp) = \mathbb{R}^n$, it is easy to show that

$$M(t) \oplus [\Delta^{-1}(t)(M(t)^\perp) \cap \tilde{M}(t)] \oplus [\Delta^{-1}(t)(\tilde{M}(t)^\perp)] = \mathbb{R}^n.$$

Thus, given $x \in \mathbb{R}^n$, we can write $x = u + v + w$ with $u \in M(t)$, $v \in \Delta^{-1}(t)(M(t)^\perp) \cap \tilde{M}(t)$ and $w \in \Delta^{-1}(t)(\tilde{M}(t)^\perp)$. Then $x'(\tilde{K}(t) - K(t))x = x'\Delta(t)[\tilde{\Pi}(t) - \Pi(t)]x = x'\Delta(t)[u + v - u] = (u' + v' + w')\Delta(t)v = v'\Delta(t)v \geqq 0$. Hence, $K(t) \leq \tilde{K}(t)$.

Now suppose that $M_0 \not\subseteq \tilde{M}_0$. Then $\Delta^{-1}(t)(\tilde{M}(t)^\perp) \not\subseteq \Delta^{-1}(t)(M(t)^\perp)$. Let $y \in \Delta^{-1}(t)(\tilde{M}(t)^\perp)$ with $y \notin \Delta^{-1}(t)(M(t)^\perp)$. Since $M(t) \oplus \Delta^{-1}(t)(M(t)^\perp) = \mathbb{R}^n$, we can write $y = u + z$ with $u \in M(t)$, $z \in \Delta^{-1}(t)(M(t)^\perp)$, and with $u \neq 0$. Then $y'(\tilde{K}(t) - K(t))y = y'\Delta(t)[\tilde{\Pi}(t) - \Pi(t)]y = -y'\Delta(t)\Pi(t)y = -y'\Delta(t)u = -u'\Delta(t)u < 0$. Thus, $\tilde{K}(t) - K(t)$ is not nonnegative definite. □

As an immediate consequence of the preceding theorem, we have the next corollary.

COROLLARY 1. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and* $\Phi(T, 0)$ *has no eigenvalues on the unit circle. Then every periodic equilibrium* $K(t)$ *satisfies*

$$K^-(t) \leqq K(t) \leqq K^+(t), \ Vt.$$

By definition, a *complete lattice* is a partially ordered set with the property that every subset has a least upper bound and a greatest lower bound. Since the set of all invariant subspaces of a given square matrix (partially ordered by inclusion) is a complete lattice, the next result follows directly from Theorem 6.

COROLLARY 2. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and* $\Phi(T, 0)$ *has no eigenvalues on the unit circle. Then the set of periodic equilibria is a complete lattice with respect to the usual partial ordering of symmetric matrices.*

The preceding corollary generalizes to the PRDE results of Coppel [8] for the equilibria of the time-invariant RDE and of Shayman [20] for the T-periodic solutions of the time-invariant RDE.

In the case where the pair $(C(\cdot), A(\cdot))$ is observable, the following result of Shayman [19] describes the signature of each periodic equilibrium.

THEOREM 7. *Suppose that* $(C(\cdot), A(\cdot))$ *is observable and there exists a periodic equilibrium* $K_1(t)$ *to the* PRDE. *Let* $A_1(t) \equiv A(t) - B(t)B'(t)K_1(t)$. *Then* $A_1(\cdot)$ *has no characteristic multipliers on the unit circle,* $K_1(t)$ *is nonsingular, and the number of positive (negative) eigenvalues of* $K_1(t)$ *is equal to the number of characteristic multipliers of* $A_1(\cdot)$ *of modulus less (greater) than* 1.

COROLLARY. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and* $(C(\cdot), A(\cdot))$ *is observable. Let* M_0 *be a* $\psi_A{}^+(T, 0)$*-invariant subspace, and let* $K(t)$ *be the corresponding periodic equilibria of the* PRDE. *Then* $K(t)$ *is nonsingular, and the number of positive eigenvalues of* $K(t)$ *is equal to the dimension of* M_0.

Proof. It follows from [13] that the controllability of $(A(\cdot), B(\cdot))$ and observability of $(C(\cdot), A(\cdot))$ imply that $\Phi(T, 0)$ has no eigenvalues on the unit circle. By Theorem 7, $K(t)$ is nonsingular, and the number of positive eigenvalues of $K(t)$ is equal to the number of eigenvalues of $\psi_{A-BB'K}(T, 0)$ inside the unit circle, or equivalently the dimension of $L^+(\psi_{A-BB'K}(T, 0))$. By Theorem 5, $L^+(\psi_{A-BB'K}(T, 0)) = M_0$. □

Remark 4. The reference [13] assumes that $\Phi(T, 0)$ has distinct eigenvalues. However, this assumption is not needed for the proof of the result referred to in our proof of the preceding corollary.

Remark 5. An immediate consequence of the preceding corollary is that if $(A(\cdot), B(\cdot))$ is controllable and $(C(\cdot), A(\cdot))$ is observable, then $K^+(t)$ is the unique positive definite periodic equilibrium, and $K^-(t)$ is the unique negative definite periodic equilibrium. Every other periodic equilibrium is nonsingular (and hence of constant signature) but indefinite.

Theorem 5 describes every periodic equilibrium of the PRDE by giving a one-to-one correspondence between the set of periodic equilibria and the set of invariant subspaces of the $n \times n$ matrix $\psi_A{}^+(T, 0)$. This result is valid even if the monodromy matrix $\Phi(T, 0)$ has multiple eigenvalues and even if it is nondiagonalizable. The result has several additional attractive features. Geometric properties of the set of generators of the periodic equilibria (i.e. the set $\{K_0\colon K(t, K_0, 0)$ is a periodic equilibrium$\}$) can be obtained from recent results on the geometry of the space of invariant subspaces of a finite-dimensional linear operator [21]. Since by Theorem 6 the bijection is order-preserving, it describes the periodic equilibria not simply as a set, but as a lattice. In other words, the classification describes the partial ordering of the periodic equilibria in addition to describing the periodic equilibria themselves. Furthermore, it follows from the Corollary to Theorem 7 that if $(C(\cdot), A(\cdot))$ is observable, the signature of each periodic equilibrium is completely determined by the dimension of the associated $\psi_A{}^+(T, 0)$-invariant subspace.

Our classification of the periodic equilibria is quite different from that given in [1]. Also, the results in that reference require that $\Phi(T, 0)$ have distinct eigenvalues and that $K^+(t)$ and $K^-(t)$ be invertible. The invertibility assumption excludes the PRDE's arising from many problems of interest such as those corresponding to nonobservable systems or to control problems with conflicting objectives.

4. Stability of periodic equilibria.

THEOREM 8. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and* $\Phi(T, 0)$ *has no eigenvalues on the unit circle. Then*

(a) $K^+(t)$ *is uniformly asymptotically stable in backward time.* $K^-(t)$ *is uniformly asymptotically stable in forward time. No other periodic equilibrium is asymptotically stable in either forward or backward time.*

(b) $K(t, K_0, t_0)$ *converges to* $K^+(t)$ *as* $t \to -\infty$ *if and only if* $K_0 > K^-(t_0)$.

(c) $K(t, K_0, t_0)$ *converges to* $K^-(t)$ *as* $t \to \infty$ *if and only if* $K_0 < K^+(t_0)$.

Proof. (a) Let $K_e(t)$ be a periodic equilibrium of the PRDE, and let $A_e(t) \equiv A(t) - B(t)B'(t)K_e(t)$. Let $\tilde{K}(t) \equiv K(t) - K_e(t)$. Then $\tilde{K}(t)$ satisfies the differential equation

$$\dot{\tilde{K}}(t) = -A_e'(t)\tilde{K}(t) - \tilde{K}(t)A_e(t) + \tilde{K}(t)B(t)B'(t)\tilde{K}(t). \tag{23}$$

By inspection, the linearization of (23) at the origin is

$$\dot{\tilde{K}} = -A_e'(t)\tilde{K}(t) - \tilde{K}(t)A_e(t). \tag{24}$$

The characteristic multipliers of this periodic linear system are $\{\lambda_i^{-1}\lambda_j^{-1}\}_{i \leqq j}$ where $\{\lambda_i\}_1^n$ are the characteristic multipliers associated with $A_e(\cdot)$. By the Corollary to Lemma 2, it follows that if $K_e(t) = K^+(t)$ every characteristic multiplier of (24) is outside the unit circle, while if $K_e(t) = K^-(t)$, every characteristic multiplier of (24) is inside the unit circle. If $K_e(t)$ is any other periodic equilibrium, (24) has at least one characteristic multiplier inside the unit circle and at least one characteristic multiplier outside the unit circle. The assertions of (a) follow immediately from the linearization principle [25, p. 127].

(b) Transform the PRDE by defining $\tilde{K}(t) = K(t) - K^-(t)$. Then $\tilde{K}(t)$ satisfies the differential equation

$$\dot{\tilde{K}}(t) = -A^{-\prime}(t)\tilde{K}(t) - \tilde{K}(t)A^-(t) + \tilde{K}(t)B(t)B'(t)\tilde{K}(t) \tag{25}$$

where $A^-(t) = A(t) - B(t)B'(t)K^-(t)$. Let $\tilde{K}(t, \tilde{K}_0, t_0)$ denote the solution of (25) which goes through the point $\tilde{K}_0$ at $t = t_0$. It suffices to show that $\tilde{K}(t, \tilde{K}_0, t_0)$ converges to $K^+(t) - K^-(t) \equiv \Delta(t)$ as $t \to -\infty$ if and only if $\tilde{K}_0 > 0$.

We claim that if $\tilde{K}_0$ is singular, then $\tilde{K}(t, \tilde{K}_0, t_0)$ is singular as long as the solution continues to exist. Similarly, if $\tilde{K}_0$ is nonsingular, then $\tilde{K}(t, \tilde{K}_0, t_0)$ is nonsingular as long as the solution continues to exist. To see this, rewrite (25) as

$$\dot{\tilde{K}}(t) = -A^{-\prime}(t)\tilde{K}(t) - \tilde{K}(t)(A^-(t) - B(t)B'(t)\tilde{K}(t)). \tag{26}$$

Thus, $\tilde{K}(t)$ satisfies a homogeneous *linear* matrix differential equation. Let $\psi_1(t, t_0)$ and $\psi_2(t, t_0)$ denote the transition matrices corresponding to $-A^{-\prime}(t)$ and $-(A^-(t) - B(t)B'(t)\tilde{K}(t))'$. It follows from (26) that

$$\tilde{K}(t) = \psi_1(t, t_0)\tilde{K}(t_0)\psi_2'(t, t_0).$$

Thus, $\tilde{K}(t)$ is singular if and only if $\tilde{K}(t_0)$ is singular, proving the claim. An immediate consequence of this is that if $\tilde{K}_0$ is nonsingular, then $\tilde{K}(t, \tilde{K}_0, t_0)$ has the same signature as $\tilde{K}_0$, since no eigenvalues can pass through 0.

Suppose that $\tilde{K}(t, \tilde{K}_0, t_0)$ converges to $\Delta(t)$ as $t \to -\infty$. Let $C = \{\Delta(t): t \in [0, T]\}$ and let D denote the subset of $S(n)$ consisting of all those symmetric matrices which are positive definite. Let $S(n) - D$ denote the complement of D in $S(n)$. Let

$$\delta = \inf_{\substack{X \in C \\ Y \in S(n) - D}} \|X - Y\|.$$

Since C and $S(n) - D$ are disjoint subsets of $S(n)$ with C compact and $S(n) - D$ closed, it follows that $\delta > 0$. Since $\tilde{K}(t, \tilde{K}_0, t_0)$ converges to $\Delta(t)$ as $t \to -\infty$, there exists $t_1 \leqq t_0$ such that $\|\Delta(t_1) - \tilde{K}(t_1, \tilde{K}_0, t_0)\| < \delta$. Hence, $\tilde{K}(t_1, \tilde{K}_0, t_0) \in D$. Since $\tilde{K}(t_1, \tilde{K}_0, t_0)$ is nonsingular, so is $\tilde{K}_0$. Since $\tilde{K}(t_1, \tilde{K}_0, t_0)$ has the same signature as $\tilde{K}_0$, $\tilde{K}_0$ must be positive definite. Thus, a necessary condition for $\tilde{K}(t, \tilde{K}_0, t_0)$ to converge to $\Delta(t)$ as $t \to -\infty$ is that $\tilde{K}_0 > 0$.

Conversely, suppose that $\tilde{K}_0 > 0$. We show that $\tilde{K}(t, \tilde{K}_0, t_0)$ exists for all $t \leqq t_0$ and converges to $\Delta(t)$ as $t \to -\infty$. Consider the linear matrix differential equation

$$\dot{P}(t) = P(t)A^{-\prime}(t) + A^{-}(t)P(t) - B(t)B'(t). \tag{27}$$

Let $\psi_{A^-}(t, t_0)$ denote the transition matrix corresponding to $A^-(t)$. Let $P(t, P_0, t_0)$ denote the solution of (27) which goes through P_0 at $t = t_0$. Then

$$P(t, P_0, t_0) = \psi_{A^-}(t, t_0)P_0\psi'_{A^-}(t, t_0) - \int_{t_0}^{t} \psi_{A^-}(t, \sigma)B(\sigma)B'(\sigma)\psi'_{A^-}(t, \sigma)\, d\sigma. \tag{28}$$

Since $\tilde{K}_0^{-1}$ is positive definite, it follows from (28) that $P(t, \tilde{K}_0^{-1}, t_0)$ is positive definite for all $t \leqq t_0$. Thus, $[P(t, \tilde{K}_0^{-1}, t_0)]^{-1}$ exists for all $t \leqq t_0$, and it is straightforward to check that $[P(t, \tilde{K}_0^{-1}, t_0)]^{-1}$ satisfies (25). By uniqueness, we have

$$\tilde{K}(t, \tilde{K}_0, t_0) = [P(t, \tilde{K}_0^{-1}, t_0)]^{-1}, \tag{29}$$

which shows that the solution $\tilde{K}(t, \tilde{K}_0, t_0)$ exists for all $t \leqq t_0$.

Since $\Delta(t)$ is a solution of (25) which is nonsingular for all t, it follows that $\Delta^{-1}(t)$ is a solution of (27). From (28), we have

$$P(t, \tilde{K}_0^{-1}, t_0) - \Delta^{-1}(t) = \psi_{A^-}(t, t_0)[\tilde{K}_0^{-1} - \Delta^{-1}(t_0)]\psi'_{A^-}(t, t_0). \tag{30}$$

Since every characteristic multiplier of $A^-(\cdot)$ is outside the unit circle, there exist positive constants c_1, c_2 such that

$$\|\psi_{A^-}(t, \tau)\| \leqq c_1 \exp\{-c_2(\tau - t)\} \quad \forall t \leqq \tau.$$

Thus, if $t \leqq t_0$, we have

$$\|P(t, \tilde{K}_0^{-1}, t_0) - \Delta^{-1}(t)\| \leqq c_1^2 \exp\{-2c_2(t_0 - t)\}\|\tilde{K}_0^{-1} - \Delta^{-1}(t_0)\|. \tag{31}$$

From this, we obtain

$$\begin{aligned} \|\tilde{K}(t, \tilde{K}_0, t_0) - \Delta(t)\| &= \|\tilde{K}(t, \tilde{K}_0, t_0)[\Delta^{-1}(t) - P(t, \tilde{K}_0^{-1}, t_0)]\Delta(t)\| \\ &\leqq \|\tilde{K}(t, \tilde{K}_0, t_0)\|\,\|\Delta(t)\| c_1^2 \exp\{-2c_2(t_0 - t)\}\|\tilde{K}_0^{-1} - \Delta^{-1}(t_0)\|. \end{aligned} \tag{32}$$

From (32) it is clear that to prove that $\tilde{K}(t, \tilde{K}_0, t_0)$ converges to $\Delta(t)$ as $t \to -\infty$, it suffices to show that $\Delta(t)$ and $\tilde{K}(t, \tilde{K}_0, t_0)$ are bounded on $(-\infty, t_0]$. The boundedness of $\Delta(t)$ follows trivially from its periodicity. From (29), it is clear that to prove that $\tilde{K}(t, \tilde{K}_0, t_0)$ is bounded, it suffices to show that $P(t, \tilde{K}_0^{-1}, t_0)$ is bounded and $\det P(t, \tilde{K}_0^{-1}, t_0)$ is bounded away from 0. By (31), $P(t, \tilde{K}_0^{-1}, t_0)$ converges to the periodic motion $\Delta^{-1}(t)$ as $t \to -\infty$, so $P(t, \tilde{K}_0^{-1}, t_0)$ is bounded on $(-\infty, t_0]$.

It remains only to show that $\det P(t, \tilde{K}_0^{-1}, t_0)$ is bounded away from 0 on $(-\infty, t_0]$. Let $F = \{\Delta^{-1}(t): t \in [0, T]\}$. Let $\alpha = \min_{t \in [0,T]} \det \Delta^{-1}(t)$. Since $\Delta^{-1}(t)$ is positive definite for all t, $\alpha > 0$. Let $U = \{X \in S(n): \det X > \frac{1}{2}\alpha\}$. Then U is an open subset of $S(n)$ which contains F. Let

$$\delta_1 = \inf_{\substack{X \in F \\ Y \in S(n) - U}} \|X - Y\|.$$

Since the compact set F is disjoint from the closed set $S(n) - U$, it follows that $\delta_1 > 0$. Since $P(t, \tilde{K}_0^{-1}, t_0)$ converges to $\Delta^{-1}(t)$ as $t \to -\infty$, there exists $t_1 \leqq t_0$ such that $\|\Delta^{-1}(t) - P(t, \tilde{K}_0^{-1}, t_0)\| < \delta_1$, $\forall t < t_1$. Thus, $P(t, \tilde{K}_0^{-1}, t_0) \in U$, $\forall t < t_1$. Hence, $\det P(t, \tilde{K}_0^{-1}, t_0) > \frac{1}{2}\alpha$, $\forall t < t_1$. Since $P(t, \tilde{K}_0^{-1}, t_0) > 0$, $\forall t \leqq t_0$, there exists $\beta > 0$ such that $\det P(t, \tilde{K}_0^{-1}, t_0) \geqq \beta$, $\forall t \in [t_1, t_0]$. This shows that $\det P(t, \tilde{K}_0^{-1}, t_0) \geqq \min(\beta, \frac{1}{2}\alpha)$, $\forall t \leqq t_0$, proving that $\det P(t, \tilde{K}_0^{-1}, t_0)$ is bounded away from 0 on $(-\infty, t_0]$. This completes the proof of (b).

(c) The proof of (c) is analogous to the proof of (b). One starts by defining $\hat{K}(t) = K(t) - K^+(t)$ and proceeding parallel to (b). The details are left to the reader. □

The preceding theorem describes all those solutions which converge to $K^+(t)$ as $t \to -\infty$ and all those solutions which converge to $K^-(t)$ as $t \to \infty$. In § 7, we will describe all those solutions which converge to a given almost periodic solution as $t \to \pm\infty$. As a special case of this result, we will determine those solutions which converge to a given periodic equilibrium other than $K^+(t)$ or $K^-(t)$.

5. Finite escape times. In this section, we determine exactly which solutions of the PRDE escape in finite forward time or finite backward time. Note that from Theorem 8, a sufficient condition for $K(t, K_0, t_0)$ to have no finite escape in forward time is that $K_0 < K^+(t_0)$, and a sufficient condition for $K(t, K_0, t_0)$ to have no finite escape in backward time is that $K_0 > K^-(t_0)$.

The next lemma generalizes to periodic linear systems a well-known result for time-invariant systems. (See e.g. [14].)

LEMMA 11. *Suppose that $(A(\cdot), B(\cdot))$ is controllable and that every characteristic multiplier of $A(\cdot)$ is inside the unit circle. Then given any t_0 and any symmetric matrix P_0, there exists $t_f \geqq t_0$ such that*

$$W_A(t_0, t_f) > P_0.$$

Proof. Consider the periodic Lyapunov differential equation (PLDE)

$$\dot{P}(t) = A(t)P(t) + P(t)A'(t) - B(t)B'(t).$$

Let $P(t, P_0, t_0)$ denote the solution of the PLDE which goes through P_0 at $t = t_0$. Then

$$P(t, P_0, t_0) = \psi_A(t, t_0)P_0\psi_A'(t, t_0) - \int_{t_0}^{t} \psi_A(t, \sigma)B(\sigma)B'(\sigma)\psi_A'(t, \sigma)\, d\sigma.$$

Consider also the SHPRDE

$$\dot{K}(t) = -A'(t)K(t) - K(t)A(t) + K(t)B(t)B'(t)K(t).$$

We have $(A(\cdot), B(\cdot))$ controllable and $K^+(t) \equiv 0$. Since $\Delta(t) > 0$, it follows that $K^-(t) < 0$. Since $K^-(t)$ is nonsingular for all t, it is straightforward to check that $K^-(t)^{-1}$ satisfies the PLDE. Thus, $P(t, K^-(t_0)^{-1}, t_0) = K^-(t)^{-1}$. Since every characteristic multiplier of $A(\cdot)$ is inside the unit circle, there exist positive constants c_1, c_2 such that

$$\|\psi_A(t, t_0)\| \leqq c_1 \exp\{-c_2(t - t_0)\} \quad \forall t \geqq t_0.$$

Hence,

$$\begin{aligned} \|P(t, P_0, t_0) - K^-(t)^{-1}\| &= \|\psi_A(t, t_0)[P_0 - K^-(t_0)^{-1}]\psi_A'(t, t_0)\| \\ &\leqq c_1^2 \exp\{-2c_2(t - t_0)\}\, \|P_0 - K^-(t_0)^{-1}\| \end{aligned} \tag{33}$$

which goes to 0 as $t \to \infty$. Let $F = \{K^-(t)^{-1} : t \in [0, T]\}$, and let G denote the subset of $S(n)$ consisting of all those symmetric matrices which are negative definite. Since the compact set F is disjoint from the closed set $S(n) - G$, it follows from (33) that there exists some $t_f \geqq t_0$ such that $P(t, P_0, t_0) \in G$, $\forall t \geqq t_f$. Thus, $P(t, P_0, t_0) < 0$, $\forall t \geqq t_f$. From this we immediately obtain

$$0 > \psi_A(t_0, t)P(t, P_0, t_0)\psi_A'(t_0, t) = P_0 - W_A(t_0, t) \quad \forall t \geqq t_f.$$

We conclude that

$$W_A(t_0, t) > P_0 \quad \forall t \geqq t_f.$$

Since P_0 is arbitrary, the proof is complete. □

THEOREM 9. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and* $\Phi(T, 0)$ *has no eigenvalues on the unit circle. Then*

(a) $K(t, K_0, t_0)$ *has no finite escape time in forward time if and only if* $K^+(t_0) \geqq K_0$.

(b) $K(t, K_0, t_0)$ *has no finite escape time in backward time if and only if* $K_0 \geqq K^-(t_0)$.

Proof. (a) Transform the PRDE by defining $\hat{K}(t) = K(t) - K^+(t)$. Then $\hat{K}(t)$ satisfies the differential equation

$$\dot{\hat{K}}(t) = -A^{+\prime}(t)\hat{K}(t) - \hat{K}(t)A^+(t) + \hat{K}(t)B(t)B'(t)\hat{K}(t) \tag{34}$$

where $A^+(t) \equiv A(t) - B(t)B'(t)K^+(t)$. Let $\hat{K}(t, \hat{K}_0, t_0)$ denote the solution of (34) which goes through the point $\hat{K}_0$ at $t = t_0$. Since $K(t, K_0, t_0) - K^+(t)$ satisfies (34), uniqueness of solutions implies that $\hat{K}(t, K_0 - K^+(t_0), t_0) = K(t, K_0, t_0) - K^+(t)$. Since $K^+(t)$ exists for all t, it follows that $K(t, K_0, t_0)$ has a finite escape time in forward time if and only if $\hat{K}(t, K_0 - K^+(t_0), t_0)$ has a finite escape time in forward time. Since $K^+(t_0) \geqq K_0$ if and only if $K_0 - K^+(t_0) \leqq 0$, (a) is equivalent to the assertion that $\hat{K}(t, \hat{K}_0, t_0)$ has no finite escape in forward time if and only if $\hat{K}_0 \leqq 0$.

The Hamiltonian matrix which corresponds to the SHPRDE (34) is

$$\hat{H}(t) = \begin{bmatrix} A^+(t) & -B(t)B'(t) \\ 0 & -A^{+\prime}(t) \end{bmatrix},$$

and the associated transition matrix is

$$\hat{\Phi}(t, t_0) = \begin{bmatrix} \psi_{A^+}(t, t_0) & -\psi_{A^+}(t, t_0) W_{A^+}(t_0, t) \\ 0 & \psi'_{A^+}(t_0, t) \end{bmatrix} \tag{35}$$

where

$$W_{A^+}(t_0, t) = \int_{t_0}^{t} \psi_{A^+}(t_0, \sigma) B(\sigma) B'(\sigma) \psi'_{A^+}(t_0, \sigma)\, d\sigma.$$

$\hat{K}(t, \hat{K}_0, t_0)$ has a finite escape time $t_e > t_0$ if and only if $\hat{\Phi}(t_e, t_0)(\gamma(\hat{K}_0)) \notin \mathscr{L}_0(n)$ and $\hat{\Phi}(t, t_0)(\gamma(\hat{K}_0)) \in \mathscr{L}_0(n)$, $\forall t \in [t_0, t_e)$. Thus, $\hat{K}(t, \hat{K}_0, t_0)$ escapes in finite forward time if and only if there exists $t_e > t_0$ such that $\hat{\Phi}(t_e, t_0)(\gamma(\hat{K}_0))$ is not complementary to $\mathrm{Sp}\begin{bmatrix} 0 \\ I \end{bmatrix}$. From (35), we have

$$\hat{\Phi}(t_e, t_0)(\gamma(\hat{K}_0)) = \mathrm{Sp}\begin{bmatrix} \psi_{A^+}(t_e, t_0) - \psi_{A^+}(t_e, t_0) W_{A^+}(t_0, t_e)\hat{K}_0 \\ \psi'_{A^+}(t_0, t_e)\hat{K}_0 \end{bmatrix}. \tag{36}$$

Let $Z(t) \equiv I - W_{A^+}(t_0, t)\hat{K}_0$. By (36), $\hat{K}(t, \hat{K}_0, t_0)$ escapes in finite forward time if and only if there exists $t_e > t_0$ such that $\psi_{A^+}(t_e, t_0) - \psi_{A^+}(t_e, t_0) W_{A^+}(t_0, t_e)\hat{K}_0$ is singular, or equivalently such that $Z(t_e)$ is singular. Let $U(t)$ denote the symmetric matrix $\hat{K}_0 Z(t)$. It is straightforward to show that $Z(t)$ is nonsingular if and only if $\ker U(t) = \ker \hat{K}_0$.

Suppose that $\hat{K}_0 \leqq 0$. Showing that $\hat{K}(t, \hat{K}_0, t_0)$ has no finite escape in forward time is equivalent to showing that $\ker U(t) = \ker \hat{K}_0$, $\forall t \geqq t_0$. Trivially we have $\ker \hat{K}_0 \subseteq \ker U(t)$. Let $t \geqq t_0$ and suppose that $x \in \ker U(t)$. It follows that

$$x'\hat{K}_0 x - x'\hat{K}_0 W_{A^+}(t_0, t)\hat{K}_0 x = 0.$$

Since $\hat{K}_0 \leqq 0$ and $W_{A^+}(t_0, t) \geqq 0$, it follows that $x'\hat{K}_0 x = 0$, which then implies that $\hat{K}_0 x = 0$. Thus, $\ker U(t) = \ker \hat{K}_0$, $\forall t \geqq t_0$.

Conversely, suppose that $\hat{K}_0$ is not negative semidefinite. In other words, $\hat{K}_0$ has at least one positive eigenvalue. Let r denote the rank of $\hat{K}_0$. Choose $M \in 0(n)$ (the group of $n \times n$ orthogonal matrices) such that the matrix $\tilde{K}_0 \equiv M' \hat{K}_0 M$ has the form

$$\tilde{K}_0 = \begin{bmatrix} R_0 & 0 \\ 0 & 0 \end{bmatrix}$$

where R_0 is $r \times r$ symmetric, nonsingular, and has at least one positive eigenvalue. Let $\tilde{W}(t) \equiv M' W_{A^+}(t_0, t) M$, and partition $\tilde{W}(t)$ as

$$\tilde{W}(t) = \begin{bmatrix} W_{11}(t) & W_{12}(t) \\ W_{21}(t) & W_{22}(t) \end{bmatrix}$$

with $W_{11}(t) r \times r$. Let $\tilde{U}(t) \equiv M' U(t) M$. Then it is trivial to check that

$$\tilde{U}(t) = \tilde{K}_0 - \tilde{K}_0 \tilde{W}(t) \tilde{K}_0 = \begin{bmatrix} R_0 - R_0 W_{11}(t) R_0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{37}$$

To show that $\hat{K}(t, \hat{K}_0, t_0)$ has a finite escape in forward time is equivalent to showing that there exists $t_e > t_0$ such that $\ker \hat{K}_0 \subsetneq \ker U(t_e)$. In other words, we must show that there exists $t_e > t_0$ such that $\operatorname{rank} U(t_e) < r$, or equivalently, such that $\operatorname{rank} \tilde{U}(t_e) < r$. By (37), this is equivalent to showing that $R_0 - R_0 W_{11}(t_e) R_0$ is singular. Since $\tilde{W}(t_0) = 0$, $R_0 - R_0 W_{11}(t_0) R_0 = R_0$ and hence has at least one positive eigenvalue.

We claim that $R_0 - R_0 W_{11}(t) R_0$ eventually becomes negative definite. By Lemma 11, given any $n \times n$ symmetric matrix P, eventually $W_{A^+}(t_0, t) - P$ becomes positive definite. Equivalently, given any $n \times n$ symmetric matrix P, eventually $\tilde{W}(t) - P$ becomes positive definite. This implies that given any $r \times r$ symmetric matrix P_{11}, $W_{11}(t) - P_{11}$ eventually becomes positive definite. Since $R_0 - R_0 W_{11}(t) R_0 = R_0(R_0^{-1} - W_{11}(t)) R_0$, it follows that $R_0 - R_0 W_{11}(t) R_0$ eventually becomes negative definite, as claimed. Since $R_0 - R_0 W_{11}(t_0) R_0$ has at least one positive eigenvalue, there exists $t_e > t_0$ such that $R_0 - R_0 W_{11}(t_e) R_0$ is singular. Hence, $\hat{K}(t, \hat{K}_0, t_0)$ has a finite escape in forward time, which completes the proof of (a).

The proof of (b) is analogous to the proof of (a) and is left to the reader. □

An assertion which is equivalent to the conclusion of (b) appears in [7, Thm. 6]. However, no assumptions are specified and no proof is given. It is easy to show that the result is not true without the assumption of controllability.

6. Almost periodic solutions. In this section, we determine all of the almost periodic solutions of the PRDE. These solutions include as special cases the constant solutions, periodic equilibria (i.e. T-periodic solutions), and periodic solutions which are not T-periodic. The approach we take consists of 2 steps. Firstly, we determine every almost periodic solution of the EPRDE—i.e. of the Riccati differential equation extended to $\mathscr{L}(n)$. Secondly, we prove that every such solution is completely contained in the subset $\mathscr{L}_0(n)$ and hence corresponds (via the embedding $\gamma : S(n) \to \mathscr{L}(n)$) to an almost periodic solution of the PRDE on $S(n)$. Thus, we obtain a description of every almost periodic solution of the PRDE on $S(n)$.

The standard definition of a complex-valued almost periodic function [9] can be generalized to a function which takes values in a complete metric space. Let (X, ρ) be a complete metric space, and let $f : \mathbb{R} \to X$ be a continuous function. We say that f is almost periodic if and only if given any sequence of real numbers $\{\alpha_n\}_1^\infty$, there exists a subsequence $\{\alpha_{n_j}\}$ such that the sequence of translates $\{f(t + \alpha_{n_j})\}$ converges uniformly in t as $j \to \infty$. This generalizes the so-called Bochner definition of an almost periodic complex-valued function. We will need several basic properties of almost periodic

functions with values in either a complete metric space or in a Banach space. These are summarized in the Appendix.

The Grassmann manifold $G^n(\mathbb{R}^{2n})$ of all n-dimensional subspaces of $\mathbb{R}^{2n}$ can be given the structure of a metric space by defining on it the so-called "gap metric" θ. (See [10] for details concerning this metric.) If $S_1, S_2 \in G^n(\mathbb{R}^{2n})$ and if P_1, P_2 are the orthogonal projections onto S_1 and S_2 respectively, then $\theta(S_1, S_2) =^{\text{def}} \|P_1 - P_2\|$ (operator norm).

Remark 6. The gap metric is widely used by analysts. However, topologists define a topology on $G^n(\mathbb{R}^{2n})$ in a different way. Let $V_n(\mathbb{R}^{2n})$ denote the set of all $2n \times n$ full rank matrices with real entries. $V_n(\mathbb{R}^{2n})$ is an open subset of $\mathbb{R}^{2n \times n}$ and thus has the standard Euclidean topology. ($V_n(\mathbb{R}^{2n})$ is known as a Stiefel manifold.) Define a mapping $q: V_n(\mathbb{R}^{2n}) \to G^n(\mathbb{R}^{2n})$ with $q(Y)$ the column space of the matrix Y. Then $G^n(\mathbb{R}^{2n})$ is given the quotient topology induced by the surjective map q. In other words, a subset U of $G^n(\mathbb{R}^{2n})$ is open if and only if $q^{-1}(U)$ is open in $V_n(\mathbb{R}^{2n})$. It is not hard to show that the quotient topology is in fact the same as the topology induced by the gap metric.

Let $\mathrm{Sp}(n, \mathbb{R})$ and $\mathrm{Gl}(n, \mathbb{R})$ denote the symplectic group and the general linear group respectively.

LEMMA 12. *Let $M \in \mathrm{Sp}(n, \mathbb{R})$ be semisimple (diagonalizable) with no eigenvalues on the unit circle. Then there exists $P \in \mathrm{Sp}(n, \mathbb{R})$ such that $P^{-1}MP$ has the form $\left[\begin{smallmatrix} D & 0 \\ 0 & (D^{-1})' \end{smallmatrix}\right]$, where (1) D is a nonsingular semisimple matrix in real canonical form; (2) every eigenvalue of D has modulus less than 1; (3) the blocks on the main diagonal of D are ordered by increasing modulus of the corresponding eigenvalues.*

Proof. Since M is symplectic, semisimple, with no eigenvalues on the unit circle, the real canonical form of M can then be taken to be of the form $\left[\begin{smallmatrix} D & 0 \\ 0 & (D^{-1})' \end{smallmatrix}\right]$ with D as described. Thus, there exists $\tilde{P} \in \mathrm{Gl}(2n, \mathbb{R})$ such that

$$M\tilde{P} = \tilde{P}\begin{bmatrix} D & 0 \\ 0 & (D^{-1})' \end{bmatrix}.$$

Let $\tilde{P} \equiv [V, W]$ with V, W each $2n \times n$, and let $R \equiv V'JW$. Then $MV = VD$ and $MW = W(D^{-1})'$. Hence, $R = V'JW = V'M'JMW = D'V'JW(D^{-1})' = D'R(D^{-1})'$. Thus, D' commutes with R.

Since $\mathrm{Sp}\, V = L^+(M)$ and $\mathrm{Sp}\, W = L^-(M)$, it follows from Lemma 3 that $V'JV = 0$ and $W'JW = 0$. Consequently,

$$\tilde{P}'J\tilde{P} = \begin{bmatrix} 0 & R \\ -R' & 0 \end{bmatrix}$$

which shows that R is nonsingular. Define $P \equiv [V, WR^{-1}]$. Then $P'JP = J$, so $P \in \mathrm{Sp}(n, \mathbb{R})$. Using the fact that D' commutes with R, it is easy to show that

$$MP = P\begin{bmatrix} D & 0 \\ 0 & (D')^{-1} \end{bmatrix}. \qquad \square$$

LEMMA 13. *Let $\mathbb{R}^{2n} = V_1 \oplus \cdots \oplus V_{2r}$ be a direct sum decomposition of $\mathbb{R}^{2n}$ with the special property that $[J(V_j)]^{\perp} = \bigoplus_{i=1, i \neq 2r-j+1}^{2r} V_i, j = 1, \cdots, 2r$. Let $\lambda_1, \cdots, \lambda_{2r}$ be nonzero real numbers (not necessarily distinct) with the property that $\lambda_j^{-1} = \lambda_{2r-j+1}$, $j = 1, \cdots, r$. Define a linear transformation P on $\mathbb{R}^{2n}$ by $Px = \lambda_j x$, $\forall x \in V_j$, $j = 1, \cdots, 2r$. Then $P \in \mathrm{Sp}(n, \mathbb{R})$.*

Proof. Let $x \in V_j$, $y \in V_i$, and suppose that $i \neq 2r - j + 1$. Then $y'JPx = \lambda_j y'Jx = 0$. Also, $y'(P^{-1})'Jx = \lambda_i^{-1} y'Jx = 0$. Now suppose that $x \in V_j$ and $y \in V_{2r-j+1}$. Then $y'JPx =$

$\lambda_j y'Jx$, and $y'(P^{-1})'Jx = \lambda_{2r-j+1}^{-1} y'Jx = \lambda_j y'Jx$. Thus, $y'JPx = y'(P^{-1})'Jx$, $\forall x \in V_j$, $\forall y \in V_i$, $\forall i, j$. Hence, $JP = (P^{-1})'J$ which proves that $P \in \mathrm{Sp}(n, \mathbb{R})$. □

We are now prepared to begin to construct all of the almost periodic solutions of the EPRDE on $\mathscr{L}(n)$. We assume that $\Phi(T, 0)$ has no eigenvalues on the unit circle. Until otherwise specified, we make no additional assumptions concerning $\Phi(T, 0)$, $A(\cdot)$, $B(\cdot)$, and $C(\cdot)$.

Let $\mu_1 < \mu_2 < \cdots < \mu_{2r}$ denote the distinct moduli of the eigenvalues of $\Phi(T, 0)$. Since $\Phi(T, 0)$ is symplectic, it follows that $0 < \mu_i < 1$ for $i = 1, \cdots, r$ and $\mu_i > 1$ for $i = r+1, \cdots, 2r$. Also, $\mu_{2r-i+1} = \mu_i^{-1}$, $\forall i$. Let $E_i(t)$ denote the direct sum of the primary components of $\Phi(t+T, t)$ corresponding to those eigenvalues of $\Phi(t+T, t)$ of modulus μ_i $(i = 1, \cdots, 2r)$. Since $\Phi(t+T, t)$ is symplectic, $\dim E_i(t) = \dim E_{2r-i+1}(t)$, $\forall i$. From the similarity relation $\Phi(t+T, t) = \Phi(t, t_0)\Phi(t_0+T, t_0)\Phi(t_0, t)$ it follows that

$$E_i(t) = \Phi(t, t_0)(E_i(t_0)). \tag{38}$$

Since $E_i(t_0)$ is $\Phi(t_0+T, t_0)$-invariant, (38) implies that the subspace $E_i(t)$ is T-periodic. Note also that $L^+(\Phi(t+T, t)) = E_1(t) \oplus \cdots \oplus E_r(t)$ and $L^-(\Phi(t+T, t)) = E_{r+1}(t) \oplus \cdots \oplus E_{2r}(t)$.

Let $l = (l_1, \cdots, l_r)$ be an r-tuple of integers such that $0 \leqq l_j \leqq \dim E_j(t)$, $j = 1, \cdots, r$. (By (38), $\dim E_j(t)$ is constant.) For each such l and each t, define a collection of subspaces of $\mathbb{R}^{2n}$ by $U(l, t) \equiv \{\bigoplus_{j=1}^r [S_j \oplus ([J(S_j)]^\perp \cap E_{2r-j+1}(t))]\colon S_j \in G^{l_j}(E_j(t)),\ j = 1, \cdots, r\}$, where $G^{l_j}(E_j(t))$ denotes the Grassmann manifold consisting of all l_j-dimensional subspaces of $E_j(t)$. The next result shows that $U(l, t)$ is actually a subset of the Lagrange-Grassmann manifold $\mathscr{L}(n)$.

LEMMA 14. $U(l, t) \subset \mathscr{L}(n)$.

Proof. Let t be fixed. It follows from Lemma 3 that

$$[J(E_j(t))]^\perp = \bigoplus_{\substack{i=1 \\ i \neq 2r-j+1}}^{2r} E_i(t), \qquad j = 1, \cdots, 2r. \tag{39}$$

Define a linear transformation P on $\mathbb{R}^{2n}$ by $Px = \mu_j x$, $\forall x \in E_j(t)$, $j = 1, \cdots, 2r$. By Lemma 13, $P \in \mathrm{Sp}(n, \mathbb{R})$.

Let $S \in U(l, t)$. Then S can be expressed in the form $S = \bigoplus_{j=1}^r [S_j \oplus ([J(S_j)]^\perp \cap E_{2r-j+1}(t))]$, where $S_j \in G^{l_j}(E_j(t))$, $j = 1, \cdots, r$. Since $S_j \subseteq E_j(t)$, it follows from (39) that $J(S_j)$ is orthogonal to S, $j = 1, \cdots, r$. Since $J([J(S_j)]^\perp \cap E_{2r-j+1}(t))$ is contained in $J(E_{2r-j+1}(t))$, it follows from (39) that $J([J(S_j)]^\perp \cap E_{2r-j+1}(t))$ is orthogonal to $\bigoplus_{i=1, i\neq j}^{2r} E_i(t)$. From this and the fact that $J([J(S_j)]^\perp \cap E_{2r-j+1}(t)) = S_j^\perp \cap J(E_{2r-j+1}(t))$, it follows that $J([J(S_j)]^\perp \cap E_{2r-j+1}(t))$ is orthogonal to S, $j = 1, \cdots, r$. We conclude that $J(S)$ is orthogonal to S.

Since $J(S)$ is orthogonal to S, it remains only to show that S is n-dimensional. Since $\sum_{j=1}^r \dim E_j(t) = n$, it suffices to show that

$$\dim S_j + \dim([J(S_j)]^\perp \cap E_{2r-j+1}(t)) = \dim E_j(t),$$

$j = 1, \cdots, r$. From the definition of P, it follows that S_j is P-invariant. Since $P \in \mathrm{Sp}(n, \mathbb{R})$, it follows that $J(S_j)$ is P'-invariant, so $[J(S_j)]^\perp$ is P-invariant. Since $E_1(t), \cdots, E_{2r}(t)$ are the eigenspaces of P, this implies that

$$[J(S_j)]^\perp = \bigoplus_{i=1}^{2r} [J(S_j)]^\perp \cap E_i(t).$$

Since $S_j \subseteq E_j(t)$, it follows from (39) that $[J(S_j)]^\perp \cap E_i(t) = E_i(t)$ provided $i \neq 2r-j+1$.

Hence

$$[J(S_j)]^{\perp} = \left(([J(S_j)]^{\perp} \cap E_{2r-j+1}(t)) \oplus \left[\bigoplus_{\substack{i=1 \\ i \neq 2r-j+1}}^{2r} E_i(t) \right] \right).$$

Equating the dimensions of the two sides in this equation gives $2n - \dim S_j = \dim [J(S_j)]^{\perp} \cap E_{2r-j+1}(t) + (2n - \dim E_{2r-j+1}(t))$. Since $\dim E_j(t) = \dim E_{2r-j+1}(t)$, we obtain $\dim S_j + \dim [J(S_j)]^{\perp} \cap E_{2r-j+1}(t) = \dim E_j(t)$, which completes the proof. □

The next result describes the invariant property of the family $\{U(l, t): t \in \mathbb{R}\}$ (l fixed) with respect to the flow of the EPRDE.

LEMMA 15. $U(l, t) = \Phi(t, t_0)(U(l, t_0))$.

Proof. Let $S \in U(l, t_0)$. Then S can be expressed in the form $S = \bigoplus_{j=1}^{r} [S_j \oplus ([J(S_j)]^{\perp} \cap E_{2r-j+1}(t_0))]$, where $S_j \in G^{l_j}(E_j(t_0))$, $j = 1, \cdots, r$. Thus,

$$\Phi(t, t_0)(S) = \bigoplus_{j=1}^{r} [\Phi(t, t_0)(S_j) \oplus \Phi(t, t_0)([J(S_j)]^{\perp} \cap E_{2r-j+1}(t_0))].$$

By (38), we have $\Phi(t, t_0)(S_j) \in G^{l_j}(E_j(t))$, $j = 1, \cdots, r$. Also,

$$\begin{aligned} \Phi(t, t_0)([J(S_j)]^{\perp} \cap E_{2r-j+1}(t_0)) &= \Phi(t, t_0)([J(S_j)]^{\perp}) \cap E_{2r-j+1}(t) \\ &= [J(\Phi(t, t_0)(S_j))]^{\perp} \cap E_{2r-j+1}(t), \end{aligned}$$

where the last equality is easily proven using the fact that $\Phi(t, t_0)$ is symplectic. Thus,

$$\Phi(t, t_0)(S) = \bigoplus_{j=1}^{r} [\Phi(t, t_0)(S_j) \oplus ([J(\Phi(t, t_0)(S_j))]^{\perp} \cap E_{2r-j+1}(t))],$$

with $\Phi(t, t_0)(S_j) \in G^{l_j}(E_j(t))$, $j = 1, \cdots, r$. Hence, $\Phi(t, t_0)(S) \in U(l, t)$, which shows that $\Phi(t, t_0)(U(l, t_0)) \subseteq U(l, t)$. Reversing the roles of t_0 and t gives $\Phi(t_0, t)(U(l, t)) \subseteq (U(l, t_0))$, or $U(l, t) \subseteq \Phi(t, t_0)(U(l, t_0))$. Thus, $\Phi(t, t_0)(U(l, t_0)) = U(l, t)$. □

COROLLARY. $\Phi(t + T, t)(U(l, t)) = U(l, t)$.

Proof. From the definition of $U(l, t)$ and the T-periodicity of the subspace $E_j(t)$, it follows that $U(l, t) = U(l, t + T)$, so the assertion is an immediate consequence of Lemma 15. □

Remark 7. Let $n_j \equiv \dim E_j(t)$, $j = 1, \cdots, r$. (As previously noted, n_j does not depend on t.) It is clear from its definition that $U(l, t)$ is an embedded submanifold of $\mathscr{L}(n)$ which is analytically isomorphic to the product of Grassmann manifolds $G^{l_1}(\mathbb{R}^{n_1}) \times \cdots \times G^{l_r}(\mathbb{R}^{n_r})$. From the well-known fact that $\dim G^k(\mathbb{R}^m) = k(m - k)$, it follows that $\dim U(l, t) = \sum_{j=1}^{r} l_j(n_j - l_j)$. Since $U(l, t + T) = U(l, t)$, we see that $U(l, t)$ is a product of Grassmann manifolds which oscillates with period T.

There is an important special case worth noting. Suppose that in addition to having no eigenvalues on the unit circle, $\Phi(T, 0)$ satisfies the condition that every eigenvalue is distinct and if two eigenvalues have the same modulus, then they are a pair of complex conjugates. It is easily seen that this additional condition is generic in the space of symplectic matrices with no eigenvalues on the unit circle. This condition implies that n_j is equal to either 1 or 2. It is equal to 1 if $\Phi(T, 0)$ has a real eigenvalue of modulus μ_j, and is equal to 2 if $\Phi(T, 0)$ has a pair of complex conjugate eigenvalues of modulus μ_j. Since $G^0(\mathbb{R}^m)$ and $G^m(\mathbb{R}^m)$ are single points for any m, and $G^1(\mathbb{R}^2)$ is the projective line which is topologically the circle, it follows that $U(l, t)$ is a torus (i.e. a product of circles) of dimension equal to the cardinality of the set $\{j: 1 \leqq j \leqq r, n_j = 2, l_j = 1\}$.

We are now prepared to describe *all* of the almost periodic solutions of the EPRDE.

THEOREM 10. *Suppose that* $\Phi(T, 0)$ *has no eigenvalues on the unit circle and is semisimple.*

(a) *If* $S_0 \in U(l, t_0)$, *then* $S(t, S_0, t_0)$ *is an almost periodic solution of the* EPRDE, *and* $S(t, S_0, t_0) \in U(l, t)$, $\forall t$.

(b) *No other solutions of the* EPRDE *are almost periodic.*

Proof. If $S_0 \in U(l, t_0)$, then $S(t, S_0, t_0) = \Phi(t, t_0)(S_0) \in U(l, t)$ by Lemma 15. By Lemma 12, there exists $P \in \mathrm{Sp}(n, \mathbb{R})$ such that

$$\Lambda \equiv P^{-1}\Phi(T, 0)P = \begin{bmatrix} D & 0 \\ 0 & (D^{-1})' \end{bmatrix}$$

where (1) D is nonsingular semisimple in real canonical form; (2) every eigenvalue of D has modulus less than 1; (3) the blocks on the main diagonal of D are ordered by increasing modulus of the corresponding eigenvalues. Then D is of the form $D = \mathrm{diag}\{\mu_1 D_1, \cdots, \mu_r D_r\}$ where D_j is $\dim E_j(0) \times \dim E_j(0)$ and is zero except for 1×1 and/or 2×2 blocks on the main diagonal. Each 1×1 block is either $[1]$ or $[-1]$, while each 2×2 block is of the form

$$\begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \quad \text{for some } \alpha \in (0, \pi).$$

We now construct a particular Floquet representation for the transition matrix $\Phi(t, t_0)$. It is a standard result that there exists a *real* matrix R such that $\Lambda^2 = e^{2RT}$. In fact, we can define R as follows: Let

$$R \equiv \begin{bmatrix} W & 0 \\ 0 & -W' \end{bmatrix}$$

with $W = \mathrm{diag}\{W_1, \cdots, W_r\}$ where W_j is obtained from D_j by (1) replacing each 1×1 block on the main diagonal of D_j (whether $[1]$ or $[-1]$) with the 1×1 block $[(1/T) \ln \mu_j]$, and (2) replacing each 2×2 block of the form

$$\begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix}$$

on the main diagonal of D_j with the 2×2 block

$$\begin{bmatrix} \frac{1}{T}\ln\mu_j & \frac{1}{T}\alpha \\ -\frac{1}{T}\alpha & \frac{1}{T}\ln\mu_j \end{bmatrix}.$$

Then $e^{2W_j T} = \mu_j^2 D_j^2$, so $e^{2WT} = D^2$. This implies that $e^{2RT} = \Lambda^2$ as required. Let $V(t) \equiv P^{-1}\Phi(t, 0)P e^{-Rt}$. Then it is straightforward to show that $V(t+2T) = V(t)$ and

$$\Phi(t, t_0) = PV(t)\, e^{R(t-t_0)} V^{-1}(t_0) P^{-1}. \tag{40}$$

Next, we modify R to obtain a new matrix as follows: Let $\hat{W} \equiv \mathrm{diag}\{\hat{W}_1, \cdots, \hat{W}_r\}$ where $\hat{W}_j \equiv W_j - (1/T) \ln \mu_j I$. Define

$$\hat{R} \equiv \begin{bmatrix} \hat{W} & 0 \\ 0 & -\hat{W}' \end{bmatrix}.$$

Since $\hat{W}_j$ (and hence $\hat{W}$) is skew-symmetric, we have

$$\hat{R} = \begin{bmatrix} \hat{W} & 0 \\ 0 & \hat{W} \end{bmatrix}.$$

Let $\hat{\Phi}(t, t_0) \equiv PV(t)\, e^{\hat{R}(t-t_0)} V^{-1}(t_0) P^{-1}$. We claim that if $S_0 \in U(l, t_0)$, then

$$\Phi(t, t_0)(S_0) = \hat{\Phi}(t, t_0)(S_0) \quad \forall t. \tag{41}$$

By the definition of $U(l, t_0)$, S_0 is of the form

$$S_0 = \bigoplus_{j=1}^{r} [S_j \oplus ([J(S_j)]^{\perp} \cap E_{2r-j+1}(t_0))] \tag{42}$$

where $S_j \in G^{l_j}(E_j(t_0))$, $j = 1, \cdots, r$. Consequently, to establish (41), it suffices to show that if M_j is any subspace of $E_j(t_0)$, then $\Phi(t, t_0)(M_j) = \hat{\Phi}(t, t_0)(M_j)$, $j = 1, \cdots, 2r$.

Now, $E_j(t_0)$ is the sum of the primary components of $\Phi(t_0 + T, t_0)$ which correspond to eigenvalues of modulus μ_j. Then $P^{-1}\Phi(0, t_0)(E_j(t_0))$ is the sum of the primary components of $P^{-1}\Phi(0, t_0)\Phi(t_0 + T, t_0)\Phi(t_0, 0)P = P^{-1}\Phi(T, 0)P = \Lambda$ which correspond to eigenvalues of modulus μ_j. From the definition of $V(t)$, it follows that $e^{-Rt_0} V^{-1}(t_0) P^{-1} = P^{-1}\Phi(0, t_0)$. Thus, $e^{-Rt_0} V^{-1}(t_0) P^{-1}(E_j(t_0))$ is the sum of the primary components of Λ which correspond to eigenvalues of modulus μ_j. It follows that $e^{-Rt_0} V^{-1}(t_0) P^{-1}(E_j(t_0))$ is the sum of the primary components of Λ^2 which correspond to eigenvalues of modulus μ_j^2. Let F_j denote this subspace. From the structure of Λ^2, R, and $\hat{R}$, it is apparent that if $x \in F_j$, then

$$e^{\hat{R}\tau} x = \mu_j^{-\tau/T}\, e^{R\tau} x, \qquad j = 1, \cdots, 2r. \tag{43}$$

Thus, $e^{\hat{R}\tau} | F_j$ is a nonzero scalar multiple of $e^{R\tau} | F_j$. This is the key observation.

If M_j is a subspace of $E_j(t_0)$, then $e^{-Rt_0} V^{-1}(t_0) P^{-1}(M_j)$ is a subspace of F_j, so for any τ it follows from (43) that

$$e^{\hat{R}\tau}(e^{-Rt_0} V^{-1}(t_0) P^{-1}(M_j)) = e^{R\tau}(e^{-Rt_0} V^{-1}(t_0) P^{-1}(M_j)). \tag{44}$$

Using (44), we have

$$\begin{aligned} e^{\hat{R}(t-t_0)} V^{-1}(t_0) P^{-1}(M_j) &= e^{\hat{R}(t-t_0)}\, e^{Rt_0}(e^{-Rt_0} V^{-1}(t_0) P^{-1}(M_j)) \\ &= e^{\hat{R}(t-t_0)}\, e^{\hat{R}t_0}(e^{-Rt_0} V^{-1}(t_0) P^{-1}(M_j)) \\ &= e^{\hat{R}t}(e^{-Rt_0} V^{-1}(t_0) P^{-1}(M_j)) \\ &= e^{R(t-t_0)} V^{-1}(t_0) P^{-1}(M_j). \end{aligned}$$

Using (40) and the definition of $\hat{\Phi}(t, t_0)$, it follows that $\hat{\Phi}(t, t_0)(M_j) = \Phi(t, t_0)(M_j)$, which establishes the claim (41).

Let $V_n^0(\mathbb{R}^{2n})$ denote the subset of $V_n(\mathbb{R}^{2n})$ which consists of all $2n \times n$ matrices with orthonormal columns. It is easily seen that $V_n^0(\mathbb{R}^{2n})$ is compact. From (41), we have for $S_0 \in U(l, t_0)$

$$S(t, S_0, t_0) = PV(t)\, e^{\hat{R}(t-t_0)} V^{-1}(t_0) P^{-1}(S_0). \tag{45}$$

Let X_0 be a $2n \times n$ matrix whose columns form an orthonormal basis for $V^{-1}(t_0) P^{-1}(S_0)$. In other words, $X_0 \in V_n^0(\mathbb{R}^{2n})$ and $q(X_0) = V^{-1}(t_0) P^{-1}(S_0)$. (See Remark 6 for the definition of q.) Since $\hat{R}$ is skew-symmetric, $e^{\hat{R}(t-t_0)} \in O(2n)$. Thus, $e^{\hat{R}(t-t_0)} X_0 \in V_n^0(\mathbb{R}^{2n})$, $\forall t$. Since $V(t)$ is periodic with period $2T$, we have $\{PV(t)\, e^{\hat{R}(t-t_0)} X_0 \colon t \in \mathbb{R}\} \subset \{PV(t)X \colon X \in V_n^0(\mathbb{R}^{2n}),\ t \in [0, 2T]\} \equiv \Omega$ which is compact since it is the image of $[0, 2T] \times V_n^0(\mathbb{R}^{2n})$ under the continuous mapping $(t, X) \mapsto PV(t)X$. Since each entry of

$e^{\hat{R}(t-t_0)}$ is periodic, it follows from AP 3 in the Appendix that $e^{\hat{R}(t-t_0)}$ is almost periodic. Since $V(t)$ is periodic, AP 4 implies that the product $PV(t)\, e^{\hat{R}(t-t_0)}X_0$ is almost periodic. The quotient map $q: V_n(\mathbb{R}^{2n}) \to G^n(\mathbb{R}^{2n})$ is continuous, so its restriction to the compact set Ω is uniformly continuous. Since $G^n(\mathbb{R}^{2n})$ is a compact (and hence complete) metric space [10], and $PV(t)\, e^{\hat{R}(t-t_0)}X_0 \in \Omega$ for all t, it follows from AP 5 that $q(PV(t)\, e^{\hat{R}(t-t_0)}X_0)$ is almost periodic. Since $S(t, S_0, t_0) = q(PV(t)\, e^{\hat{R}(t-t_0)}X_0)$, this proves that $S(t, S_0, t_0)$ is almost periodic.

The proof of (a) is now complete. The proof of (b) is deferred to Remark 15 in the next section. □

Remark 8. Theorem 10 shows that there is a one-to-one correspondence between the set of almost periodic solutions of the EPRDE and the disjoint union $\cup_l U(l, 0)$, which is a union of submanifolds of $\mathscr{L}(n)$ each of which is topologically a product of Grassmann manifolds.

It is important to keep in mind that Theorem 10 characterizes every almost periodic solution of the EPRDE—i.e. of the *extended* PRDE on $\mathscr{L}(n)$. In general, there is not a one-to-one correspondence between the set of almost periodic solutions of the PRDE (on the space $S(n)$ of $n \times n$ symmetric matrices) and the set of almost periodic solutions of the EPRDE. The problem is that an almost periodic solution of the EPRDE may intersect the subset $\mathscr{L}(n) - \mathscr{L}_0(n)$ of $\mathscr{L}(n)$, or may be completely contained in this subset. In either case, there is no corresponding almost periodic solution of the PRDE on $S(n)$.

The next lemma is an absolutely crucial result, and is rather surprising. It implies that if $(A(\cdot), B(\cdot))$ is controllable, then every almost periodic solution of the EPRDE is completely contained in $\mathscr{L}_0(n)$ and hence corresponds to an almost periodic solution of the PRDE on $S(n)$. Thus, in the presence of controllability, the almost periodic solutions of the PRDE are in one-to-one correspondence with the almost periodic solutions of the EPRDE (via the embedding $\gamma: S(n) \to \mathscr{L}(n)$), and can be completely described using Theorem 10.

LEMMA 16. *Suppose that* $\Phi(T, 0)$ *has no eigenvalues on the unit circle and* $(A(\cdot), B(\cdot))$ *is controllable. Let* $S \in \mathscr{L}(n)$ *and suppose that* S *is of the form* $S = S^+ \oplus S^-$ *where* $S^+ \subseteq L^+(\Phi(t+T, t))$ *and* $S^- \subseteq L^-(\Phi(t+T, t))$. *Then* $S \in \mathscr{L}_0(n)$.

Proof. From the definition of $K^+(t)$ and $K^-(t)$, we have

$$L^+(\Phi(t+T, t)) = \mathrm{Sp}\begin{bmatrix} I \\ K^+(t) \end{bmatrix} \quad \text{and} \quad L^-(\Phi(t+T, t)) = \mathrm{Sp}\begin{bmatrix} I \\ K^-(t) \end{bmatrix}.$$

Let $k \equiv \dim S^+$. Then there exist $n \times k$ and $n \times (n-k)$ full rank matrices D^+, D^- such that

$$S^+ = \mathrm{Sp}\begin{bmatrix} D^+ \\ K^+(t)D^+ \end{bmatrix} \quad \text{and} \quad S^- = \mathrm{Sp}\begin{bmatrix} D^- \\ K^-(t)D^- \end{bmatrix}.$$

Since $S \in \mathscr{L}(n)$, we must have $J(S^+) \perp S^-$, which implies that $(D^-)'\Delta(t)D^+ = 0$. Now,

$$S = \mathrm{Sp}\begin{bmatrix} D^+ & D^- \\ K^+(t)D^+ & K^-(t)D^- \end{bmatrix},$$

so $S \in \mathscr{L}_0(n)$ iff the $n \times n$ matrix $[D^+ \;\; D^-]$ is nonsingular. Suppose $\exists y \in \mathbb{R}^k$ and $z \in \mathbb{R}^{n-k}$ such that

$$[D^+ \quad D^-]\begin{bmatrix} y \\ z \end{bmatrix} = 0.$$

Then $D^+y + D^-z = 0$, so $D^+y = -D^-z$. Premultiplying both sides by $z'(D^-)'\Delta(t)$ gives $0 = z'(D^-)'\Delta(t)D^+y = -z'(D^-)'\Delta(t)D^-z$. Since $\Delta(t) > 0$, this implies that $D^-z = 0$ and hence $D^+y = 0$ as well. Since D^+ and D^- each have full rank, $y = 0$ and $z = 0$. Thus, $[D^+ \ D^-]$ is nonsingular, which completes the proof. □

COROLLARY. *Suppose that* $\Phi(T, 0)$ *has no eigenvalues on the unit circle and* $(A(\cdot), B(\cdot))$ *is controllable. Then* $U(l, t) \subset \mathscr{L}_0(n)$, $\forall l$, $\forall t$.

Proof. By Lemma 14, $U(l, t) \subset \mathscr{L}(n)$, and it follows from the definition of $U(l, t)$ that each $S \in U(l, t)$ is of the form $S = S^+ \oplus S^-$ with $S^+ \subseteq L^+(\Phi(t+T, t))$ and $S^- \subseteq L^-(\Phi(t+T, t))$. Thus, the assertion is an immediate consequence of Lemma 16. □

The following result describes every almost periodic solution of the PRDE.

THEOREM 11. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and that* $\Phi(T, 0)$ *is semisimple and has no eigenvalues on the unit circle.* (a) *Each* $S_0 \in \sqcup_l U(l, t_0)$ *determines a unique almost periodic solution of the* PRDE *which is given by* $\gamma^{-1}(\Phi(t, t_0)(S_0)) \equiv K(t, \gamma^{-1}(S_0), t_0)$. (b) *Furthermore, every almost periodic solution is obtained in this way.*

Proof. Let $S_0 \in U(l, t_0)$. By Theorem 10, $S(t, S_0, t_0) \equiv \Phi(t, t_0)(S_0)$ is an almost periodic solution of the EPRDE. Furthermore, $S(t, S_0, t_0) \in U(l, t)$, $\forall t$. Define a mapping $\alpha : [0, T] \times U(l, 0) \to \mathscr{L}(n)$ by $\alpha(t, S) \equiv \Phi(t, 0)(S)$. By Lemma 15, the image of α is $\cup_{t \in \mathbb{R}} U(l, t)$. Since α is a continuous mapping and $[0, T] \times U(l, 0)$ is compact, it follows that the image of α is compact. By the Corollary to Lemma 16, $\cup_{t \in \mathbb{R}} U(l, t) \subseteq \mathscr{L}_0(n)$. The mapping $\gamma^{-1} : \mathscr{L}_0(n) \to S(n)$ is real-analytic. Restricted to the compact subset $\cup_{t \in \mathbb{R}} U(l, t)$, γ^{-1} is uniformly continuous. Thus, it follows from AP 5 that $\gamma^{-1}(S(t, S_0, t_0))$ is almost periodic. The proof that every almost periodic solution of the PRDE is of this form is deferred to Remark 16 in § 7. □

Remark 9. It follows from Theorem 11 that (under the stated hypotheses) the almost periodic solutions of the PRDE are in one-to-one correspondence with the points of $\sqcup_l U(l, 0)$, which is a disjoint union of products of Grassmann manifolds. $S_0 \in \sqcup_l U(l, 0)$ corresponds to the almost periodic solution $\gamma^{-1}(\Phi(t, 0)(S_0))$. Comparing this result with Theorem 3 shows that the periodic equilibria (T-periodic solutions) of the PRDE are precisely those almost periodic solutions $\gamma^{-1}(\Phi(t, 0)(S_0))$ for which the subspace S_0 is $\Phi(T, 0)$-invariant.

Remark 10. In theory, Theorem 11 can be used to compute any almost periodic solution of the PRDE. Given a subspace $S_0 \in U(l, 0)$, choose any basis matrix $\begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}$ for S_0 with X_0, Y_0 each $n \times n$. Let

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \Phi(t, 0) \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}.$$

Then the almost periodic solution $\gamma^{-1}(\Phi(t, 0)(S_0))$ is given by $Y(t)X^{-1}(t)$ where the matrix inverse exists for all t.

In the case where $(C(\cdot), A(\cdot))$ is observable, we can determine the signature of every almost periodic solution.

THEOREM 12. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable,* $(C(\cdot), A(\cdot))$ *is observable, and* $\Phi(T, 0)$ *is semisimple. Let* $S_0 \in U(l, t_0)$, *and let* $K(t) \equiv \gamma^{-1}(\Phi(t, t_0)(S_0))$ *be the corresponding almost periodic solution of the* PRDE. *Then* $K(t)$ *is nonsingular and has exactly* $\sum_{j=1}^{r} l_j$ *positive eigenvalues.*

Proof. It follows from [13] that the controllability of $(A(\cdot), B(\cdot))$ and observability of $(C(\cdot), A(\cdot))$ imply that $\Phi(T, 0)$ has no eigenvalues on the unit circle. Let $S(t) \equiv \Phi(t, t_0)(S_0)$. Since $S_0 \in U(l, t_0)$, we know from Theorem 10 that $S(t) \in U(l, t)$, $\forall t$. Let t be fixed. From the definition of $U(l, t)$ it follows that $S(t)$ is expressible as $S(t) = S^+ \oplus S^-$ with $S^+ \subseteq L^+(\Phi(t+T, t))$ and $S^- \subseteq L^-(\Phi(t+T, t))$. Consequently, if k denotes the

dimension of S^+, there exist $n \times k$ and $n \times (n-k)$ full rank matrices D^+, D^- such that

$$S^+ = \operatorname{Sp}\begin{bmatrix} D^+ \\ K^+(t)D^+ \end{bmatrix} \quad \text{and} \quad S^- = \operatorname{Sp}\begin{bmatrix} D^- \\ K^-(t)D^- \end{bmatrix}.$$

The condition that $J(S)^\perp = S$ implies that $(D^-)'\Delta(t)D^+ = 0$, so $\operatorname{Sp} D^- = \Delta^{-1}(t)(\operatorname{Sp} D^+)^\perp$.

Now,

$$S(t) = \operatorname{Sp}\begin{bmatrix} D^+ & D^- \\ K^+(t)D^+ & K^-(t)D^- \end{bmatrix}.$$

By Lemma 16, $S(t) \in \mathscr{L}_0(n)$, so the $n \times n$ matrix $[D^+ \ D^-]$ is nonsingular. Equivalently, $\operatorname{Sp} D^+$ and $\Delta^{-1}(t)(\operatorname{Sp} D^+)^\perp$ are complementary subspaces of $\mathbb{R}^n$. Let P be the projection onto $\operatorname{Sp} D^+$ along $\Delta^{-1}(t)(\operatorname{Sp} D^+)^\perp$. It is easily shown that

$$S(t) = \operatorname{Sp}\begin{bmatrix} I \\ K^+(t)P + K^-(t)(I-P) \end{bmatrix}$$

which implies that

$$K(t) = K^+(t)P + K^-(t)(I-P). \tag{46}$$

By the Corollary to Theorem 7, $K^+(t) > 0$ and $K^-(t) < 0$. Consequently, if $x \in \operatorname{Sp} D^+$ with $x \neq 0$, then $x'K(t)x = x'K^+(t)x > 0$, while if $x \in \Delta^{-1}(t)(\operatorname{Sp} D^+)^\perp$ with $x \neq 0$, then $x'K(t)x = x'K^-(t)x < 0$. It follows from a standard linear algebra result that $K(t)$ is nonsingular and the number of positive eigenvalues of $K(t)$ is equal to $\dim \operatorname{Sp} D^+$. Since $\dim \operatorname{Sp} D^+ = \dim S^+ = \sum_{j=1}^r l_j$, the proof is complete. □

7. Convergence to almost periodic solutions. In the preceding section, we gave a complete description of the almost periodic solutions of the PRDE (Theorem 11). In this section, we will determine the asymptotic behavior of *every* solution of the PRDE. In particular, we will show that under mild assumptions, a solution of the PRDE either escapes in finite forward (backward) time or converges to an almost periodic solution as $t \to \infty$ ($t \to -\infty$). Our results describe exactly which almost periodic solution a given solution approaches. We will obtain these results by first deriving the corresponding results for the EPRDE—i.e. for the extended differential equation on the Lagrange-Grassmann manifold.

We assume that $\Phi(T, 0)$ has no eigenvalues on the unit circle. Unless otherwise specified, we make no additional assumptions concerning $\Phi(T, 0)$, $A(\cdot)$, $B(\cdot)$, and $C(\cdot)$. For each t, define a flag (i.e. increasing sequence) of subspaces $0 = M_0(t) \subset M_1(t) \subset \cdots \subset M_{2r-1}(t) \subset M_{2r}(t) = \mathbb{R}^{2n}$ by

$$M_j(t) \equiv \bigoplus_{i=1}^{j} E_i(t), \qquad j = 0, \cdots, 2r.$$

From (38) it follows that $M_j(t)$ is T-periodic. We call the T-periodic flag $\{M_j(t)\}_{j=0}^{2r}$ the *stable flag* associated with the PRDE. Define a second T-periodic flag $\{N_j(t)\}_{j=0}^{2r}$ by

$$N_j(t) \equiv \bigoplus_{i=1}^{j} E_{2r-i+1}(t), \qquad j = 0, \cdots, 2r.$$

We call $\{N_j(t)\}_{j=0}^{2r}$ the *unstable flag* associated with the PRDE. As will be shown below, the stable (unstable) flag plays a critical role in the description of the asymptotic behavior of solutions of the PRDE as $t \to \infty$ ($t \to -\infty$).

Let $\eta_j(t)$ denote the linear projection onto $E_j(t)$ along $\bigoplus_{i=1,i\neq j}^{2r} E_i(t)$. For each $S\in\mathscr{L}(n)$, define

$$\Pi^+(t)(S)\equiv\bigoplus_{j=1}^{2r}\eta_j(t)(S\cap M_j(t))$$

and

$$\Pi^-(t)(S)\equiv\bigoplus_{j=1}^{2r}\eta_{2r-j+1}(t)(S\cap N_j(t)).$$

The next result shows that $\Pi^+(t)$ and $\Pi^-(t)$ map $\mathscr{L}(n)$ into itself. In fact, they each map $\mathscr{L}(n)$ onto the union of the Grassmannians $\{U(l,t)\}_l$. (t is fixed here.)

LEMMA 17. *Let $S\in\mathscr{L}(n)$ and let t be fixed. Let $l_j\equiv\dim S\cap M_j(t)-\dim S\cap M_{j-1}(t)$, $j=1,\cdots,2r$, let $l'_j\equiv\dim S\cap N_{2r-j+1}(t)-\dim S\cap N_{2r-j}(t)$, $j=1,\cdots,2r$, and let $l\equiv(l_1,\cdots,l_r)$ and $l'\equiv(l'_1,\cdots,l'_r)$. Then $\Pi^+(t)(S)\in U(l,t)$ and $\Pi^-(t)(S)\in U(l',t)$.*

Proof. Construct a basis for S by starting with a basis for $S\cap M_1(t)$ and extending successively to bases for $S\cap M_2(t),\cdots,S\cap M_{2r}(t)=S$. Since $M_j(t)=M_{j-1}(t)\oplus E_j(t)$, this basis can be expressed in the form $\{v_{ij}+w_{ij}: j=1,\cdots,l_i;\ i=1,\cdots,2r\}$ where $v_{ij}\in E_i(t)$, $w_{ij}\in M_{i-1}(t)$, and the set $\{v_{ij}: j=1,\cdots,l_i;\ i=1,\cdots,2r\}$ is linearly independent. Then

$$\Pi^+(t)(S)=\operatorname{Sp}\{v_{ij}: j=1,\cdots,l_i;\ i=1,\cdots,2r\}.$$

Note that this shows that $\dim\Pi^+(t)(S)=l_1+\cdots+l_{2r}=\dim S=n$.

Let $S_i\equiv\operatorname{Sp}\{v_{ij}: j=1,\cdots,l_i\}$, $i=1,\cdots,2r$. Then $S_i\in G^{l_i}(E_i(t))$, $i=1,\cdots,2r$, and $\Pi^+(t)(S)=\bigoplus_{i=1}^{2r}S_i$. To show that $\Pi^+(t)(S)\in U(l,t)$, it suffices to show that $S_{2r-i+1}=[J(S_i)]^\perp\cap E_{2r-i+1}(t)$, $i=1,\cdots,r$. In fact since $\Pi^+(t)(S)$ and $\bigoplus_{i=1}^{r}S_i\oplus([J(S_i)]^\perp\cap E_{2r-i+1}(t))$ are each n-dimensional, it suffices to show that $S_{2r-i+1}\subseteq[J(S_i)]^\perp\cap E_{2r-i+1}(t)$, $i=1,\cdots,r$. Since $S_{2r-i+1}\subseteq E_{2r-i+1}(t)$, it remains only to show that $S_{2r-i+1}\perp J(S_i)$, $i=1,\cdots,r$. Since $S\in\mathscr{L}(n)$, $J(S)\perp S$. Consequently, given any j and k with $1\leqq j\leqq l_i$ and $1\leqq k\leqq l_{2r-i+1}$, we have $0=(v_{ij}+w_{ij})'J(v_{2r-i+1,k}+w_{2r-i+1,k})$. Since $J(E_q(t))\perp E_p(t)$ provided $p\neq 2r-q+1$, it follows that $w'_{ij}Jv_{2r-i+1,k}=0$, $w'_{ij}Jw_{2r-i+1,k}=0$, and $v'_{ij}Jw_{2r-i+1,k}=0$. Thus, $0=v'_{ij}Jv_{2r-i+1,k}$, which shows that $S_{2r-i+1}\perp J(S_i)$, as required. This completes the proof that $\Pi^+(t)(S)\in U(l,t)$.

The proof that $\Pi^-(t)(S)\in U(l',t)$ can be constructed by analogy to the above and is left to the reader. □

Remark 11. It is clear from the definitions that the restrictions of $\Pi^+(t)$ and $\Pi^-(t)$ to $\sqcup_l U(l,t)$ are equal to the identity mapping. Since Lemma 17 shows that $\Pi^+(t)$ and $\Pi^-(t)$ map $\mathscr{L}(n)$ *into* $\sqcup_l U(l,t)$, it follows that they each map $\mathscr{L}(n)$ *onto* $\sqcup_l U(l,t)$.

COROLLARY. *Let $S\in\mathscr{L}(n)$. Then*

(a) $\Pi^+(t)(S)\in U(l,t)$ *if and only if* $\dim S\cap M_j(t)-\dim S\cap M_{j-1}(t)=l_j$, $j=1,\cdots,r$.

(b) $\Pi^-(t)(S)\in U(l',t)$ *if and only if* $\dim S\cap N_{2r-j+1}(t)-\dim S\cap N_{2r-j}(t)=l'_j$, $j=1,\cdots,r$.

Proof. If $\dim S\cap M_j(t)-\dim S\cap M_{j-1}(t)=l_j$, $j=1,\cdots,r$, then $\Pi^+(S)\in U(l,t)$ by Lemma 17. The converse follows immediately from the fact that the sets $\{U(l,t)\}_l$ are pairwise disjoint. This proves (a). The proof of (b) is similar. □

The next lemma describes some useful properties of the mappings $\eta_j(t)$, $\Pi^+(t)$, and $\Pi^-(t)$.

LEMMA 18. (a) $\eta_j(t)\Phi(t,t_0)=\Phi(t,t_0)\eta_j(t_0)$. (b) $\Pi^+(t)\Phi(t,t_0)=\Phi(t,t_0)\Pi^+(t_0)$. (c) $\Pi^-(t)\Phi(t,t_0)=\Phi(t,t_0)\Pi^-(t_0)$.

Proof. (a) Let $x \in E_j(t)$. By (38), $\Phi(t_0, t)x \in E_j(t_0)$, so $\Phi(t, t_0)\eta_j(t_0)\Phi(t_0, t)x = \Phi(t, t_0)\Phi(t_0, t)x = x$. Let $y \in \bigoplus_{i=1, i\neq j}^{2r} E_i(t)$. Then $\Phi(t_0, t)y \in \bigoplus_{i=1, i\neq j}^{2r} E_i(t_0)$, so $\Phi(t, t_0)\eta_j(t_0)\Phi(t_0, t)y = 0$. Thus, $\Phi(t, t_0)\eta_j(t_0)\Phi(t_0, t)$ is the projection onto $E_j(t)$ along $\bigoplus_{i=1, i\neq j}^{2r} E_i(t)$, which proves (a).

(b) Let $S \in \mathscr{L}(n)$. Then

$$\begin{aligned}
\Phi(t, t_0)(\Pi^+(t_0)(S)) &= \Phi(t, t_0)\left[\bigoplus_{i=1}^{2r} \eta_i(t_0)(S \cap M_i(t_0))\right] \\
&= \bigoplus_{i=1}^{2r} \eta_i(t)(\Phi(t, t_0)(S \cap M_i(t_0))) \\
&= \bigoplus_{i=1}^{2r} \eta_i(t)[\Phi(t, t_0)(S) \cap \Phi(t, t_0)(M_i(t_0))] \\
&= \bigoplus_{i=1}^{2r} \eta_i(t)[\Phi(t, t_0)(S) \cap M_i(t)] \\
&= \Pi^+(t)(\Phi(t, t_0)(S)),
\end{aligned}$$

where the second equality follows from (a). The proof of (c) is analogous to the proof of (b). □

Remark 12. Lemma 18 implies that $\Pi^+(t)(S(t, S_0, t_0)) = S(t, \Pi^+(t_0)(S_0), t_0)$. Since $\Pi^+(t_0)(S_0) \in U(l, t_0)$ for some l, it follows from this equality together with Theorem 10 that the image of the solution $S(t, S_0, t_0)$ under the time-varying T-periodic mapping $\Pi^+(t)$ is an almost periodic solution, and $\Pi^+(t)(S(t, S_0, t_0)) \in U(l, t)$, $\forall t$. It turns out that $\Pi^+(t)(S(t, S_0, t_0))$ is the motion to which $S(t, S_0, t_0)$ converges as $t \to \infty$. This is proved below. Corresponding statements apply to $\Pi^-(t)(S(t, S_0, t_0))$ as $t \to -\infty$.

In order to prove the next lemma, we need 2 standard facts: (1) Let (X, ρ_X) and (Y, ρ_Y) be metric spaces, and suppose that $f: X \to Y$ is uniformly continuous. If $\rho_X(x_1(t), x_2(t)) \to 0$ as $t \to \infty$, then $\rho_Y(f(x_1(t)), f(x_2(t))) \to 0$ as $t \to \infty$. (2) Let $\mathbb{R}^n$ be endowed with the Euclidean topology. If C is a compact subset of $\mathbb{R}^n$ and U is an open subset of $\mathbb{R}^n$ which contains C, then there exists an open set V with compact closure $\bar{V}$ such that $C \subset V \subset \bar{V} \subset U$.

Recall from § 6 that $V_n(\mathbb{R}^{2n})$ consists of all $2n \times n$ full rank matrices, and $V_n^0(\mathbb{R}^{2n})$ is the compact subset consisting of those matrices whose columns are orthonormal. Let $\|Z\|_1$ denote any of the equivalent norms on the Euclidean space of $2n \times n$ real matrices. Recall that the gap metric on $G^n(\mathbb{R}^{2n})$ is defined by $\theta(S_1, S_2) \equiv \|P_1 - P_2\|$ (operator norm), where P_1, P_2 are the orthogonal projections onto S_1, S_2. Also, $q: V_n(\mathbb{R}^{2n}) \to G^n(\mathbb{R}^{2n})$ is defined by $q(X) \equiv \operatorname{Sp} X$.

LEMMA 19. *Let $X(t) \in V_n^0(\mathbb{R}^{2n})$, let $Y(t) \in V_n(\mathbb{R}^{2n})$ (for each $t \in \mathbb{R}$), and suppose that $\|X(t) - Y(t)\|_1 \to 0$ as $t \to \infty$. Then $\theta(q(X(t)), q(Y(t))) \to 0$ as $t \to \infty$.*

Proof. $V_n^0(\mathbb{R}^{2n})$ is a compact subset of $\mathbb{R}^{2n \times n}$, and $V_n(\mathbb{R}^{2n})$ is an open subset of $\mathbb{R}^{2n \times n}$ which contains $V_n^0(\mathbb{R}^{2n})$. Thus, there exists an open set V with compact closure such that $V_n^0(\mathbb{R}^{2n}) \subset V \subset \bar{V} \subset V_n(\mathbb{R}^{2n})$. Let $\hat{q}$ denote the restriction of q to $\bar{V}$. Since q is continuous and $\bar{V}$ is compact, $\hat{q}$ is uniformly continuous. Since $V_n^0(\mathbb{R}^{2n})$ is compact and is disjoint from the closed set V^c (the complement of V in $\mathbb{R}^{2n \times n}$, there exists $\varepsilon > 0$ such that $\|W - Z\|_1 \geqq \varepsilon$, $\forall W \in V_n^0(\mathbb{R}^{2n})$, $\forall Z \in V^c$. Since $\|X(t) - Y(t)\|_1 \to 0$ as $t \to \infty$, there exists T_1 such that if $t > T_1$, then $\|X(t) - Y(t)\|_1 < \varepsilon$. Since $X(t) \in V_n^0(\mathbb{R}^{2n})$, it follows that $Y(t) \in V$, $\forall t > T_1$. Then for $t > T_1$, we have $\theta(q(X(t)), q(Y(t))) = \theta(\hat{q}(X(t)), \hat{q}(Y(t))) \to 0$ as $t \to \infty$ by the uniform continuity of $\hat{q}$. □

Remark 13. A simple fact concerning the gap metric which is needed below is the following: If $S_1, S_2 \in G^n(\mathbb{R}^{2n})$ and $A \in \mathrm{Gl}(2n, \mathbb{R})$, then $\theta(A(S_1), A(S_2)) \leqq \|A\| \|A^{-1}\| \theta(S_1, S_2)$. To prove this fact, let P_1, P_2 denote the orthogonal projections onto S_1, S_2 respectively. Then AP_1A^{-1}, AP_2A^{-1} are projections (not necessarily orthogonal) onto $A(S_1)$, $A(S_2)$ respectively. It can be shown [10, p. 361] that this implies that $\theta(A(S_1), A(S_2)) \leqq \|AP_1A^{-1} - AP_2A^{-1}\|$. Hence, $\theta(A(S_1), A(S_2)) \leqq \|A\| \|A^{-1}\| \|P_1 - P_2\| = \|A\| \|A^{-1}\| \theta(S_1, S_2)$ as claimed.

The next result describes the asymptotic behavior of *every* solution of the EPRDE on the Lagrange-Grassmann manifold. It shows that every solution converges to an almost periodic solution as $t \to \infty$ ($t \to -\infty$), and describes the limiting solution.

THEOREM 13. *Suppose that* $\Phi(T, 0)$ *has no eigenvalues on the unit circle and is semisimple. Let* $S_0 \in \mathscr{L}(n)$. *Then*

(a) $\theta(S(t, S_0, t_0), S(t, \Pi^+(t_0)(S_0), t_0)) \to 0$ *as* $t \to \infty$.

(b) $\theta(S(t, S_0, t_0), S(t, \Pi^-(t_0)(S_0), t_0)) \to 0$ *as* $t \to -\infty$.

Proof. (a) We can easily reduce to the case where $t_0 = 0$. To see this, let $S_1 \equiv \Phi(0, t_0)(S_0)$. Then $S(t, S_0, t_0) = S(t, S_1, 0)$. By Lemma 18(b), it follows that $S(t, \Pi^+(t_0)(S_0), t_0) = S(t, \Phi(0, t_0)\Pi^+(t_0)(S_0), 0) = S(t, \Pi^+(0)(S_1), 0)$. Consequently, it suffices to show that $\theta(S(t, S_1, 0), S(t, \Pi^+(0)(S_1), 0)) \to 0$ as $t \to \infty$.

Let $\hat{S}_1 \equiv \Pi^+(0)(S_1)$ and suppose that $\hat{S}_1 \in U(l, 0)$. Let P, $V(t)$, R, $\hat{R}$, and Λ be as defined in the proof of Theorem 10. Using (40) and Remark 13, it follows that

$$\begin{aligned}\theta(S(t, S_1, 0), S(t, \hat{S}_1, 0)) &= \theta(PV(t)\, e^{Rt}P^{-1}(S_1), PV(t)\, e^{Rt}P^{-1}(\hat{S}_1)) \\ &\leqq \|PV(t)\| \|V^{-1}(t)P^{-1}\| \theta(e^{Rt}P^{-1}(S_1), e^{Rt}P^{-1}(\hat{S}_1)).\end{aligned}$$

Let $\alpha_1 \equiv \max_{t \in [0, 2T]} \|PV(t)\|$ and let $\alpha_2 \equiv \max_{t \in [0, 2T]} \|V^{-1}(t)P^{-1}\|$. Since $V(t)$ is periodic with period $2T$, we have $\theta(S(t, S_1, 0), S(t, \hat{S}_1, 0)) \leqq \alpha_1\alpha_2 \theta(e^{Rt}P^{-1}(S_1), e^{Rt}P^{-1}(\hat{S}_1))$, $\forall t$. Consequently, it suffices to show that $\theta(e^{Rt}P^{-1}(S_1), e^{Rt}P^{-1}(\hat{S}_1)) \to 0$ as $t \to \infty$.

Since $\Pi^+(0)(S_1) \in U(l, 0)$, it follows from the proof of Lemma 17 that S_1 has a basis of the form $\{v_{ij} + w_{ij} : j = 1, \cdots, l_i;\ i = 1, \cdots, 2r\}$ where $v_{ij} \in E_i(0)$, $w_{ij} \in M_{i-1}(0)$, and $\{v_{ij}\}$ is a basis for $\Pi^+(0)(S_1) \equiv \hat{S}_1$. $E_i(0)$ is the sum of the primary components of $\Phi(T, 0)$ which correspond to eigenvalues of modulus μ_i, so $P^{-1}(E_i(0))$ is the sum of the primary components of $P^{-1}\Phi(T, 0)P = \Lambda$ which correspond to eigenvalues of modulus μ_i. From the structure of Λ, it follows that $P^{-1}(E_i(0)) \perp P^{-1}(E_k(0))$ whenever $i \neq k$. Consequently, $P^{-1}v_{ij} \perp P^{-1}v_{kp}$ whenever $i \neq k$. However, by using the Gram-Schmidt process, it is clear that the basis $\{v_{ij} + w_{ij}\}$ can be chosen in such a way that $P^{-1}v_{ij} \perp P^{-1}v_{ip}$ and such that $P^{-1}v_{ij}$ has unit length. Thus, without loss of generality, we can assume that $\{P^{-1}v_{ij}\}$ is an orthonormal basis for $P^{-1}(\hat{S}_1)$.

From (43) we have $e^{Rt}x = \mu_i^{t/T}\, e^{\hat{R}t}x$, $\forall x \in P^{-1}(E_i(0))$. Thus, $e^{Rt}P^{-1}(\hat{S}_1) = \mathrm{Sp}\{e^{Rt}P^{-1}v_{ij}\} = \mathrm{Sp}\{\mu_i^{t/T}\, e^{\hat{R}t}P^{-1}v_{ij}\} = \mathrm{Sp}\{e^{\hat{R}t}P^{-1}v_{ij}\}$. Since $\{P^{-1}v_{ij}\}$ is orthornormal and $e^{\hat{R}t}$ is an orthogonal matrix, $\{e^{\hat{R}t}P^{-1}v_{ij}\}$ is an orthonormal basis for $e^{Rt}P^{-1}(\hat{S}_1)$. We also have $e^{Rt}P^{-1}(S_1) = \mathrm{Sp}\{e^{Rt}P^{-1}(v_{ij} + w_{ij})\} = \mathrm{Sp}\{\mu_i^{t/T}\, e^{\hat{R}t}P^{-1}v_{ij} + e^{Rt}P^{-1}w_{ij}\} = \mathrm{Sp}\{e^{\hat{R}t}P^{-1}v_{ij} + \mu_i^{-t/T}\, e^{Rt}P^{-1}w_{ij}\}$. Now, $w_{ij} \in M_{i-1}(0) = E_1(0) \oplus \cdots \oplus E_{i-1}(0)$, so $P^{-1}w_{ij} \in P^{-1}(E_1(0)) \oplus \cdots \oplus P^{-1}(E_{i-1}(0))$. Let $P^{-1}w_{ij} = x_{ij}^1 + \cdots + x_{ij}^{i-1}$ with $x_{ij}^k \in P^{-1}(E_k(0))$. Then $e^{Rt}P^{-1}w_{ij} = \sum_{k=1}^{i-1} \mu_k^{t/T}\, e^{\hat{R}t}x_{ij}^k$. Since $e^{\hat{R}t}$ is orthogonal and $\mu_k < \mu_i$ for $k < i$, it follows immediately that $\mu_i^{-t/T}\, e^{Rt}P^{-1}w_{ij} \to 0$ as $t \to \infty$. Let $X(t)$ denote the $2n \times n$ matrix whose columns are $\{e^{\hat{R}t}P^{-1}v_{ij}\}$, and let $Y(t)$ denote the $2n \times n$ matrix whose columns are $\{e^{\hat{R}t}P^{-1}v_{ij} + \mu_i^{-t/T}\, e^{Rt}P^{-1}w_{ij}\}$. Then $X(t) \in V_n^0(\mathbb{R}^{2n})$, $Y(t) \in V_n(\mathbb{R}^{2n})$, and $\|X(t) - Y(t)\|_1 \to 0$ as $t \to \infty$. By Lemma 19, $\theta(q(X(t)), q(Y(t))) \to 0$ as $t \to \infty$. Thus, $\theta(e^{Rt}P^{-1}(\hat{S}_1), e^{Rt}P^{-1}(S_1)) \to 0$ as $t \to \infty$, completing the proof of (a).

The proof of (b) is analogous to the proof of (a). □

Remark 14. Given $S_0 \in \mathscr{L}(n)$, the limiting almost periodic solution to which the solution $S(t, S_0, t_0)$ converges as $t \to \infty$ can be computed in 2 different ways. The first way is to compute the initial condition $\Pi^+(t_0)(S_0)$ and then compute the solution $S(t, \Pi^+(t_0)(S_0), t_0)$ corresponding to this initial condition. The second way is to compute the solution $S(t, S_0, t_0)$ and then compute its image under the periodically time-varying mapping $\Pi^+(t)$. Since $S(t, \Pi^+(t_0)(S_0), t_0) = \Pi^+(t)(S(t, S_0, t_0))$ (Remark 12), the 2 methods produce the same limiting solution.

Remark 15. We can now complete the proof of Theorem 10 by proving (b). Let $S_0 \in \mathscr{L}(n)$ and suppose that $S_0 \notin U(l, t_0)$ for any l. Then $\Pi^+(t_0)(S_0) \neq S_0$. By Theorem 13, $\theta(S(t, S_0, t_0), S(t, \Pi^+(t_0)(S_0), t_0)) \to 0$ as $t \to \infty$. Since $S(t, \Pi^+(t_0)(S_0), t_0)$ is almost periodic and is not identically equal to $S(t, S_0, t_0)$, it follows from AP 6 that $S(t, S_0, t_0)$ is not almost periodic. This completes the proof.

We now describe the asymptotic behavior of *every* solution of the PRDE on the space $S(n)$ of symmetric matrices.

THEOREM 14. *Suppose that* $(A(\cdot), B(\cdot))$ *is controllable and that* $\Phi(T, 0)$ *is semisimple and has no eigenvalues on the unit circle.*

(a) *If* $K_0 \not\leqq K^+(t_0)$, *then* $K(t, K_0, t_0)$ *escapes at a finite time* $t_e > t_0$.

(b) *If* $K_0 \leqq K^+(t_0)$, *then* $K(t, K_0, t_0)$ *exists for all* $t \in [t_0, \infty)$ *and converges to the almost periodic solution* $K(t, \gamma^{-1}(\Pi^+(t_0)(\gamma(K_0))), t_0)$ *as* $t \to \infty$.

(c) *If* $K_0 \not\geqq K^-(t_0)$, *then* $K(t, K_0, t_0)$ *escapes at a finite time* $t_e < t_0$.

(d) *If* $K_0 \geqq K^-(t_0)$, *then* $K(t, K_0, t_0)$ *exists for all* $t \in (-\infty, t_0]$ *and converges to the almost periodic solution* $K(t, \gamma^{-1}(\Pi^-(t_0)(\gamma(K_0))), t_0)$ *as* $t \to -\infty$.

Proof. Theorem 9 establishes all of the assertions concerning the existence and nonexistence of finite escape times. In fact, these results are valid even if $\Phi(T, 0)$ is not semisimple. Given any $K_0 \in S(n)$, $\Pi^+(t_0)(\gamma(K_0)) \in U(l, t_0)$ and $\Pi^-(t_0)(\gamma(K_0)) \in U(l', t_0)$ for some l and l'. Consequently, Theorem 11 establishes the almost periodicity of the solutions $K(t, \gamma^{-1}(\Pi^+(t_0)(\gamma(K_0))), t_0)$ and $K(t, \gamma^{-1}(\Pi^-(t_0)(\gamma(K_0))), t_0)$. Thus, it remains only to prove the assertions in (b) and (d) concerning convergence.

Suppose that $K_0 \leqq K^+(t_0)$. Let $S_0 \equiv \gamma(K_0)$ and let $S_1 \equiv \Pi^+(t_0)(\gamma(K_0))$. By Theorem 13, $\theta(S(t, S_0, t_0), S(t, S_1, t_0)) \to 0$ as $t \to \infty$. Since $K_0 \leqq K^+(t_0)$, $S(t, S_0, t_0) \in \mathscr{L}_0(n)$ for all $t \geqq t_0$. Since $K(t, K_0, t_0) = \gamma^{-1}(S(t, S_0, t_0))$ and $K(t, \gamma^{-1}(\Pi^+(t_0)(\gamma(K_0))), t_0) = \gamma^{-1}(S(t, S_1, t_0))$, we need to show that $\|\gamma^{-1}(S(t, S_0, t_0)) - \gamma^{-1}(S(t, S_1, t_0))\| \to 0$ as $t \to \infty$.

Let $C \equiv \cup_{t \in \mathbb{R}} U(l, t) = \cup_{t \in [0, T]} U(l, t)$. C is compact since it is the image of $U(l, 0) \times [0, T]$ under the continuous mapping $(S, t) \mapsto \Phi(t, 0)(S)$. By the Corollary to Lemma 16, $C \subset \mathscr{L}_0(n)$. Since C and $\mathscr{L}(n) - \mathscr{L}_0(n)$ are disjoint closed sets in the metric space $\mathscr{L}(n)$ (which is a normal topological space), there exist disjoint open sets U, V such that $C \subset U$ and $\mathscr{L}(n) - \mathscr{L}_0(n) \subset V$. In particular, this implies that $C \subset U \subset \bar{U} \subset \mathscr{L}_0(n)$. Since C and $\mathscr{L}(n) - U$ are nonempty disjoint compact sets, there exists $\delta > 0$ such that $\theta(S_\alpha, S_\beta) \geqq \delta$, $\forall S_\alpha \in C$, $\forall S_\beta \in \mathscr{L}(n) - U$. Since $\theta(S(t, S_0, t_0), S(t, S_1, t_0)) \to 0$ as $t \to \infty$ and $S(t, S_1, t_0) \in C, \forall t$, there must exist $T_1 \geqq t_0$ such that $S(t, S_0, t_0) \in U, \forall t \geqq T_1$. Restricted to the compact set $\bar{U}$, the continuous mapping γ^{-1} is uniformly continuous. Consequently, the fact that $\theta(S(t, S_0, t_0), S(t, S_1, t_0)) \to 0$ as $t \to \infty$ implies that $\|\gamma^{-1}(S(t, S_0, t_0)) - \gamma^{-1}(S(t, S_1, t_0))\| \to 0$ as $t \to \infty$ as required. This completes the proof of (b). The proof of the convergence asserted in (d) is completely analogous. □

Remark 16. We can now complete the proof of Theorem 11 by proving (b). Let $K_0 \in S(n)$, and suppose that $K(t, K_0, t_0)$ is *not* of the form $\gamma^{-1}(\Phi(t, t_0)(S_0))$ for some $S_0 \in \sqcup_l U(l, t_0)$. In other words, $\gamma(K_0) \notin \sqcup_l U(l, t_0)$. By Theorem 14, $K(t, K_0, t_0)$ either escapes at a finite time $t_e > t_0$ or converges to the almost periodic solution $K(t, \gamma^{-1}(\Pi^+(t_0)(\gamma(K_0))), t_0)$ as $t \to \infty$. Since $\gamma(K_0) \notin \sqcup_l U(l, t_0)$, $\Pi^+(t_0)(\gamma(K_0)) \neq$

$\gamma(K_0)$. Since a pair of distinct almost periodic functions cannot converge (AP 6), we conclude that in either case, $K(t, K_0, t_0)$ is not almost periodic. This completes the proof.

Remark 17. Parts (b) and (c) of Theorem 8 give necessary and sufficient conditions for $K(t, K_0, t_0)$ to converge to the periodic equilibrium $K^+(t)$ as $t \to -\infty$ and to the periodic equilibrium $K^-(t)$ as $t \to \infty$. If one makes the additional assumption that $\Phi(T, 0)$ is semisimple, these results can be recovered as easy consequences of Theorem 14. Let $K_0 \in S(n)$. Then by Theorem 14, $K(t, K_0, t_0)$ converges to $K^-(t)$ as $t \to \infty$ if and only if $K_0 \leqq K^+(t_0)$ and $\gamma^{-1}(\Pi^+(t_0)(\gamma(K_0))) = K^-(t_0)$. From the definition of $K^-(t)$, it follows that $\{\gamma(K^-(t_0))\} = U(l, t_0)$ for the case where $l = (l_1, \cdots, l_r) = (0, \cdots, 0)$. By Lemma 17, it follows that $\Pi^+(t_0)(\gamma(K_0)) = \gamma(K^-(t_0))$ if and only if $\dim \gamma(K_0) \cap M_j(t_0) = 0$, $j = 1, \cdots, r$, or equivalently, if and only if $\dim \gamma(K_0) \cap M_r(t_0) = 0$. But, $M_r(t_0) = \gamma(K^+(t_0))$, so this condition can be expressed as $\gamma(K_0) \cap \gamma(K^+(t_0)) = 0$. It is trivial to check that this condition is satisfied if and only if $K^+(t_0) - K_0$ is nonsingular. Thus, $K(t, K_0, t_0)$ converges to $K^-(t)$ as $t \to \infty$ if and only if $K_0 \leqq K^+(t_0)$ and $K^+(t_0) - K_0$ is nonsingular, or equivalently, if and only if $K_0 < K^+(t_0)$. This establishes part (c) of Theorem 8. Part (b) of Theorem 8 can be proved from Theorem 14 by an analogous argument.

8. Conclusions. In this paper, we have provided a rather complete description of the phase portrait of the matrix Riccati equation with periodic coefficients. We have generalized to the periodic case many of the key results in the theory of time-invariant Riccati equations. As noted in § 1, almost all of these results can be extended to the case of a Riccati equation arising from a periodic control problem with conflicting objectives.

By generalizing to the periodic case a well-known theorem of J. C. Willems, we have classified every periodic equilibrium. In the case where the system is observable, our results describe the signature of each periodic equilibrium. The stability of each periodic equilibrium is determined, and necessary and sufficient conditions are given for convergence to $K^+(t)$ as $t \to -\infty$ and for convergence to $K^-(t)$ as $t \to \infty$. We also give necessary and sufficient conditions for a solution to have a finite escape in either forward or backward time. All of these results are valid even if the monodromy matrix $\Phi(T, 0)$ is nondiagonalizable.

In §§ 6 and 7, we impose the requirement that $\Phi(T, 0)$ be diagonalizable, but we continue to allow multiple eigenvalues. We give an exact description of the oscillating Grassmannian manifolds on which all of the almost periodic solutions occur. An arbitrary solution either escapes in finite forward (backward) time or converges to an almost periodic solution as $t \to \infty$ ($t \to -\infty$). We give an explicit formula for the limiting almost periodic solution. Since we have also given necessary and sufficient conditions for a solution to have a finite escape time, these results describe the asymptotic behavior of *every* solution of the PRDE.

Appendix. It is straightforward to show that many of the basic properties of complex-valued almost periodic functions generalize to the case where the values are in a complete metric space. (For properties requiring algebraic structure, one takes X to be a Banach space.) By making obvious changes in the proofs of the corresponding results for complex-valued almost periodic functions in [9, Chaps. 1, 2], the following properties can be verified:

AP 1. Let (X, ρ) be a complete metric space and let $f: \mathbb{R} \to X$ be continuous. Then f is almost periodic if and only if given any $\varepsilon > 0$, there exists $L = L(\varepsilon)$ such that given any s, there exists $\tau \in [s, s + L]$ such that $\rho(f(t), f(t + \tau)) < \varepsilon$ for all $t \in \mathbb{R}$.

AP 2. Let (X, ρ) be a complete metric space and let $f: \mathbb{R} \to X$ be a continuous periodic function. Then f is almost periodic.

AP 3. Let $\{(X_i, \rho_i)\}_{i=1}^r$ be complete metric spaces and let $f_i: \mathbb{R} \to X_i$ be almost periodic, $i = 1, \cdots, r$. If $f: \mathbb{R} \to X_1 \times \cdots \times X_r$ is defined by $f(t) = (f_1(t), \cdots, f_r(t))$, then f is almost periodic (relative to the product metric).

AP 4. Let $\{X_i\}_{i=1}^3$ be Banach spaces, and let $\alpha: X_1 \times X_2 \to X_3$ be a product mapping. (That is α is bilinear and $\|\alpha(x_1, x_2)\| \leqq \|x_1\| \, \|x_2\|$.) If $f_i: \mathbb{R} \to X_i$ $(i = 1, 2)$ are almost periodic and if $f: \mathbb{R} \to X_3$ is defined by $f(t) = \alpha(f_1(t), f_2(t))$, then f is almost periodic.

AP 5. Let (X_1, ρ_1) and (X_2, ρ_2) be complete metric spaces, and let $f: \mathbb{R} \to X_1$ be almost periodic. Let Ω be a subset of X_1 which contains the image of f. If $F: \Omega \to X_2$ is uniformly continuous, then $F \circ f$ is almost periodic.

AP 6. Let (X, ρ) be a complete metric space, and let $f_i: \mathbb{R} \to X$ be almost periodic, $i = 1, 2$. If $\rho(f_1(t), f_2(t)) \to 0$ as $t \to \infty$, then $f_1(t) = f_2(t)$, $\forall t$.

REFERENCES

[1] E. BEKIR AND R. S. BUCY, *Periodic equilibria for matrix Riccati equations*, Stochastics, 2 (1976), pp. 1-104.

[2] S. BITTANTI et al., *H-controllability and observability of linear periodic systems*, Proc. IEEE Conference on Decision and Control, San Antonio, TX, 1983, pp. 1376-1379.

[3] S. BITTANTI, P. COLANERI AND G. GUARDABASSI, *Periodic solutions of periodic Riccati equations*, IEEE Trans. Automat. Control, 29 (1984), pp. 665-667.

[4] S. BITTANTI et al., *Periodic systems: controllability and the matrix Riccati equation*, this Journal, 16 (1978), pp. 37-40.

[5] R. W. BROCKETT, *Finite-Dimensional Linear Systems*, John Wiley, New York, 1970.

[6] P. BRUNOVSKY, *Controllability and linear closed-loop controls in linear periodic systems*, J. Differential Equations, 6 (1969), pp. 296-313.

[7] J. R. CANABAL, *Periodic geometry of the Riccati equation*, Stochastics, 1 (1974), pp. 432-440.

[8] W. A. COPPEL, *Matrix quadratic equations*, Bull. Austral. Math. Soc., 10 (1974), pp. 377-401.

[9] A. M. FINK, *Almost Periodic Differential Equations*, Springer-Verlag, Berlin, 1974.

[10] I. GOHBERG, P. LANCASTER AND L. RODMAN, *Matrix Polynomials*, Academic Press, New York, 1982.

[11] R. HERMANN, *Cartanian Geometry, Nonlinear Waves, and Control Theory, Part* A, Interdisciplinary Mathematics, Vol. XX, Math. Science Press, Brookline, MA, 1979.

[12] G. A. HEWER, *Periodicity, detectability and the matrix Riccati equation*, this Journal, 13 (1975), pp. 1235-1251.

[13] H. KANO AND T. NISHIMURA, *Periodic solution of matrix Riccati equations with detectability and stabilizability*, Int. J. Contr., 29 (1979), pp. 471-487.

[14] C. MARTIN, *Finite escape time for Riccati differential equations*, Systems and Control Letters, 1 (1981), pp. 127-131.

[15] ———, *Grassmannian manifolds, Riccati equations and feedback invariants of linear systems*, in Geometrical Methods for the Theory of Linear Systems, C. Byrnes and C. Martin, eds., Reidel, Dordrecht, 1980.

[16] D. A. SANCHEZ, *A note on periodic solutions of Riccati-type equations*, SIAM J. Appl. Math., 17 (1969), pp. 957-959.

[17] ———, *Computing periodic solutions of Riccati differential equations*, Appl. Math. Comp., 6 (1980), pp. 283-287.

[18] C. R. SCHNEIDER, *Global aspects of the matrix Riccati equation*, Math. Systems Theory, 7 (1973), pp. 281-286.

[19] M. A. SHAYMAN, *Inertia theorems for the periodic Liapunov equation and periodic Riccati equation*, Systems and Control Letters, 4 (1984), pp. 27-32.

[20] ———, *On the periodic solutions of the matrix Riccati equation*, Math. Systems Theory, 16 (1983), pp. 267-287.

[21] ———, *On the variety of invariant subspaces of a finite-dimensional linear operator*, Trans. Amer. Math. Soc., 274 (1982), pp. 721-747.

[22] ———, *Phase portrait of the matrix Riccati equation*, this Journal.

[23] ———, *Phase portrait of the matrix Riccati equation: Global structure*, Proc. Conference on Information Science and Systems, Johns Hopkins Univ., Baltimore, 1983, pp. 263-270.

[24] J. C. WILLEMS, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automat. Control, 16 (1971), pp. 621-634.

[25] J. L. WILLEMS, *Stability Theory of Dynamical Systems*, John Wiley, New York, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

006

ON THE PARAMETRIZATION OF LINEAR CONSTANT SYSTEMS*

L. BARATCHART†

Abstract. This paper introduces special factorizations for the transfer functions of linear constant systems, and shows that these factorizations allow parametrizations of systems with given cyclic structure in the state space. This last set is proved to be a submanifold of the manifold of systems of given order, thereby stressing some links between our approach and the classical construction due to Hazewinkel and Kalman.

Key words. parametrization of linear systems, transfer function factorization, differential structure

Introduction. Geometric aspects of the parametrization problem for linear constant systems, of given order n, have been first studied by Kalman [19], [20], whose construction allows one to consider the set S_n of such systems as a quasi-projective variety over an algebraically closed field K, as well as an analytic manifold when $K = \mathbf{R}$ or $\mathbf{C}$. These have been further studied by several authors, using techniques of algebraic geometry (e.g. constructing moduli spaces, geometric quotients [7], [8], [12], [14], [22]), or of differential geometry and algebraic topology (e.g. looking for continuous sections, classifying spaces, computing topological invariants [6], [7], [9], [13], [15], [16]). The fact that (H, F, G) triples form a principal Gl_n bundle over S_n is of importance in many of these works; this may contribute to explain why they are not much concerned with the external representation of systems (except [6] which deals with the scalar case). Now, Kalman's construction gives coordinates on the manifold in terms of (H, F, G) "canonical forms," and such forms, as well as some external analogues in terms of $D^{-1}N$ (or ND^{-1}) factorizations of the transfer function, have been examined from an algebraic point of view by other authors, in particular for identification purposes [1], [5]. As is well known, a given system admits generally many representations of this type, in accordance with the fact that charts have to overlap, since the differential structure is not trivial [13]. Another canonical form, related to the Kronecker indices of pairs (F, G) (or (H, F)), is due in essence to Kalman in its state-space formulation [21], and in its external version to Beghelli-Guidorzi [4] (see also [1], [11]). Its uniqueness, for a given system, makes it algebraically attractive, but it is geometrically very different from the preceding ones, since it can no longer represent coordinates on S_n. More precisely, it parametrizes an open subset of S_n when the Kronecker indices are in "generic configuration," and a submanifold of strictly lower dimension in other cases (these "Kronecker submanifolds," each isomorphic to some K^l, play an essential role in Helmke's computation of mod-2 Betti numbers of S_n [15]).

In this paper, we deal with another partition of S_n into submanifolds which, to our knowledge, have not been studied before, namely those consisting in systems with a given cyclic structure in the state-space. Assuming $K = \mathbf{R}$ (but nothing is to be changed if $K = \mathbf{C}$), we prove that this set is a submanifold of S_n (a related result is given in [22], cf. § 5), and compute its dimension by exhibiting an atlas in terms of (H, F, G) triples, which is shown to be related to another system of charts, in terms of certain transfer function factorizations $RD^{-1}N$ (R a scalar matrix). The structure of D, especially when the number of invariants is small, allows its explicit inversion, and the parameters appear then in the transfer function itself more explicitly than in

* Received by the editors July 7, 1983, and in revised form March 5, 1984.
† INRIA, Route des Lucioles, Sophia Antipolis, 06560 Valbonne, France.

the previously mentioned works. In particular, the dense open set of cyclic systems admits a rather simple description. We point out here that Beghelli-Guidorzi's factorization allows a nice description of Kronecker submanifolds, since the parameters are just the coefficients of the entries of D and N; however, D has generically no special structure, so that the expression of $D^{-1}N$ itself is quite complicated.

The existence of our factorization could be proved combining results of this paper with an algebraic identity given in [1] (cf. § 5). However, we derive it here from a factorization theorem for regular polynomial matrices which is new, and stated in § 2; § 1 recalls Kalman's construction and Fuhrmann realization theory, together with some classical results in linear algebra. Nothing is used in differential geometry beyond the basic definitions, which are assumed to be familiar to the reader.

1. Preliminaries. We let p, n, m be positive integers, chosen once and for all. By a system, we shall mean in the sequel an input-output map $S: u(t) \to y(t)$, which may be represented by some state-space equations:

$$\dot{x} = Fx + Gu, \qquad y = Hx \tag{1}$$

where H, F, G are real matrices in $\mathbf{R}^{p\times n}$, $\mathbf{R}^{n\times n}$, $\mathbf{R}^{n\times m}$ respectively. Equivalently, a system may thus be considered an equivalence class of triples (H, F, G) which induce the same S. These triples are said to be equivalent, and any of them is called a realization of S. We will restrict our attention to systems of minimal dimension n, and we denote this set by $\mathbf{S}_n$. It is well known that if we consider the reachability and observability matrices of any realization (H, F, G) of S,

$$\mathbf{R}_{FG} = (G, FG, \cdots, F^{n-1}G), \qquad \mathbf{O}_{HF} = \begin{bmatrix} H \\ HF \\ \vdots \\ HF^{n-1} \end{bmatrix}$$

then S belongs to $\mathbf{S}_n$ if and only if $\mathbf{R}_{FG}$ and $\mathbf{O}_{HF}$ are of full rank. Triples verifying this condition will be said canonical, and in that case, (H, F, G) is equivalent to (H', F', G') if and only if

$$H' = HP^{-1}, \quad F' = PFP^{-1}, \quad G' = PG$$

for some unique regular matrix P. We then denote

$$(H', F', G') = (H, F, G)^P. \tag{2}$$

Now if we define $\mathbf{L}_n$ to be the open set, in $\mathbf{R}^{pn+nn+nm}$, of canonical triples, $\mathbf{S}_n$ identifies with the equivalence classes of $\mathbf{L}_n$ for the preceding relation. Let us denote by $\Pi: \mathbf{L}_n \to \mathbf{S}_n$ the canonical map. We endow $\mathbf{L}_n$ with its natural topology, and $\mathbf{S}_n$ with the quotient topology, defining O in $\mathbf{S}_n$ to be open if and only if $\Pi^{-1}(O)$ is open. We shall denote by Gl_n the set of regular matrices in $\mathbf{R}^{n\times n}$. Π is an open mapping, because if U is open in $\mathbf{L}_n$, and $P \in Gl_n$, $U^P = \{u^P;\ u \in U\}$ is open, and hence $\Pi^{-1}(\Pi(U)) = \bigcup_P U^P$ is also open. Using this we now show that $\mathbf{S}_n$ is Hausdorff. Let (H, F, G) and (H', F', G') be in $\mathbf{L}_n$, and consider the linear equations

$$P\mathbf{R}_{FG} = \mathbf{R}_{F'G'}, \qquad \mathbf{O}_{HF} = \mathbf{O}_{H'F'}P. \tag{3}$$

It is easily seen that our two systems are equivalent if and only if (3) is solvable. Now if it is not, extracting a Cramer's subsystem (e.g. in the first one), we conclude that the other equations do not depend linearly on that one, or equivalently that some determinant is nonzero. This will remain so under small changes in the coefficients of

H, F, G, H', F', G', and therefore there exist two open sets O and O', containing (H, F, G) and (H', F', G') respectively, such that $\Pi(O)$ and $\Pi(O')$ are disjoint. This means that $\mathbf{S}_n$ is Hausdorff.

Now that $\mathbf{S}_n$ has been made into a topological space, the following construction, due to Hazewinkel and Kalman [13], [20], allows to endow it with an analytic manifold structure, of dimension $n(m+p)$.

Let (H, F, G) be a realization of $S \in \mathbf{S}_n$. We denote by $g_1, \cdots, g_m$ the columns of G, so that each column of $\mathbf{R}_{FG}$, is of the form $C_{ij} = F^i g_j$, $i \in \{0, \cdots, n-1\}$, $j \in \{1, \cdots, m\}$. From the definition of $\mathbf{S}_n$, n among these columns are independent. In fact, it is not difficult to see that these n columns may be chosen so that any time we select C_{ij}, we have already selected C_{kj}, for $k < i$. Our selection ν of pairs (i, j) is called nice, meaning that $(i, j) \in \nu$ implies $(k, j) \in \nu$ for $k < i$, and we will say that S admits ν as nice selection, meaning that the submatrix P_ν of $\mathbf{R}_{FG}$ corresponding to the indices of ν is regular. Because (2) implies (3), this notion does not depend on our choice of (H, F, G), but only on S, and so is well defined. Now we define

$$(H_\nu, F_\nu, G_\nu) = (H, F, G)^{P_\nu^{-1}} \tag{4}$$

and our new realization is easily seen to have the following form,

$$F_\nu = \begin{bmatrix} 0 & & & x & & & & x & \\ 1 & \ddots & & x & & & & x & \\ & \ddots & 0 & \vdots & & & & \vdots & \\ & & 1 & x & & & & x & \cdots \\ & & & x & 0 & & & & \\ & & & & 1 & \ddots & & & \\ & & & \vdots & & \ddots & 0 & \vdots & \\ & & & x & & & 1 & x & \end{bmatrix}, \quad G_\nu = \begin{bmatrix} & 1 & & 0 & \\ & 0 & & 0 & \\ & \vdots & & \vdots & \\ X & 0 & X & 0 & \cdots \\ & 0 & & 1 & \\ & 0 & & 0 & \\ & \vdots & & \vdots & \\ & 0 & & 0 & \end{bmatrix}$$

where the 0's and 1's occur in G_ν precisely in the place of the columns of G which have been selected in ν, the 1's being in correspondence with the first row of each diagonal companion block in F_ν.

If P'_ν denotes the analogue of P_ν for $\mathbf{R}_{F',G'}$, it is easily checked that (2) implies $P'_\nu = PP_\nu$, and therefore that (H_ν, F_ν, G_ν) does not depend on (H, F, G) but only on S, and is the only realization of S such that F and G have the preceding form.

Let $\mathbf{S}_n^\nu$ be the set of $S \in \mathbf{S}_n$ admitting ν as a nice selection. This set is open, because it is the image under Π of the open set consisting in triples (H, F, G) such that P_ν is regular.

We now define $\psi_\nu : \mathbf{S}_n^\nu \to \mathbf{R}^{n(m+p)}$ such that $\psi_\nu(S)$ is the list, arranged in any conventional order, of the coefficients of H_ν, and those coefficients of F_ν and G_ν which were represented by x's in the preceding diagram. ψ_ν is injective, and its image V_ν is obviously open. By definition of the quotient topology, it is easy to see that ψ_ν is a homeomorphism onto V_ν. It remains to prove that if ν, μ are nice selections, the following map is analytic

$$\psi_\nu \circ \psi_\mu^{-1} : \psi_\mu(\mathbf{S}_n^\mu \cap \mathbf{S}_n^\nu) \to V_\nu$$

but this follows from the fact that $(H_\mu, F_\mu, G_\mu) \to (H_\mu, F_\mu, G_\mu)^{P_\nu^{-1}}$ is rational and therefore analytic. Our charts on $\mathbf{S}_n$ will then be the $(\mathbf{S}_n^\nu, \psi_\nu)$.

Before we go further on, we need some notation and classical results about polynomial matrices that we merely state. We denote by $\mathbf{R}[z]$ and $\mathbf{R}(z)$ the set of real polynomials and rational fractions respectively. If $r \in \mathbf{R}[z]$, $\deg(r)$ means the degree

of r. A fraction $p/q \in \mathbf{R}(z)$ is said to be strictly proper if $\deg(p) < \deg(q)$, and a rational matrix if all its entries are. For the sake of brevity, we shall simply say "proper" instead of "strictly proper." We call a polynomial matrix with constant determinant unimodular, and two polynomial matrices are equivalent if they differ by right and left multiplication by a unimodular matrix. Left (right) equivalence allows only left (right) multiplication in the above definition. Any polynomial matrix N is right equivalent to a matrix of the following form [18]:

$$\begin{bmatrix} & X & \cdots \\ & a_2 & \\ a_1 & & O \end{bmatrix} \quad \text{with } \deg(a_i) > \deg(\text{other element in the same row}).$$

Furthermore if Δ_i is the g.c.d. of the i-rowed minors constructed with the i last rows of $\mathbf{N}$, then $\prod_{j=1}^{i} a_i = \Delta_i$. In particular, if none of the a_i's is zero, they are uniquely defined [2]. Such a form is called a Hermite form for N. Of course, we could have chosen the opposite inclination for the "diagonal," so that the minors would have been built on the first rows of N. We get dual results with left equivalence. If in addition N is square and regular (the result can be made more general, but we will not need it), it is equivalent to a diagonal matrix $D = \operatorname{diag}\{d_1, \cdots, d_k\}$, where $d_i \mid d_{i+1}$ ($\mid$ means divides). The d_i's are uniquely determined by these conditions, and if Δ_i is the g.c.d. of all i-rowed minors in N, then $d_i = \Delta_i/\Delta_{i-1}$ ($\Delta_0 = 1$ conventionally).

Now if we consider the $\mathbf{R}[z]$-module $\mathbf{R}^k[z]$ and its submodule $N\mathbf{R}^k[z] = \{Nx;\ x \in \mathbf{R}^k[z]\}$, the preceding result shows that the quotient module: $\mathbf{R}^k[z]/N\mathbf{R}^k[z]$ is isomorphic to the following direct sum: $\sum_i \mathbf{R}[z]/(d_i)$, (d_i) being the ideal generated by d_i in $\mathbf{R}[z]$. In particular, its dimension as a vector space over $\mathbf{R}$ is

$$\sum_i \deg(d_i) = \deg(\det N). \tag{5}$$

The polynomials d_i are referred to as the invariants of N.

An essentially equivalent formulation of this result is the following. Let $F \in \mathbf{R}^{k\times k}$. If $r \in \mathbf{R}[z]$, $v \in \mathbf{R}^k$, we write $r \cdot v$ for the endomorphism $r(F)$ applied to v. This makes $\mathbf{R}^k$ into an $\mathbf{R}[z]$-module. The period of v will be the monic polynomial r of least degree such that $r \cdot v = 0$. We denote by $\{v_1, \cdots, v_s\}$ the submodule spanned by the v_i's. We have a direct sum: $\mathbf{R}^k = \sum_i \{v_i\}$ and if r_i is the period of v_i, then $r_i \mid r_{i+1}$. The r_i's are the same in all such decomposition, and are called the invariants of F. They are in fact the invariants of the polynomial matrix $zI - F$ that are nontrivial (different from 1). Moreover, two matrices are conjugates if and only if they have the same invariants.

Let us go back to our systems. To each $S \in \mathbf{S}_n$, we associate its transfer function, that is the proper rational $p \times m$ matrix: $T(z) = H(zI - F)^{-1}G$ where (H, F, G) is any realization of S, and it is well-known that distinct systems give rise to distinct transfer function. Conversely, for every proper rational matrix $T(z)$, we can write

$$T(z) = H(zI - F)^{-1}G \tag{6}$$

and all triples satisfying (6) give rise to the same system via (1). However, this system may not be in $\mathbf{S}_n$. We now examine this point.

If we write $T(z) = N'D^{-1}N$ where N', D, N are polynomial matrices, the following procedure, due to Fuhrmann, lets us construct a realization of $T(z)$, that is a triple (H, F, G) such that (6) holds [10]: for any fraction p/q in $\mathbf{R}(z)$, $(p/q)_1$ denotes the coefficient of z^{-1} in the development of p/q as a formal power series: $p/q = \sum_{i>i_0} a_i z^{-i}$. If we write $p = aq + r$ by Euclidean division, we equivalently have $(p/q)_1 = (zr/q)(\infty)$. If N is a matrix, N_1 is the matrix obtained by performing the above operation on its

entries. If $D \in \mathbf{R}^{k\times k}[z]$ in our factorization of $T(z)$, we call X the quotient module of D: $\mathbf{R}^k[z]/D\mathbf{R}^k[z]$. For every x in $\mathbf{R}^k[z]$, we denote by $\underline{x}$ the class of x in X. We consider three maps:

$$
\begin{aligned}
&F: X \to X, && F(\underline{x}) = z\underline{x} (= \underline{zx} \text{ by definition}),\\
&G: \mathbf{R}^m \to X, && G(u) = \underline{Nu},\\
&H: X \to \mathbf{R}^p, && H(\underline{x}) = (N'D^{-1}x)_1
\end{aligned} \tag{7}
$$

(this last map is well defined for any element in the class of x yields the same result). By (5), $\dim_{\mathbf{R}}(X) = \deg(\det D)$, and fixing a basis in X, we can identify our maps with their matrices. It is then easy to see that (6) holds. Moreover, it is known that deg (det D) is minimal in such a factorization, if and only if (D, N), (N', D) are left and right coprime respectively, that is if we have two Bezout equations: $DA + NB = I$, $CN' + ED = I$ where A, B, C, E are polynomial matrices. Since $H(zI - F)^{-1}G$ is a factorization of the preceding type, it is clear that $T(z)$ will be associated to some $S \in \mathbf{S}_n$ if and only if deg (det D) has minimal value n among all possible factorization. This minimum is called the McMillan degree of $T(z)$, and what precedes shows that the set of $p \times m$ proper rational matrices of McMillan degree n is in bijection with $\mathbf{S}_n$. We can therefore endow it with the analytic structure described previously. However, this does not give an explicit parametrization in terms of the entries of $T(z)$, due to the involved character of (6).

Now all matrices F which occur in realizations of $S \in \mathbf{S}_n$ have the same invariants, since they are conjugates, hence we may define the cyclic structure of S as the list, l, arranged in increasing order, of the degrees $n_1, \cdots, n_r$ of its invariants. We have $\sum n_i = n$. We define $\mathbf{S}_{n,l}$ as the set of $S \in \mathbf{S}_n$ whose cyclic structure is l. The aim of this paper is to prove that $\mathbf{S}_{n,l}$ is a submanifold of $\mathbf{S}_n$, and to construct an atlas on it in terms of certain transfer-function factorizations. This does not lead to a global parametrization of $\mathbf{S}_n$, but only of submanifolds which partition it. However, it allows a rather simple description of the open and dense subset of cyclic systems (where there is only a single invariant). Our construction is based on an algebraic result, proved in the next section. In § 3 we apply it to transfer-function factorization, before finally proceeding with the above mentioned construction.

2. A factorization theorem. We will prove a result which plays an intermediate role between the classical reduction theorems for polynomial matrices quoted in the first paragraph. Everything in this paragraph goes through without change over any infinite field instead of $\mathbf{R}$ [2].

THEOREM 1. *Let M be a regular matrix in $\mathbf{R}[z]^{p\times p}$. There exist a unimodular matrix U and $R \in Gl_p$, such that $UMR = D = (d_{i,j})$ satisfies*:

i) $d_{i,j} = 0$ *if* $i > j$ (D *is upper triangular*);

ii) $d_{i,i} \mid d_{k,l}$ *for* $k \geqq i$ *and every* l;

iii) $\deg(d_{j,j}) > \deg(d_{i,j})$ *for* $i < j$ *and every* j.

Furthermore, the diagonal elements $d_{i,i}$ are the invariants of M.

Before proving Theorem 1, we will need several lemmas.

LEMMA 1. *If $r_1, \cdots, r_k$ have no common divisor in $\mathbf{R}[z]$, and r_k is nonzero, there exist $\lambda_1, \cdots, \lambda_{k-1} \in \mathbf{R}$ such that $\lambda_1 r_1 + \cdots + \lambda_{k-1} r_{k-1}$ is prime to r_k. Furthermore one of the λ_i's may be chosen arbitrarily in $\mathbf{R}$-$\{0\}$.*

Proof. Let $u_1, \cdots, u_s$ be the prime factors of r_k in $\mathbf{R}[z]$. Let V be the vector space spanned by the r_i's over $\mathbf{R}$. Let W_i be the ideal (u_i) viewed as a vector space over $\mathbf{R}$, and $V_i = W_i \cap V$. Were the first assertion not true, there would exist for any q in V an

i such that $u_i \mid q$, hence $V = \cup_i V_i$. Because a vector space over an infinite field cannot be a finite union of proper subspaces (cf. e.g. [17]), we have $V = V_{i_0}$ for some i_0, and u_{i_0} is a common divisor to the r_i's, a contradiction. For the second assertion, it is sufficient to show that we may choose $\lambda_1 = 1$. Suppose not, and fix some λ_i's for $1 < i$. For $t \in \mathbf{R}$, define $Q(t) = t(\sum_{i>1} \lambda_i r_i) + r_1$; $Q(t)$ is never prime to r_k. The u_i's being finite in number, one of them divides $Q(t)$ and $Q(t')$ for distinct t t', and by difference $\sum_{i>1} \lambda_i r_i$. But by the first part of the proof some combination of the r_i's is prime to r_k, and we see now that λ_1 can be neither 0 nor 1, a contradiction. Q.E.D.

LEMMA 2. *Let* $D = \begin{bmatrix} d_1 & d_2 \\ 0 & d_3 \end{bmatrix}$ *be a* 2×2 *regular polynomial matrix. There exist unimodular matrices* U *and* $R \in Gl_2$ *with* $\det(R) = 1$, *such that*

$$UDR = \begin{bmatrix} \mu & \lambda \\ 0 & d_1 d_3/\mu \end{bmatrix} \quad \text{where } \mu = \text{g.c.d.}\,(d_1, d_2, d_3).$$

Proof. Let us write

$$U = \begin{bmatrix} \alpha & \beta \\ \delta & \gamma \end{bmatrix}, \qquad R = \begin{bmatrix} x & y \\ z & t \end{bmatrix}.$$

Computing UDR, we find

$$UDR = \begin{bmatrix} x\alpha d_1 + z(\alpha d_2 + \beta d_3) & y\alpha d_1 + t(\alpha d_2 + \beta d_3) \\ x\delta d_1 + z(\delta d_2 + \gamma d_3) & y\delta d_1 + t(\delta d_2 + \gamma d_3) \end{bmatrix};$$

we put $d_i = d_i' \mu$ and choose $\delta = -z d_3'$, $\gamma = x d_1' + z d_2'$, so that UDR is upper triangular. By Lemma 1, we may choose x and a nonzero z, such that $x d_1' + z d_2'$ is prime to d_3'. Then δ and γ are coprime, and we choose α and β such that $\det(U) = 1$. The first entry of U is now easily computed to be μ. Finally, choosing for example $t = 0$, $y = -1/z$, we have $\det(R) = 1$. Q.E.D.

From now on, we shall, for notational simplicity, call a (U, R) transformation a left multiplication by U (unimodular) and a right multiplication by R (regular).

LEMMA 3. *If* $D = (d_{i,j})$ *is an upper triangular regular polynomial matrix, and if for some* i $d_{i,i}$ *does not divide all the elements of its row, there exists a* (U, R) *transformation such that if we write* $D' = (d'_{i,j}) = UDR$, D' *will still be upper triangular, with* $\deg(d'_{i,i}) < \deg(d_{i,i})$, $d'_{k,k} = d_{k,k}$ *if* $k < i$. *Moreover, if* $k < i$, *each* $d'_{k,l}$ *will be a linear combination of the* $d_{k,j}$'s.

Proof. Suppose $d_{i,i}$ does not divide $d_{i,k}$. Exchanging the columns $i+1$ and k, we can recover a triangular matrix by the Hermite procedure mentioned in § 1 (type U transformation), hence we may suppose $i + 1 = k$. Applying Lemma 2 to the matrix

$$\begin{bmatrix} d_{i,i} & d_{i,i+1} \\ 0 & d_{i+1,i+1} \end{bmatrix}$$

we find two matrices U_2 and R_2. Putting

$$U = \begin{bmatrix} I & 0 & 0 \\ 0 & U_2 & 0 \\ 0 & 0 & I \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} I & 0 & 0 \\ 0 & R_2 & 0 \\ 0 & 0 & I \end{bmatrix}$$

where the I's are identity matrices of appropriate sizes, we define a (U, R) transformation satisfying our requirements. Q.E.D.

LEMMA 4. *Let* D *be as in Lemma* 3, *but suppose this time that* $d_{i,i}$ *does not divide every* $d_{j,j}$ *for* $j > i$. *The conclusion of Lemma* 3 *remains valid.*

Proof. Let k be the smallest index $> i$, such that $d_{i,i}$ does not divide $d_{k,k}$. Then by definition, $d_{k-1,k-1}$ does not divide $d_{k,k}$. Treating the matrix

$$D=\begin{bmatrix} d_{k-1,k-1} & d_{k-1,k} \\ 0 & d_{k,k} \end{bmatrix}$$

as we did in Lemma 3, we obtain D' such that $d'_{k-1,k-1} | d_{k,k}$, $\deg(d'_{k-1,k-1}) < \deg(d_{k-1,k-1})$. If $k=i+1$, we are done. If not, $d'_{i,i}=d_{i,i}$ does not divide $d'_{k-1,k-1}$, and we can replace k by $k-1$. In a finite number of steps we are left to the case $k=i+1$. Q.E.D.

LEMMA 5. *If D is as before, there exists a (U, R) transformation such that $D'=UDR$ satisfies $d'_{1,1} | d'_{i,j}$ for every i, j.*

Proof. Applying Lemma 3, we may assume $d_{1,1} | d_{1,l}$ for every l, because the degree of $d'_{1,1}$ decreases at each step. Applying this lemma to the second row, we may assume that $d_{2,2} | d_{2,l}$, for every l, and still $d_{1,1} | d_{1,l}$, because the $d_{1,l}$ are replaced by linear combinations of themselves. Continuing that way, we may assume $d_{i,i} | d_{i,l}$ for every i and l. If then $d_{1,1}$ does not divide $d_{i,i}$, we apply Lemma 4, to get a $d_{1,1}$ of lower degree. The procedure stops in a finite number of steps. Q.E.D.

Proof of Theorem 1. By the Hermite procedure we can assume $M=(m_{i,j})$ to be upper triangular. We use induction on the size p of M. If $p=1$, there is nothing to prove. If not, we can assume by Lemma 5 that $m_{1,1} | m_{i,j}$. We define $\underline{M}$ to be the $(p-1, p-1)$ matrix obtained by deleting the first row and column of M. By the induction hypothesis we associate with $\underline{M}$ two matrices $\underline{U}$ and $\underline{R}$. Putting

$$U=\begin{bmatrix} 1 & 0 \\ 0 & \underline{U} \end{bmatrix} \quad \text{and} \quad R=\begin{bmatrix} 1 & 0 \\ 0 & \underline{R} \end{bmatrix},$$

we see that all our requirements are satisfied, except possibly condition iii), but it is an easy consequence of euclidean division (in fact the basis of Hermite procedure), that we can ensure this by U-type transformation without losing ii). The $d_{i,i}$'s are the invariants of M, because we can write $D=\operatorname{diag}\{d_{i,i}\}V$, with V upper triangular with 1's on the diagonal, hence unimodular. Q.E.D.

Let us observe that if $d_{i,i}=1$, then $d_{k,i}=0$ for $k<i$ by condition iii). A particularly simple case is the cyclic one (only $d_{p,p}$ may differ from 1). The matrix D has then the following form

$$D=\begin{bmatrix} 1 & & & a_1 \\ & 1 & O & \vdots \\ & & \ddots & \vdots \\ O & & & a_p \end{bmatrix}$$

with $\deg(a_i) < \deg(a_p)$ for $i<p$. Note that D^{-1} is then easily computed: it is obtained by replacing a_i by $-a_i/a_p$ for $i<p$ in the above expression of D, and a_p by $1/a_p$. More generally, if D is under the form of the theorem, the element (i,j) of D^{-1} may be written $P_{i,j}/d_{j,j}$, with $P_{i,j} \in \mathbf{R}[z]$, as follows easily from condition ii) by induction on the formulae $\sum_k d_{i,k} d^{-1}_{k,j} = \delta_{i,j}$. In the next paragraph, $P_{i,j}$ will have the above meaning (an explicit expression for it can be found in [2] but this is of no use in the sequel).

Let us introduce some notations, that will be of use in the rest of the paper. When dealing with a matrix D, we will call it special if it is of the preceding type. According to ii), we will then denote $d_{i,j}=q_{i,j}d_{i,i}$, and we will write d_i instead of $d_{i,i}$.

It is clear that Theorem 1 can be stated in a transposed form, finding U^*, R^*, such that $D^* = R^*MU^* = (d^*_{i,j})$ is under a form transposed of that of D, that is: D^* is lower triangular, conditions ii) and iii), being replaced by

ii)* $d^*_{i,i} | d^*_{k,l}$ for $l \geqq i$ and every k;

iii)* $\deg(d^*_{j,j}) > \deg(d^*_{j,i})$ for $i < j$ and every j.

These forms will be called t-special and will be denoted with "*" as superscript. We shall write as above $d^*_{i,j} = q^*_{i,j} d^*_{j,j}$ and d^*_i for $d^*_{i,i}$.

Examples. Consider the matrix

$$D = \begin{bmatrix} 1 & z+1 & z^2+1 \\ 2 & 4z & 4z^2-z-6 \\ 0 & z-1 & z^2-1 \end{bmatrix}.$$

If we set

$$U = \begin{bmatrix} 20z+81 & -10z-40 & 20z+69 \\ -4z-16 & 2z+8 & -4z-14 \\ z^2+2z-7 & -z^2/2-z+7/2 & z^2+3z/2-6 \end{bmatrix}$$

and

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -2 \\ 0 & 1/2 & 0 \end{bmatrix},$$

U is unimodular and we have

$$UDR = \begin{bmatrix} 1 & 0 & 20z-24 \\ 0 & 1 & 4-4z \\ 0 & 0 & z^2-3z+2 \end{bmatrix}$$

and D is cyclic. Since we also have

$$\begin{bmatrix} -4z-3 & 2z+2 & -4z-3 \\ -z-1 & z/2+1/2 & -z-1/2 \\ -2z^2+2z+8 & z^2-z-4 & -2z^2+3z+6 \end{bmatrix} D \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2z \\ 0 & 1 & z/2-1/2 \\ 0 & 0 & z^2-3z+2 \end{bmatrix},$$

we see that a special form associated with D is not unique; in fact, it is not difficult to see that infinitely many such forms exist.

Consider now

$$D = \begin{bmatrix} z^2+3z+2 & z^2+3z+2 & z^2+2z+1 \\ z^2+3z-1 & z^2+2z+1 & z^2+2z-1 \\ z^2+2z+4 & z+1 & z^2+2z+3 \end{bmatrix}.$$

Setting

$$U = \begin{bmatrix} z & -z-1 & 1 \\ 1-z & z & -1 \\ -1 & 1 & 1 \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix},$$

U is unimodular, we have

$$UDR = \begin{bmatrix} 1 & 0 & 2z+4 \\ 0 & z+1 & -2(z+1) \\ 0 & 0 & (z+1)^2 \end{bmatrix},$$

and D has then 2 invariants.

3. Application to transfer function factorization.

A. General case. Among the factorizations described in § 1, we consider those of the type

$$T = D^{-1}N \tag{8}$$

where T is the transfer function of some $S \in \mathbf{S}_n$, D, N are left coprime polynomial matrices. It is well-known that for given T, such a thing exists. Applying Theorem 1 to the matrix D allows us to deduce from (8) another factorization: $T = RD^{-1}N$ where D is now in a special form, R being in Gl_p. Such a factorization will be called special. The diagonal elements of D are then the invariants of the system, and the list of their degrees its cyclic structure. Now we take into account the properness of T. Because R is real, this is equivalent to the properness of $D^{-1}N$. If we consider that D is given, this will impose some structure for each column of N. We first introduce two more notations: if $s/q \in \mathbf{R}(z)$, and $s = aq + r$ is euclidean division, then $s/q = a + r/q$. We define $E(s/q) = a$, $PP(s/q) = r/q$, and call them respectively the *entire* and *proper* parts of s/q. Putting now $N = (n_{i,j})$, $M_{i,k} = d_k/d_i$ for $k > i$ (so that $M_{i,k}$ is polynomial), the properness of the product of D^{-1} by the jth column of N means simply that $(n_{i,j} + \sum_{k=i+1}^{p} n_{k,j} P_{i,k}/M_{i,k})/d_i$ is a proper fraction for all i. Defining $f_i(x_{i+1}, \cdots, x_p) = \sum_{k=i+1}^{p} x_k P_{i,k}/M_{i,k}$ ($f_p = 0$ conventionally), we see that

$$n_{i,j} = -E(f_i(n_{i+1,j}, \cdots, n_{p,j})) + r_{i,j} \tag{9}$$

where $r_{i,j} \in \mathbf{R}[z]$ is such that $\deg(r_{i,j}) < \deg(d_i)$ with $\deg(0) = -\infty$. Conversely, given any set $(r_{i,j})$ of polynomials satisfying the above degree conditions, there is a unique matrix N, defined recursively by (9), such that $D^{-1}N$ is proper; we then have the following proposition.

PROPOSITION 1. *If D is in special form, with the preceding notation, the set V_D of $p \times m$ polynomial matrices N such that $D^{-1}N$ is proper is a nm-dimensional vector space over $\mathbf{R}$. If $N = (n_{i,j})$ is in V_D, the coefficients of the $r_{i,j}$'s defined by* (9) *are coordinates in V_D, and the jth column of $D^{-1}N$ is*

$$\begin{bmatrix} (r_{1,j} + PP(f_1(n_{2,j}, \cdots, n_{p,j})))/d_1 \\ \vdots \\ (r_{p,j})/d_p \end{bmatrix}.$$

Proof. Putting $r_{i,j} = \delta_{k,i}\delta_{l,j}z^t$ (where δ is the Kronecker symbol), for $1 \leqq k \leqq p$, $1 \leqq l \leqq m$, $0 \leqq t < \deg(d_k)$ in (9), we define matrices which are easily seen to form a basis of V_D. The coordinates of any matrix in this basis are obviously the coefficients of its r_{ij}'s. The expression for $D^{-1}N$ follows from (9). Q.E.D.

If we analyse the causality of a product $N^*(D^*)^{-1}R^*$ where D^* is in t-special form, we get of course transposed results, introducing parameters $r^*_{i,j}$ such that $\deg(r^*_{i,j}) < \deg(d^*_j)$. Such a factorization will be called t-special. The space V_{D^*}, dual to V_D, is then pn-dimensional.

Now, our factorization $T = RD^{-1}N$ was minimal by definition. Conversely, we want to know when this is the case for a given special factorization. It is the object of the next theorem.

THEOREM 2. *Let D, N be as above, and r be the number of nontrivial invariants of D. Let Δ_i be the g.c.d. of the minors constructed with the last i rows of N. D, N are left coprime if and only if Δ_i is prime to d_{p-i+1} for $1 \leqq i \leqq r$.*

Proof. We first note that N must have at least r columns, for the state module, isomorphic to the quotient module of D: $X = \mathbf{R}^p[z]/D\mathbf{R}^p[z]$ (cf. § 1) cannot be spanned by less than r elements, so that the assertion of the theorem makes sense. We make then two remarks:

1) (D, N left coprime) is equivalent to (D, NU left coprime for some unimodular matrix U). This is obvious from Bezout equation.

2) If T is an upper triangular unimodular matrix, (so that its diagonal elements are real numbers), (D, N left coprime) is also equivalent to (D, TN left coprime).

To see this we note—a fact that will be sometimes tacitly used in the sequel—that $D\mathbf{R}^p[z]$ is made of those polynomial vectors whose ith component is multiple of d_i. This implies easily, according to the form of D, and T, that TD may be written as DU. Taking determinants, we see that U is unimodular. Multiplying some Bezout equation between D and N by T on the left and T^{-1} on the right yields then the result.

We go back to our theorem, and prove sufficiency. By Remark 1, we may assume N is in Hermite form (cf. § 1):

$$N = \begin{bmatrix} & X & & \cdots \\ & & n_{p-1} & \\ n_p & & & \end{bmatrix}$$

where $n_{p-i+1} = \Delta_i/\Delta_{i-1}$ for $1 \leqq i \leqq r$ ($\Delta_0 = 1$). Given $Y \in \mathbf{R}^p[z]$, we consider the equation

$$NX = Y \bmod \begin{bmatrix} d_1 \\ \vdots \\ d_p \end{bmatrix}$$

where X is to be found in $\mathbf{R}^m[z]$, the ith equation being in $\mathbf{R}[z]/(d_i)$. If $d_i = 1$, the corresponding equation is trivial, so we need only consider the case when $i > p - r$. But n_i is then prime to d_i by hypothesis, hence invertible in $\mathbf{R}[z]/(d_i)$, and the equation is solvable, proceeding recursively from the last row up to the first one. Solving $NA = I + W$ (where W has its ith row divisible by d_i, hence may be written DB for some polynomial matrix B) yields a Bezout equation.

Suppose conversely D, N left coprime, N as above. If $\mu = \text{g.c.d.}\,(n_p, d_p)$, we see that the matrix $\left[\begin{smallmatrix} I & 0 \\ 0 & \mu \end{smallmatrix}\right]$ is a common left factor to D and N, and therefore μ must be constant. Hence n_p, d_p are coprime, and $n_p a + d_p b = c$ is solvable in a and b for every c in $\mathbf{R}[z]$. This implies that by left multiplication by a matrix of the type described in remark 2, we can make every element of the first column of $\mathbf{N}$ (except n_p) a multiple of d_p. Now if n_{p-1} (which has remained unchanged under the preceding operation) is not prime to d_{p-1}, and μ is their g.c.d., the matrix

$$\begin{bmatrix} I & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

is again a nontrivial common left factor to D and N. Continuing this procedure yields

n_i prime to d_i for every i. This shows, by the divisibility conditions $d_i | d_{i+1}$, that Δ_i is prime to d_{p-i+1} for $1 \leqq i \leqq r$. Q.E.D.

We now study what matrices R can occur in special factorizations of a given T. For this, the next two propositions will be technical results. From here on, our special (t-special) factorizations are always assumed to be coprime, and r will always denote the number of nontrivial invariants under consideration, as in theorem 2. Given a special factorization $D^{-1}N$, we investigate whether it may be written as a t-special factorization $N^*(D^*)^{-1}$. When it actually is the case, by realization theory (§ 1), D and D^* have the same nontrivial invariants (although their numbers of invariants "1" differ if m is not equal to p).

PROPOSITION 2. *If κ_i is the minor constructed with the last i rows and columns of N, a sufficient condition for*

$$D^{-1}N = N^*(D^*)^{-1} \tag{10}$$

to be solvable is:

$$g.c.d.\,(\kappa_i,\ d_{p-i+1}/d_{p-i}) = 1 \quad \text{for } 1 \leqq i \leqq r \tag{11}$$

where $d_k = 1$ if $k \leqq 0$ by convention.

(*The condition is also necessary* [2], *but this will not be used here.*)

Proof. Equation (10) can be read $ND^* = DN^*$. Had we found a solution (N^*, D^*) where D^* satisfies all our requirements except iii)*, right multiplying both matrices by an appropriate unimodular matrix, as in the end of the proof of Theorem 1, yields a complete solution. Hence we can forget about the degree conditions. Now (cf. Remark 2) the problem is to find D^* such that the ith row of ND^* is divisible by d_i, hence we may assume $p = r$. We must have $d^*_{m-j} = d_{r-j}$ if $0 \leqq j < r$, $d^*_j = 1$ otherwise, and we are left to choose the $q^*_{i,j}$'s, and necessarily $q^*_{i,j} = 0$ if $i < m - r + 1$. Because d^*_j divides the jth column of ND^*, the problem reduces to those elements $x_{i,j}$ of ND^* with $r - i < m - j$. We may therefore assume $m > 1$. For $0 \leqq i < r$ and $0 \leqq j < m$, we let $k = \inf(j-1, r-1)$, and we have then:

$$x_{r-i,m-j} = d^*_{m-j}\left(\sum_{s=m-k}^{m} n_{r-i,s}q^*_{s,m-j} + n_{r-i,m-j}\right).$$

The equations to be solved may be written, as in the proof of Theorem 2, as a linear system of congruences:

$$\begin{bmatrix} n_{r-k,m-k} & \cdots & n_{r-k,m} \\ \vdots & & \vdots \\ n_{r,m-k} & \cdots & n_{r,m} \end{bmatrix} \begin{bmatrix} q^*_{m-k,m-j} \\ \vdots \\ q^*_{m,m-j} \end{bmatrix} = -\begin{bmatrix} n_{r-k,m-j} \\ \vdots \\ n_{r,m-j} \end{bmatrix} \bmod \begin{bmatrix} d_{r-k}/d^*_{m-j} \\ \vdots \\ d_r/d^*_{m-j} \end{bmatrix}$$

for $1 \leqq j < m$. Note that since $d^*_{m-j} = d_{r-(k+1)}$, the fractions inside the modulo symbol are actually polynomials. We rewrite the above equation under the more compact form

$$N_{k+1}q^*_j = p_j \bmod (d_{r-k}/d^*_{m-j}, \cdots, d_r/d^*_{m-j})$$

where the meaning of our new notation is obvious. Since, by hypothesis, $\det(N_{k+1})$ is prime to d_{r-k}/d^*_{m-j} (and is therefore a unit in $\mathbf{R}[z]/(d_{r-k}/d^*_{m-j})$), we can find X_{k+1} such that $N_{k+1}X_{k+1} = p_j \bmod (d_{r-k}/d^*_{m-j}, \cdots, d_{r-k}/d^*_{m-j})$, but only the first modulo is then correct. Thus we have $N_{k+1}X_{k+1} = p_j + Wd_{r-k}/d^*_{m-j}$, where $W \in \mathbf{R}^{k+1}[z]$. If $k = 0$, we are done. Otherwise, let $\mathbf{W}$ be the vector of $\mathbf{R}^k[z]$ obtained from W by deleting the first component. Again by hypothesis, we find X_k such that $N_kX_k =$

$-\mathbf{W}+W'd_{r-(k-1)}/d_{r-k}$; we put

$$Y_k = X_k d_{r-k}/d^*_{m-j} \quad \text{and} \quad X'_{k+1} = X_{k+1} + \begin{bmatrix} 0 \\ Y_k \end{bmatrix}.$$

Now we see that $N_{k+1}X'_{k+1} = p_j \bmod (d_{r-k}/d^*_{m-j}, d_{r-(k-1)}/d^*_{m-j}, \cdots, d_{r-(k-1)}/d^*_{m-j})$; in other words the second modulo has become correct too. Continuing in this fashion, we solve the equation, because at each step, thanks to the divisibility conditions between the d_i's, the moduli that are already correct remain so. Q.E.D.

Because (11) remains unchanged under transposition, the condition on N^* for (10) to be solvable (N^* and D^* being now given) is the same, replacing p by m. In both cases, it will be referred to as the minors condition.

The next proposition relates the minors condition to an intrinsic property of the system underlying a special factorization. We introduce one more definition: in a $\mathbf{R}[z]$-module, we define the period of an element v modulo a submodule M as the monic polynomial q of smallest degree such that $qv \in M$ (it is actually the period in the quotient module, and it exists at least when our module is finite-dimensional over $\mathbf{R}$).

PROPOSITION 3. *Let T be the transfer function of $S \in \mathbf{S}_n$, and $p_1, \cdots, p_r$ be its invariants. The following conditions are equivalent*:

a) *for some realization (H, F, G) of S, the columns $(g_1, \cdots, g_m)$ of G are such that in the state module $(\mathbf{R}^n$ under the action of F cf.* § 1):

$-g_m$ *has period* p_r,

$-g_{m-1}$ *has period* $p_{r-1} \bmod \{g_m\}$,

$\cdots$

$-g_{m-(r-1)}$ *has period* $p_1 \bmod \{g_{m-(r-2)}, \cdots, g_m\}$;

b) *for any realization of S,* a) *holds*;

c) *for some special factorization $T = RD^{-1}N$, the minors condition holds*;

d) *for any special factorization,* c) *holds.*

Proof. a) is obviously equivalent to b), because it does not depend on the basis chosen in the state space; it is also obvious that d) implies c). Let us prove that a) implies d): we first note that a) implies by induction

$$\dim_{\mathbf{R}} \{g_{m-i}, \cdots, g_m\} = \sum_{j=0}^{i} \deg (p_{r-j}) \quad \text{for } 0 \leqq i \leqq r-1 \tag{12}$$

and let us denote the above integer by k_i. Let $T = RD^{-1}N$ be a special factorization, and consider the Fuhrmann triple associated with it, in any $\mathbf{R}$-basis of the state-module X (the quotient module of D cf. § 1, (7)). The column vectors of G represent then the equivalence classes, in X, of the corresponding columns of N. Let $M_i = (N_{m-i}, \cdots, N_m)$ be the matrix made of the last $i+1$ columns of N, and $\underline{\mathbf{M}}_i$ the square submatrix of M_i made of its last $i+1$ rows. By (12), the columns of M_i span a submodule V_i of X whose dimension over $\mathbf{R}$ is k_i. Right multiplying M_i by a unimodular matrix, we put it under Hermite form (§ 1), so that $\underline{\mathbf{M}}_i$ now reads:

$$\underline{\mathbf{M}}_i = \begin{bmatrix} & & t_m \\ & X & \\ & \iddots & \\ t_{m-i} & & O \end{bmatrix}.$$

Note that neither V_i, nor $\det(\underline{\mathbf{M}}_i)$—up to the multiplication by a nonzero real number—has changed. Now it is clear (cf. Remark 2, Proof of Theorem 2) that p_{r-j}, $0 \leqq j \leqq i$, is an annihilator (i.e. a multiple of the period) for the $(j+1)$th column of M_i. But $\dim_{\mathbf{R}} V_i$ is equal at most to the sum of the dimensions of the submodules spanned by each column of M_i, that is the sum of the degrees of the periods. This last number cannot therefore be less than k_i, so that p_{r-j} is actually the period of the $(j+1)$th column.

Let $\mu = \text{g.c.d.}\,(t_m, p_{r-i}/p_{r-(i+1)})$ where, we recall, $p_0 = 1$ conventionally. Then p_{r-i}/μ is an annihilator of the $(i+1)$th column of M_i, hence by what precedes, $\mu = 1$. Now, using a device already mentioned in the proof of Theorem 2, we can ensure that $p_{r-i}/p_{r-(i+1)}$ divides every element of the first row of $\underline{\mathbf{M}}_i$, except t_m. Let now $\nu = \text{g.c.d.}\,(t_{m-1}, p_{r-(i-1)}/p_{r-(i+1)})$. Then $p_{r-(i-1)}/\nu$ is an annihilator of the ith column of M_i, because it writes $p_{r-(i+1)}p_{r-(i-1)}/(\nu p_{r-(i+1)})$. Therefore $\nu = 1$. Again we can ensure that $p_{r-(i-1)}/p_{r-(i+1)}$ divides every element of the second row of $\underline{\mathbf{M}}_i$, except t_{m-1}. Continuing in this way, we prove that t_{m-j} is prime to $p_{r-(i-j)}/p_{r-(i+1)}$. The product $\prod_{j=0}^{i} t_{m-j} = \det(\underline{\mathbf{M}}_i)$ is a fortiori prime to $p_{r-i}/p_{r-(i+1)}$. But this is precisely the minors condition.

We now prove that c) implies a). We consider again the Fuhrmann triple associated with our special factorization, in some basis (over $\mathbf{R}$) of the state module, and write as usual $N = (n_{i,j})$, and N_i for the ith column of N. As before, g_i represent the equivalence class of N_i in X for $1 \leqq i \leqq m$.

Now p_r is an annihilator for g_m, since it annihilates X itself. Let q be the period of g_m (whence $q \mid p_r$), and $\mu = p_r/q$. Then by definition of q: $p_{r-j} \mid q n_{p-j,m}$ for $0 \leqq j \leqq r-1$, or equivalently:

$$\mu \mid n_{p-j,m}(p_r/p_{r-j}). \tag{13}$$

In particular, $\mu \mid n_{p,m}$, but $n_{p,m}$ is prime to p_r/p_{r-1} (minors condition), hence μ is prime to p_r/p_{r-1} also, and since $\mu \mid p_r$, we see that $\mu \mid p_{r-1}$. If $r = 1$, then $p_{r-1} = 1$ conventionally, hence $\mu = 1$. Otherwise, suppose by induction that $\mu \mid n_{p-k,m}$ for $0 \leqq k < j$, and that μ is prime to p_r/p_{r-j}, for some $j < r$. By (13), μ divides $n_{p-j,m}$, and with the notations of Proposition 2, μ divides the minor κ_{j+1}, whence μ is prime to $p_{r-j}/p_{r-(j+1)}$ by the minors condition. Therefore μ is prime to $p_r/p_{r-(j+1)}$. This shows by induction that μ is prime to p_r, hence $\mu = 1$, thereby proving that g_m has period p_r. If $r = 1$, we are done. We now proceed to prove a) in the case of a column g_{m-i}, where i satisfies $1 \leqq i \leqq r-1$. Using the notations of the first part of the proof, we consider the matrix $M_{(i-1)}$ and its submatrix $\underline{\mathbf{M}}_{(i-1)}$. Right multiplying as before $M_{(i-1)}$ by a unimodular matrix V, we get a new matrix $M'_{(i-1)}$ with $\underline{\mathbf{M}}'_{(i-1)}$ under Hermite form, its diagonal element in the $(i-j)$th column being, as before, denoted by t_{m-j}. Now the product $\prod_{j=k}^{i-1} t_{m-j}$, where $0 \leqq k \leqq i-1$, is the g.c.d. of the $(i-k)$-rowed minors built with the last $(i-k)$ rows of $\underline{\mathbf{M}}_{(i-1)}$ (§ 1), hence divides κ_s for $i-k \leqq s \leqq i$; by the minors condition, it must be prime to each $p_{r-(s-1)}/p_{r-s}$, and therefore to their product $p_{r-(i-k-1)}/p_{r-i}$. Then t_{m-k} is a fortiori prime to $p_{r-(i-k-1)}/p_{r-i}$. Using a now standard procedure, we can add to N_{m-i} a combination of the columns of $M'_{(i-1)}$, (hence also of $M_{(i-1)}$), to get a polynomial vector N'_{m-i} whose $(p-j)$th component is a multiple of p_{r-j}/p_{r-i}, for $0 \leqq j \leqq i-1$. Now, since they are congruent modulo V_{i-1}, we can replace N_{m-i} by N'_{m-i} in our proof, and it is clear that this does not modify the minors κ_j. We observe that p_{r-i} is now an annihilator for N_{m-i}; let q be its period and $\mu = p_{r-i}/q$. Then, denoting $n_{j,m-i}$ by n_j for simplicity

$$p_{r-j} \mid q n_{p-j} \quad \text{for } 0 \leqq j \leqq r-1, \tag{14}$$

whence for $j \leqq i$, $p_{r-i} \mid q n_{p-j}$, and μ divides n_{p-j}. We can now use the same argument

as we did in the proof that g_m has period p_r: from (14) we get

$$\mu \mid n_{p-j} p_{r-i}/p_{r-j} \quad \text{for } i \leqq j \leqq r-1. \tag{15}$$

On the other hand, we have seen that μ divides each n_{p-j} for $j \leqq i$, hence divides κ_{i+1}, and so it is prime to $p_{r-i}/p_{r-(i+1)}$ by the minors condition. Since it divides p_{r-i} by definition, it must divide $p_{r-(i+1)}$. If $i = r-1$, this implies $\mu = 1$. Otherwise, if we suppose that for some j, $i \leqq j < r-1$ we have $\mu \mid n_{p-k}$ for $0 \leqq k \leqq j$ and μ prime to p_{r-i}/p_{r-j}, then $\mu \mid \kappa_{j+1}$, hence is prime to $p_{r-j}/p_{r-(j+1)}$ by the minors condition, hence also to $p_{r-i}/p_{r-(j+1)}$, and by (15) $\mu \mid n_{p-(j+1)}$. Therefore, by induction, we conclude that $\mu \mid p_0$, that is $\mu = 1$. Q.E.D.

Considering Proposition 3, it makes sense to say that $S \in \mathbf{S}_n$ verifies the minors condition. When working with t-special factorizations, we will say that S verifies the dual minors condition, that is to say its tranpose verifies the minors condition. The final result in this section is Theorem 3.

THEOREM 3. *If T is the transfer function of $S \in \mathbf{S}_n$, the set O of all $P \in Gl_m$ such that TP verifies the minors condition is an open dense subset of $\mathbf{R}^{m \times m}$. A t-special factorization $T = N^*(D^*)^{-1}P^{-1}$ exists if and only if $P \in O$.*

Proof. Let $T = RD^{-1}N$ be a special factorization. The minors condition between D and NP can be expressed by saying that some polynomials (the resultants) in the coefficients of the d_i's (the invariants), the coefficients of the $n_{i,j}$'s, and those of P, are nonzero. This shows that O is the complement in Gl_m of the set of zeroes of some polynomial, hence is open, and will be dense if nonvoid. To show this, we construct a matrix P: we consider some realization (H, F, G) of S, and an invariant decomposition of the state module (§ 1) $X = \sum_{i=1}^{r} \{v_i\}$ where v_i has period p_i. Writing $G = (g_1, \cdots, g_m)$ in columns, we have $g_i = \sum_k Q_i^k v_k$ where $Q_i^k \in \mathbf{R}[z]$. But the g_i's span the state module too (reachability), whence $v_r = \sum_i R_i g_i$ with $R_i \in \mathbf{R}[z]$. Substituting g_i, we get $v_r = (\sum_i R_i Q_i^r) v_r + x$ where $x \in \{v_1, \cdots, v_{r-1}\}$ (this is interpreted as the null submodule if $r = 1$). By direct sum we get $x = 0$, and $\sum R_i Q_i^r = 1$ modulo p_r. By Lemma 1, § 2, we find $(\lambda_i)_{1 \leqq i \leqq m}$ in $\mathbf{R}$ such that $\sum \lambda_i Q_i^r$ is prime to p_r, and $\lambda_m = 1$. The vector $g'_m = \sum \lambda_i g_i$ then has period p_r, for its component on v_r is prime to p_r. We choose (λ_i) as mth column for P. If $r = 1$, we put 1's on the diagonal and 0's elsewhere. Otherwise, the quotient module $Y = X/\{g'_m\}$ is easily seen to be isomorphic to the direct sum $\sum_{i=1}^{r-1} \{v_i\}$, and generated by the classes of $g_1, \cdots, g_{m-1}$. We define $\underline{G} = (g_1, \cdots, g_{m-1})$, and we may suppose by induction on m (note that when $m = 1$, then $r = 1$) that we have constructed a $(m-1) \times (m-1)$ upper triangular matrix $\underline{P}$, with 1's on the diagonal, such that the classes of the columns vectors g'_i of $\underline{G}\underline{P}$ satisfy a) of Proposition 3 in Y (whose invariants are $p_1, \cdots, p_{r-1}$). Then taking the matrix $\begin{bmatrix} \underline{P} \\ 0 \end{bmatrix}$ for the $m-1$ first columns of P yields an upper triangular matrix, with 1's on the diagonal (hence regular), such that the matrix $G' = GP$ satisfies condition a) of Proposition 3 in X. Therefore, TP satisfies the minors condition, and $P \in O$.

Now if $P \in O$, and if we write $T = RD^{-1}N$, then $D^{-1}(NP)$ satisfies the minors condition, hence may be written $N^*(D^*)^{-1}$ by Proposition 2. Considering RN^* as a new matrix N^* yields the desired t-special factorization of T. Suppose conversely that $TP = N^*(D^*)^{-1}$. We denote by (e_i) the canonical basis of $\mathbf{R}^m$, and by $\underline{x}$ the class of x in $\mathbf{R}^m[z]/D^*\mathbf{R}^m[z]$. Let p_i be the period of $\underline{e}_i$ modulo $\{\underline{e}_{i+1}, \cdots, \underline{e}_m\}$. Then $p_i e_i + \sum_{k>i} r_k e_k \in D^*\mathbf{R}^m[z]$, for some polynomials r_k. It is however obvious from the t-special form of D^* that this implies $d_i \mid p_i$. Since $d_i(e_i + \sum_{k>i} q^*_{k,i} e_k)$ is a left multiple of D^* (it is its ith column), we see that $p_i \mid d_i$, so $p_i = d_i$. This shows that for any Fuhrmann realization of $N^*(D^*)^{-1}$, condition a) of Proposition 3 holds. Therefore, TP satisfies the minors conditions. Q.E.D.

Note that proving that O is nonempty, could be done simply by showing that the resultant polynomial mentioned at the beginning of the proof is not the zero polynomial. The present proof however is constructive, and also shows easily that P may be chosen triangular as we did, or orthogonal, symmetric etc. By transposition, Theorem 3 gives conditions on a matrix R to allow a special factorization T, and combining both kind of factorization we can ensure that two matrices R and R^* allow a symmetric factorization $T = RD^{-1}NR^*$ where $D^{-1}N = N^*(D^*)^{-1}$ is solvable [2]. This, however, will not be used here.

Example. Consider the transfer function

$$T = \begin{bmatrix} (z^2+5z+2)/(z+1)^3 & (z-1)/(z+1)^2 \\ z^2/(z+1)^3 & 1/(z+1)^2 \end{bmatrix}.$$

It admits the special factorization $T = D^{-1}N$, where

$$D = \begin{bmatrix} z+1 & (z+1)(z+3) \\ 0 & (z+1)^3 \end{bmatrix}, \qquad N = \begin{bmatrix} z+2 & 2 \\ z^2 & z+1 \end{bmatrix}.$$

In accordance with Proposition 1, this may be rewritten as

$$T = \begin{bmatrix} 1/(z+1) & -(z+3)/(z+1)^3 \\ 0 & 1/(z+1)^3 \end{bmatrix} \cdot \begin{bmatrix} 1-E(-(z+3)z^2/(z+1)^2) & 1-E(-(z+3)(z+1)/(z+1)^2) \\ z^2 & z+1 \end{bmatrix}.$$

Since $\det N = -z^2+3z+2$ while z^2 and $z+1$ are coprime, we see from Theorem 2 that our factorization is minimal. The minors condition is clearly not satisfied, so that by Theorem 3 a t-special factorization of T with $R^* = I$ does not exist. However, it is easily seen that $\left[\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right]$ belongs to O, and as expected from Theorem 3, we have a t-special factorization:

$$T = \begin{bmatrix} 2z+2 & z^2+5z+1 \\ z & z^2+z+1 \end{bmatrix} \begin{bmatrix} z+1 & 0 \\ z(z+1) & (z+1)^3 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}^{-1}.$$

Because the results proved in this section take a simpler form when the system is cyclic, we rephrase them separately in this special (but generic, cf. § 4) case. This is done in the next subsection.

B. Cyclic case. We first recall that a cyclic system (resp. matrix) is one with a single (nontrivial) invariant. A cyclic special form differs from an identity matrix only by its last column, which will be, as in § 1, denoted ${}^t(a_1, \cdots, a_p)$; a_p is then the invariant polynomial. In the remainder of this section, D will always stand for a cyclic special form.

1) If N is a polynomial matrix such that $D^{-1}N$ is proper (that is $N \in V_D$), Proposition 1 shows that N is entirely defined by its last row, and putting $r_{p,j} = b_j$ for simplicity, we see that its jth column N_j is given by

$$N_j = \begin{bmatrix} E(a_1 b_j/a_p) \\ \vdots \\ E(a_i b_j/a_p) \\ \vdots \\ b_j \end{bmatrix},$$

and we get for the jth column of $D^{-1}N$

$$(D^{-1}N)_j = \begin{bmatrix} -PP(a_1 b_j / a_p) \\ \vdots \\ -PP(a_{p-1} b_j / a_p) \\ b_j / a_p \end{bmatrix} \quad \text{with } \deg(b_j) < \deg(a_p) = n.$$

2) By Theorem 2, the factorization $D^{-1}N$ is minimal if and only if g.c.d. $(b_1, \cdots, b_m, a_p) = 1$.

3) The minors condition reads now: "b_m, a_p are coprime." According to Proposition 3, this means the system can be controlled by its last input.

4) The assertion that O is nonempty in Theorem 3 is then a well-known fact in system theory that a cyclic system over an infinite field can be controlled by a linear combination of its inputs. This is equivalent to Lemma 1. Transposing Theorem 3, we see for instance that cyclic systems of the form $D^{-1}N$ are precisely those which are observable by their last output, i.e. such that an identically zero pth output implies a zero initial state.

Example. Consider the system S given by the state space realization

$$H = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 0 & -1 \\ 1 & 0 & -2 \\ 0 & 1 & -2 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & 1 \\ 2 & 1 \\ 1 & 0 \end{bmatrix}.$$

This triple is easily seen to be canonical, and S is obviously cyclic and observable by its last output. Its transfer function T may therefore be written $D^{-1}N$, and we have indeed:

$$T = \begin{bmatrix} 4z^2+12z+11 & 2z^2-z-3 \\ z+2 & z^2+2z \end{bmatrix} (z^3+2z^2+2z+1)^{-1}$$

$$= \begin{bmatrix} 1 & -z^2-4z-6 \\ 0 & z^3+2z^2+2z+1 \end{bmatrix}^{-1} \begin{bmatrix} -1 & -(z+3) \\ z+2 & z^2+z \end{bmatrix}.$$

As expected, we have

$$-1 = E((-z^2-4z-6)(z+2)/(z^3+2z^2+2z+1))$$

and

$$-(z+3) = E((-z^2-4z-6)(z^2+z)/(z^3+2z^2+2z+1)).$$

Note that S is not controllable by its last input, in accordance with the fact that z^2+z and $z^3+2z^2+2z+1 = (z+1)(z^2+z+1)$ are not coprime, that is the minors condition does not hold.

4. An analytic structure for $\mathbf{S}_{n,l}$. We first prove that $\mathbf{S}_{n,l}$ is a submanifold of $\mathbf{S}_n$. Let P be in Gl_m. We consider the mapping $\phi_P : \mathbf{S}_n \to \mathbf{S}_n$ such that if T is the transfer function of S, $\phi_P(S)$ has transfer function TP. It is easily seen that ϕ_P is analytic: if ν, μ are nice selection for S and $\phi_P(S)$ respectively, if (H, F, G) and (H', F', G') are realizations of these two systems and using the notation of § 1, we get

$$(H'_\mu, F'_\mu, G'_\mu) = (H_\nu, F_\nu, G_\nu P)^L, \tag{16}$$

where L is the submatrix of $\mathbf{R}_{F_\nu, G_\nu P}$ corresponding to the indices of μ. In a neighborhood of S, μ will remain a nice selection for $\phi_P(S)$, because this just means that some determinant is nonzero, and will remain locally so. Therefore (16) remains valid locally,

showing that ϕ_P is rational, hence analytic. Because ϕ_P^{-1} is just $\phi_{P^{-1}}$ it has analytic inverse. By what precedes, $(\phi_P^{-1}(\mathbf{S}_n^\nu), \psi_\nu \circ \phi_P)$ is a chart on $\mathbf{S}_n$, for every nice selection ν. We choose some cyclic structure $l=(n_1, \cdots, n_r)$, and denote by ν_0 the following nice selection $\{(i,j);\ m-r+1 \leqq j \leqq m, 0 \leqq i < n_{r-m+j}\}$, where, we recall, this means we select in the reachability matrix the columns $F^i g_j$ with (i,j) as above. If the g_i's are as in Proposition 3, ν_0 is nice for S, because the $F^i g_j$ with $(i,j) \in \nu_0$ are then independent over $\mathbf{R}$, so that Proposition 3 and Theorem 3 show that every $S \in \mathbf{S}_{n,l}$ belongs to $\phi_P^{-1}(\mathbf{S}_n^{\nu_0})$ for some appropriate P and we analyse $\mathbf{S}_{n,l}$ locally with the chart $(\phi_P^{-1}(\mathbf{S}_n^{\nu_0}), \psi_{\nu_0} \circ \phi_P)$, denoted by (V_P, θ_P). Observe that $S \in V_P$ belongs to $\mathbf{S}_{n,l}$ if and only if $\phi_P(S)$ does, for ϕ_P preserves the cyclic structure. This means that $\theta_P(\mathbf{S}_{n,l} \cap V_p) = \psi_{\nu_0}(\mathbf{S}_{n,l} \cap \mathbf{S}_n^{\nu_0})$ and to characterize this last set, we shall use the next proposition. Let us first introduce a definition: according to our previous notations (cf. § 1), F_{ν_0} is a matrix consisting in blocks $B_{i,j}$, of size $n_i \times n_j$ respectively, where $1 \leqq i \leqq r$, $1 \leqq j \leqq r$, each diagonal block $B_{i,i}$ being a companion matrix (that is a matrix with 1's under the diagonal, parameters in the last column, and 0's elsewhere), and each nondiagonal block $B_{i,j}$, $i \neq j$, being a matrix with parameters in the last column and 0's elsewhere. Writing $(a_0, \cdots, a_{n_i-1})$ for the transpose of the last column of $B_{i,j}$, we define the associate polynomial of this block to be

$$z^{n_i} - \sum_{k=0}^{n_i-1} a_k z^k \quad \text{if } i=j \quad \text{and} \quad \sum_{k=0}^{n_i-1} a_k z^k \quad \text{if } i \neq j$$

(so that the associate polynomial of a diagonal block is its companion polynomial). In the sequel, $P_{i,j}$ will always denote the associate polynomial of the block $B_{i,j}$ of the matrix F_{ν_0} under consideration.

PROPOSITION 4. *A necessary and sufficient condition for F_{ν_0} as above to have cyclic structure l is that it satisfies*

$$P_{i,j} = 0 \quad \text{if } i<j, \qquad P_{j,j} \mid P_{i,k} \quad \text{if } i \geqq j,\ k \geqq j. \tag{17}$$

The $P_{i,i}$'s are then the invariants of F_{ν_0}.

Proof. We first prove necessity. We denote by $(e_i)_{1 \leqq i \leqq n}$ the vectors of the canonical basis of $\mathbf{R}^n$, and for $1 \leqq j \leqq r$ we define $t_j = 1 + \sum_{i<j} n_i$ and $w_j = e_{t_j}$. By hypothesis, we can write $\mathbf{R}^n$, viewed as a $\mathbf{R}[z]$-module by means of the action of F_{ν_0}, as a direct sum $\sum_{i=1}^r \{v_i\}$ where v_i has period p_i, the p_i's being the invariants. Were some $P_{i,r}$ nonzero for $i<r$, the first n_r+1 powers of F_{ν_0} applied to w_r would yield independent vectors, so the period of w_r would have degree greater than n_r, which is impossible because p_r annihilates the whole module. Now since w_r has period $P_{r,r}$ of maximal degree n_r, we have $P_{r,r} = p_r$, and (17) holds for $j=r$. If $r=1$, we are done. Suppose by induction that (17) and the relation $P_{j,j} = p_j$ have been proved for every $j > k \geqq 1$, for some $k \leqq r-1$. We consider the modules $V_i = \{w_i, \cdots, w_r\}$ defined for $k+1 \leqq i \leqq r$, and the modules $V_i' = p_k V_i$. For every $i<r$ as above (if any), we have by induction hypothesis: $(p_i/p_k)(p_k w_i) = p_i w_i = \sum_{s=i+1}^r P_{s,i} w_s$, this last sum being in $p_i V_{i+1}$, hence a fortiori in V_{i+1}'. Therefore, the period q of $p_k w_i$ modulo V_{i+1}' divides p_i/p_k. But since, by induction hypothesis, w_i has period p_i modulo V_{i+1}, and qp_k is an annihilator for w_i modulo V_{i+1}', we see that p_i divides qp_k, hence $q = p_i/p_k$. Hence the dimension over $\mathbf{R}$ of the quotient module V_i'/V_{i+1}' is $n_i - n_k$. Now the dimension over $\mathbf{R}$ of V_r' is clearly $n_r - n_k$, since the period of $p_k w_r$ is p_r/p_k, and we get by addition $\dim_{\mathbf{R}} V_{k+1}' = \sum_{s=k+1}^r (n_s - n_k)$. This last number is also equal to the dimension of $p_k \mathbf{R}^n$ because this module is the direct sum of the $\{p_k v_i\}$'s (obvious). By inclusion and equality of dimension over $\mathbf{R}$,

we conclude that $V'_{k+1} = p_k \mathbf{R}^n$. Therefore we can write

$$p_k w_k = \sum_{i=k+1}^{r} Q_i p_k w_i \tag{18}$$

with $Q_i \in \mathbf{R}[z]$, and by what precedes, we may ensure upon euclidean division by p_i/p_k that $\deg(Q_i) < n_i - n_k$. Now the period $P_{k,k}$ of w_k modulo $\{w_1, \cdots, w_{k-1}, w_{k+1}, \cdots, w_r\}$ has degree n_k. Hence, by (18), $P_{k,k} = p_k$ and (17) holds obviously for $j = k$. Necessity is thus proved by induction.

We prove sufficiency: by (17) we may write $P_{i,k} = Q_{i,k} P_{k,k}$ and

$$P_{k,k} w_k = \sum_{i=k+1}^{r} Q_{i,k} P_{k,k} w_i \quad \text{for every } k,\ 1 \leqq k \leqq r,$$

the above sum being zero if $k = r$. Let $w'_k = w_k - \sum Q_{i,k} w_i$ if $k < r$, and $w'_r = w_r$. Then w'_k has period $P_{k,k}$: this is obvious if $k = r$, and otherwise comes from the fact that $P_{k,k}$ is an annihilator, while w_k has period $P_{k,k}$ modulo V_{k+1} by hypothesis. Furthermore the sum $\sum \{w'_k\}$ is direct, for if $\sum R_k w'_k = 0$ with $R_k \in \mathbf{R}[z]$, then $R_1 w'_1 \in V_2$, whence also $R_1 w_1 \in V_2$, and therefore $P_{1,1} | R_1$, and $R_1 w'_1 = 0$. By induction we get $R_k w'_k = 0$ for every k. Because $P_{i,i} | P_{j,j}$ for $i \leqq j$ by (17), this proves that the $P_{k,k}$'s are the invariants of F_{ν_0}, and since $\deg(P_{k,k}) = n_k$, F_{ν_0} has cyclic structure l. Q.E.D.

We see from Proposition 4 that for F_{ν_0} associated with $S \in \mathbf{S}_{n,l} \cap \mathbf{S}_n^{\nu_0}$, we can write

$$P_{i,j} = Q_{i,j} P_{j,j} \quad \text{for } i > j \quad \text{and} \quad P_{j,j} = Q_{j,j} P_{j-1,j-1} \quad \text{for } j > 1.$$

This allows us to define a map $g : \mathbf{S}_{n,l} \cap \mathbf{S}_n^{\nu_0} \to \mathbf{R}^N$ which to S associates the coefficients of $P_{1,1}$, of the $Q_{i,j}$'s defined above, and of the parameters of G_{ν_0} and H_{ν_0} associated with S.

We set $O_P = \mathbf{S}_{n,l} \cap V_P$, and $\chi_P = g \circ \phi_P : O_P \to \mathbf{R}^N$. $(O_P)_{P \in Gl_n}$ is an open covering of $\mathbf{S}_{n,l}$, and we can now prove the already announced result.

THEOREM 4. *$\mathbf{S}_{n,l}$ is a submanifold of $\mathbf{S}_n$, of dimension $N = n(m+p) - \sum_{i=1}^{r-1} (2i+1) n_{r-i}$ where r is as usual the number of invariants (if $r = 1$ the above sum is 0), n_i being the degree of the ith invariant. $(O_P, \chi_P)_{P \in Gl_n}$ defined above is an atlas for $\mathbf{S}_{n,l}$.*

Proof. Using the notation of § 1, we let S be in $\mathbf{S}_n^{\nu_0}$ and H, F, G be a realization of S. For simplicity, we write H_0, F_0, G_0 instead of H_{ν_0}, F_{ν_0}, G_{ν_0}. The parameters in H_0 and G_0 will be denoted by $p(H_0)$, $p(G_0)$. The parameters in F_0 are, according to our notations, the coefficients of the $P_{i,j}$'s; $\psi_{\nu_0}(S) \in V_{\nu_0}$ consists of all parameters in H_0, F_0, G_0, and will be denoted by $(p(H_0), p(G_0), P_{i,j})$. For every i, j, $i > j$, we write by euclidean division

$$P_{i,j} = Q_{i,j} P_{j,j} + R_{i,j},$$

and if $j > 1$,

$$P_{j,j} = Q_{j,j} P_{j-1,j-1} + R_{j,j}.$$

We consider the map $f : V_{\nu_0} \to \mathbf{R}^{n(m+p)}$ such that $f(p(H_0), p(G_0), P_{i,j}) = (p(H_0), p(G_0), (P_{i,j})_{i<j}, P_{1,1}, (Q_{i,j})_{i \geqq j}, (R_{i,j})_{i \geqq j})$. In other words, we choose as parameters, instead of the coefficients of the $P_{i,j}$'s with $i > j$ or $i = j > 1$, the coefficients of their quotients and rests upon division by $P_{j,j}$ if $i > j$, and by $P_{j-1,j-1}$ if $i = j$. Since $P_{j,j}$ is monic, the coefficients of the $Q_{i,j}$'s and $R_{i,j}$'s are given by universal polynomial formulae in those of the $P_{i,j}$'s. Therefore f is analytic. It is also injective, since f^{-1} is readily computed. By invariance of the domain [24], f is open, and its image W is an open set. Finally, f^{-1} is obviously analytic, and this shows that $(V_P, f \circ \theta_P)$ is a chart on $\mathbf{S}_n$ for every $P \in Gl_m$. By Proposition 4, $f \circ \theta_P(V_P \cap S_{n,l})$ consists in those elements

of W such that all $R_{i,j}$'s, and $P_{i,j}$ for $i<j$, are zero, hence our chart splits, and $\mathbf{S}_{n,l}$ is a submanifold of S_n, of dimension $n(m+p)-k$ where k is the number of parameters contained in the $R_{i,j}$'s and the $P_{i,j}$'s that are zero. For fixed j, this number is equal to $\sum_{i<j} n_i + n_{j-1} + n_j(r-j)$ where $n_0 = 0$ conventionally. Hence the coefficient of n_i in k is $2(r-i)+1$ if $i<r$ and 0 if $i=r$. This yields the desired value for N. It is moreover obvious from the proof that (O_P, χ_P) is a system of charts for $\mathbf{S}_{n,l}$. Q.E.D.

We observe that the only case when $N = n(m+p)$ is that of cyclic systems, when $r=1$. The fact that cyclic systems form a submanifold of maximal dimension implies it is an open set. However, it is straightforward that it is an open and dense subset of $\mathbf{S}_n$ from the two following facts: first $\mathbf{S}_n^{\nu_0}$, which is open by definition, consists entirely in cyclic systems in that case; and second, any system with distincts poles (eigenvalues of F) in $\mathbf{C}$ is cyclic, while it is possible to approach arbitrarily a given polynomial with a simple rooted one.

We now perform another construction to endow $\mathbf{S}_{n,l}$ with an analytic structure. Let S be in $\mathbf{S}_n^{\nu_0} \cap \mathbf{S}_{n,l}$. By Propositions 3 and 4, S satisfies the minors condition, whence, denoting by T its transfer function, we may write by Theorem 3 a t-special factorization: $T = N^*(D^*)^{-1}$. We want to take as parameters the "independent" coefficients in N^* and D^*, that is, with the notation of §§ 2, 3, the coefficients of d^*_{m-r+1}, of d^*_j/d^*_{j-1} for $j > m-r+1$, of the $q^*_{i,j}$'s, and of the $r^*_{i,j}$'s. To do this, we must prove, among other things, that N^* and D^* are unique. A direct proof of this fact can be found in [2]; however, we shall derive it here from the uniqueness of H_{ν_0}, F_{ν_0}, G_{ν_0} associated with S. This is contained in the following and final result.

THEOREM 5. *To every $S \in (\mathbf{S}_n^{\nu_0} \cap \mathbf{S}_{n,l})$, with transfer function T, is associated a unique pair N^*, D^* such that $T = N^*(D^*)^{-1}$ is a t-special factorization. If $\sigma : \mathbf{S}_n^{\nu_0} \cap \mathbf{S}_{n,l} \to \mathbf{R}^N$ is the map such that $\sigma(S)$ is the list, arranged in any conventional order, of the coefficients of d^*_{m-r+1}, of d^*_j/d^*_{j-1} for $m-r+1<j \leqq m$ (if any), of the $q^*_{i,j}$'s and the $r^*_{i,j}$'s, then $(O_P, \sigma \circ \phi_P)$, for P in Gl_m, is a system of charts for $\mathbf{S}_{n,l}$, and endows it with its analytic structure of submanifold of $\mathbf{S}_n$.*

Proof. We have already seen that a t-special factorization exists for T, say $N^*(D^*)^{-1}$. Associating with S its realization (H_0, F_0, G_0)-notation of the proof of Theorem 4—we know by realization theory and Proposition 4 that $P_{j,j} = d_{m-r+j}$ for $1 \leqq j \leqq r$, while $d_k = 1$ for $1 \leqq k \leqq m-r$ (if any). We can write by definition $D^* = U \operatorname{diag}\{d_i\}$ where U is a lower triangular matrix with 1's on its diagonal (hence unimodular), whose element (i,j) is $q^*_{i,j}$. We put $U^{-1} = (s_{i,j})$. It is still lower triangular with 1's on the diagonal, and we have

$$T = N^* M^{-1} U^{-1} \quad \text{with } M = \operatorname{diag}\{d_i\}. \tag{19}$$

From the formulae $s_{l,j} = -\sum_{l \geqq k > j} s_{l,k} q^*_{k,j}$, it is readily seen that $\deg(s_{i,j}) < \deg(d_i) - \deg(d_j)$ if $i > j$, because this is true for $q^*_{i,j}$. We denote by C_i the ith column of U^{-1}, and we associate with (19) its Fuhrmann realization (H, F, G), by choosing as a basis of the state module $\mathbf{R}^m[z]/M\mathbf{R}^m[z]$ the classes of the $z^j C_i$'s, for $m-r<i \leqq m$ and $0 \leqq j < \deg(d_i)$. Note that these constitute indeed a basis because $S \in S_n^{\nu_0}$. Furthermore, (H, F, G) and (H_0, F_0, G_0) coincide by uniqueness of the latter. With the notations of § 1 to denote equivalence classes in the quotient module of M, we see from Proposition 4 that $\underline{C}_i$ has period d_i modulo $\{\underline{C}_{i+1}, \cdots, \underline{C}_m\}$ if $m-r<i \leqq m$, and that we may write with the notations introduced in the proof of Theorem 4

$$d_i \underline{C}_i = \sum_{k=i+1}^{m} Q_{k,i} d_i \underline{C}_k.$$

Because U^{-1} is triangular with 1's on the diagonal, this yields in particular $d_i s_{i+1,i} =$

$d_iQ_{i+1,i} \bmod d_{i+1}$, whence $s_{i+1,i} = Q_{i+1,i}$ by comparison of the degrees. Suppose inductively that $s_{k,i}$ has been computed from the $Q_{u,v}$'s and d_j's, if $i < k < i+l$ and $m-r < i \leqq m$; we then have $d_i s_{i+l,i} = \sum_{k=i+1}^{i+l} Q_{k,i} d_i s_{i+l,k}$ modulo d_{i+l}, and all the $s_{i+l,k}$'s are known in the above sum. Now the degree of $s_{i+l,i}$ allows us to determine it upon euclidean division of the sum by d_{i+l} (we assume $i+l \leqq m$, but otherwise there is no need to compute $s_{i+l,i}$). This shows by induction that $s_{i,j}$ is uniquely and rationally determined from F_0, provided $m-r < j$. If $1 \leqq j \leqq m-r$ (if any), we have $\underline{C}_j = \sum_{k=m-r+1}^{m} T_k \underline{C}_k$ where the T_k's are known polynomials, whose coefficients are to be read in G_0. Therefore $s_{i,j} = \sum T_k s_{i,k}$ modulo d_i, and the $s_{i,k}$'s are known in this sum. Upon euclidean division by d_i, $s_{i,j}$ is again uniquely and rationally determined. This gives U^{-1}, hence U, hence D^*, and $N^* = TD^*$. The map σ is thus well defined, rational, and obviously injective. We now count the number of parameters in N^* and D^*. N^* provides pn coefficients for the $r_{i,j}^*$'s, while D^* provides n_r coefficients for the d_i's, $(m-r)n$ coefficients for the $q_{i,j}^*$'s with $j < m-r+1$, and $\sum_{s=1}^{r-1} \sum_{r \geqq k > s} (n_k - n_s)$ for the other $q_{i,j}^*$'s. This last sum is equal to $\sum_{i=0}^{r-1} n_{r-i}\,(r-2i-1)$. The total number of parameters is now easily seen to be N (cf. Theorem 4), showing that the image B of σ is a subset of $\mathbf{R}^N$. B is easily seen to be open (it must be so anyway by invariance of the domain), and the above considerations show that the map $\sigma \circ \phi_P \circ \chi_P^{-1} = \sigma \circ g^{-1} : \chi_P(O_P) \to B$, is rational, hence analytic. It remains to prove that its inverse is also analytic, but this is easily seen, because for $S \in \mathbf{S}_n^{\nu_0} \cap \mathbf{S}_{n,l}$, as we have seen above, H_0, F_0, G_0 are precisely the matrices of the Fuhrmann maps associated with our t-special factorization $N^*(D^*)^{-1}$, with basis $z^j \underline{C}_i$ in the state space, showing that $\psi_{\nu_0} \circ \sigma^{-1}$ is rational, hence also $g \circ \sigma^{-1}$. Q.E.D.

As mentioned in § 3, we might choose our parameters in special factorizations $D^{-1}N$, and even impose that $D^{-1}N = N^*(D^*)^{-1}$ is solvable, i.e. both minors condition and its dual are satisfied. The analytic structure thus obtained is similar to the present one [2].

5. Relation to other works.

A. About special factorizations. The existence of the factorizations introduced in this paper depends on Theorem 1, which is used in § 3 to analyse the external representation of a given system. They are related later only with state-space description, through Propositions 3 and 4. An alternative proof of this existence, proceeding in some sense in reverse order, may be sketched as follows. We start with an algebraic identity given in [1, (5.5a)]

$$(zI - F_\nu)^{-1} G_\nu = W_\nu E_\nu^{-1},$$

where ν is a nice selection, W_ν, E_ν polynomial matrices, E_ν being built with the associate polynomials $P_{i,j}$ of F_ν. If l is some cyclic structure and ν the associated nice selection (denoted by ν_0 in § 4), Proposition 4 shows at once that when F_ν has cyclic structure l, E_ν is under t-special form. Applying Theorem 3 and Proposition 3 then yields the existence of t-special factorizations.

B. About the submanifolds $\mathbf{S}_{n,l}$. The set $\mathbf{W}_{n,l}$ consisting in those elements of $\mathbf{S}_{n,l}$ whose invariant polynomials have simple roots is clearly an open (dense) subset of $\mathbf{S}_{n,l}$, hence a submanifold of $\mathbf{S}_n$, of dimension N. This fact can also be deduced from a result of Kanewsky [22, (1.5), (1.6)], asserting that systems (over $\mathbf{C}$) with given Jordan structure in the state space form a smooth subvariety (hence a submanifold) of $\mathbf{S}_n$, of dimension $N - n_r$. One might for instance choose the roots of the rth invariant as n_r supplementary parameters, and the product structure thus obtained yields the desired result for $\mathbf{W}_{n,l}$ (over $\mathbf{C}$). However, the analysis near a multiple root is more difficult,

since the Jordan structure is locally subject to change, and the parametrization of the various branches is not helpful if we do not know on which one we are.

C. Application to L^2 approximation. The results obtained here in the cyclic case have been used in [2] to approach the problem of approximating a (possibly non-rational) L^2 sequence of matrices, F, by a stable rational cyclic one of given state space dimension, T, the error in the L^2 norm being minimal.

The technique consists merely in differentiating $\|F-T\|^2$ with respect to the parameters introduced in this paper. Identifying the vector space spanned by the derivatives with some space related to V_D and V_D^* (cf. Proposition 1), one obtains a necessary condition on T in terms of orthogonality between certain spaces, which generalizes the equation obtained in [23] to treat the scalar case.

This, together with a heuristic algorithm, has been specialized in [3] to the case when F is almost rational, for identification purposes.

6. Conclusion. In this paper, we present an attempt to parametrize systems by means of special factorizations $RD^{-1}N$ of their transfer functions. Roughly speaking, we want D to be triangular, so that we can control deg (det D), that is the dimension of the state space. We then need another matrix R to get divisibility conditions in D, that allow us in turn to find free parameters in N. But we restrict R to be scalar, because we also want to take causality into account, which is easier to analyse when we have only two factors in the product. It turns out that once R is chosen, D and N are uniquely determined, just as the choice of some nice selection ν makes (H_ν, F_ν, G_ν) unique. However, imposing the divisibility conditions prevents us from parametrizing topologically the whole manifold. Instead, we have to limit ourselves to certain submanifolds, namely the $\mathbf{S}_{n,l}$'s. The parametrization in the cyclic case is by far the simplest one, in spite of its maximum number of parameters.

REFERENCES

[1] A. C. ANTOULAS, *On canonical forms for linear constant systems*, Internat. J. Control, 33 (1981), pp. 95-122.

[2] L. BARATCHART, *Une structure différentielle pour certaines classes de systèmes, application a l'approximation L^2*, Thèse de docteur ingénieur, Ecole Nat. Sup. des Mines de Paris, December 1982.

[3] L. BARATCHART AND S. STEER, *Sur l'identification des systèmes cycliques*, Proc. 6th International Conference on Analysis & Optimization of Systems, Nice, Lecture Notes in Control and Information Sciences 63, A. Bensoussan and J. L. Lions, eds, Springer-Verlag, Berlin, 1984.

[4] S. BEGHELLI AND R. GUIDORZI, *A new input-output canonical form for multivariable systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 692-696.

[5] ———, *Transformations between input-output multistructural models: properties and applications*, Internat. J. Control, 37 (1983), pp. 1385-1400.

[6] R. W. BROCKETT, *Some geometric questions in the theory of linear systems*, IEEE Trans. Automat. Control, AC-21 (1976), pp. 449-455.

[7] C. I. BYRNES AND T. E. DUNCAN, *On certain topological invariants arising in system theory*, in New Directions in Applied Mathematics, P. Hilton and G. Young, eds, Springer, New York, 1981, pp. 29-71.

[8] C. I. BYRNES AND N. HURT, *On the moduli of linear dynamical systems*, Adv. in Math. Studies in Analysis, 4 (1979), pp. 83-122.

[9] D. F. DELCHAMPS, *The geometry of spaces of linear systems with an application to the identification problem*, Ph.D. thesis, Harvard Univ., Cambridge, Ma., 1982.

[10] P. A. FUHRMANN, *Algebraic system theory: an analyst's point of view*, J. Franklin Institute, 301 (1976), pp. 521-540.

[11] R. GUIDORZI, *Invariants and canonical forms for systems: structural and parametric identification*, Automatica, 17 (1981), pp. 117-133.

[12] M. HAZEWINKEL, *Moduli and canonical forms for linear dynamical systems* III: *the algebraic-geometric case*, Proc. 1976 Ames Research Center Conference on Geometric Control Theory, C. Martin and R. Hermann, eds., Math Sciences Press, 1977, pp. 291-336.

[13] ———, *Moduli and canonical forms for linear dynamical systems* II: *the topological case*, Math. Syst. Theory, 10 (1977), pp. 363-385.

[14] M. HAZEWINKEL AND R. E. KALMAN, *On invariants, canonical forms and moduli for linear, constant, finite dimensional, dynamical systems*, in Proc. CNR-CISM Symposium on Algebraic System Theory, Udine 1975, Lecture Notes in Economics and Mathematical Theory 131, Springer, New York, 1976, pp. 48-60.

[15] U. HELMKE, *Zur Topologie des Raumes linearer Kontrollsysteme*, Ph.D. thesis, Univ. Bremen, 1982.

[16] ———, *The topology of the space of linear systems*, Proc. 21st IEEE Conference on Decision and Control, 1982, pp. 948-949.

[17] M. HEYMANN, *Structure and Realization Problems in the Theory of Dynamical Systems*, C.I.S.M. Courses and Lectures 204, Springer, New York, 1975.

[18] N. JACOBSON, *Lectures in Abstract Algebra Vol.* II: *Linear Algebra*, Springer, New York, 1953.

[19] R. E. KALMAN, *Global structure of classes of linear dynamical systems*, Lectures of the NATO Advanced Study Institute on Geometric and Algebraic Methods for Nonlinear Systems, London, 1971.

[20] ———, *Algebraic geometric description of the class of linear systems of constant dimension*, 8th Annual Princeton Conference on Information Sciences and Systems, 1974.

[21] ———, *Kronecker invariants and feedback*, in Ordinary Differential Equations, L. Weiss ed., Academic Press, New York, 1972.

[22] D. KANEWSKY, *On the algebro-geometric structure of the moduli space of completely reachable systems*, J. Pure Appl. Algebra, 29 (1983), pp. 143-150.

[23] E. ROSENCHER, *Approximation rationnelle des filtres à un ou deux indices: une approche hilbertienne*, thèse de docteur ingénieur, Univ. Paris IX, 1978.

[24] M. SPIVAK, *A Comprehensive Introduction to Differential Geometry Vol.* I, Publish or Perish Press, Berkeley, CA, 1970.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

007

STOCHASTIC APPROXIMATION IN HILBERT SPACE: IDENTIFICATION AND OPTIMIZATION OF LINEAR CONTINUOUS PARAMETER SYSTEMS*

H. J. KUSHNER† AND A. SHWARTZ†

Abstract. We treat a class of stochastic approximations (with small constant gain ε), with values in a Hilbert space. The problem and algorithm arise, e.g., when one seeks to iteratively identify the transfer function of a linear system (continuous parameter) or to adaptively optimize the transfer function of a stochastic linear system. Weak convergence methods are used to prove convergence of the interpolated sequence as $\varepsilon \to 0$, and to characterize the equations satisfied by the limit.

Projected and unprojected cases are dealt with. In one important case, convergence to a constrained optimum is proved when $\varepsilon n \to \infty$ as $\varepsilon \to 0$. The normalized error sequence is analyzed. It is shown that the limit (as $\varepsilon \to 0$) of the interpolated normalized error sequence satisfies a linear integral equation driven by a Hilbert space Wiener process. Many of the calculations and results are useful for approximation problems for distributed systems with nonwhite noise inputs.

Key words. stochastic approximation, optimization in Hilbert space, recursive estimation, system identification, adaptive stochastic centrol

1. Introduction. consider a system with transfer function $K(\cdot)$, input $z(\cdot)$, sampling interval Δ and output (at sampling time $n\Delta$)

$$y_n = \int_0^T K(\tau) z(n\Delta - \tau)\, d\tau + \psi_n,$$

where $\{\psi_n\}$ is a stationary (observation noise) sequence with mean zero, and independent of $z(\cdot)$. For notational convenience (w.l.o.g.), we set $T = 1$ henceforth. We prove a number of results concerning the recursive algorithm (1.1) and its "projected" form (1.3) for estimation of the transfer function $K(\cdot)$: convergence, rate of convergence and behavior of the estimates for large times. As will be pointed out below, the techniques and ideas are useful in many other cases.

The basic algorithm (1.1) is an obvious distributed parameter analogue of an algorithmic form which is commonly used in the finite dimensional case. $K_n^\varepsilon(\cdot)$ is the nth recursive estimate of $K(\cdot)$.

$$(1.1)\quad K_{n+1}^\varepsilon(u) = K_n^\varepsilon(u) - \varepsilon z(n\Delta - u)\left[\int_0^1 K_n^\varepsilon(s) z(n\Delta - s)\, ds - y_n\right], \quad \varepsilon > 0, \quad K_0^\varepsilon = K_0.$$

As in the finite-dimensional case, (1.1) can be viewed either as an algorithm for identifying K or for optimizing the transfer function of a linear system. Define $H = L_2[0, 1]$, with inner product $\langle x, y\rangle$ and norm $|x|$.

Throughout we assume the following (although they are not always needed). The process $z(\cdot)$ is stationary, mean zero, $E|z(t)|^{4+\alpha} < \infty$ for some $\alpha > 0$ and $E|\psi_n|^{2+\alpha} < \infty$ for some $\alpha > 0$. Let $R(u) = Ez(u+s)z(s)$ be continuous and define the operator R from H to H by $Rf = g$, where $g(u) = \int_0^1 R(u-s) f(s)\, ds$, $u \in [0, 1]$. Assume that R has a complete orthonormal set $\{e_i\}$ of eigenfunctions with eigenvalues $\{\lambda_i\}$. If not, then

* Received by the editors August 18, 1983, and in revised form June 15, 1984. This research was supported in part by the Air Force Office of Scientific Research under contract AF-AFOSR 81-0116, by the National Science Foundation under contract ECS82-11476 and by the Office of Naval Research under contract N00014-76-C-0279-P6.

† Lefschetz Center for Dynamical Systems, Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912.

we augment the set of eigenfunctions in order to complete them, and all the results, except possibly those in § 4 continue to hold. Note $R(0)=\sum_i \lambda_i<\infty$. Define $\delta K_n^\varepsilon(u)=K_n^\varepsilon(u)-K(u)$, and let K_n^ε, δK_n^ε, K and Z_n denote the H-valued random variables whose point values are $K_n^\varepsilon(u)$, $\delta K_n^\varepsilon(u)$, $K(u)$ and $z(n\Delta-u)$, resp.

Rewrite (1.1) in the Hilbert space form

$$K_{n+1}^\varepsilon = K_n^\varepsilon - \varepsilon Z_n(\langle \delta K_n^\varepsilon, Z_n\rangle - \psi_n), \qquad K_0^\varepsilon = K_0. \tag{1.2}$$

Define the *piecewise constant interpolations* $K^\varepsilon(\cdot)$ and $\delta K^\varepsilon(\cdot)$ by $K^\varepsilon(t)=K_n^\varepsilon$ and $\delta K^\varepsilon(t)=\delta K_n^\varepsilon$, resp., on the time interval $[n\varepsilon, n\varepsilon+\varepsilon)$. We also discuss the projected algorithm (1.3), where $\pi_1=$ projection onto the unit ball in H (any size ball will do just as well).

$$K_{n+1}^\varepsilon = \pi_1[K_n^\varepsilon - \varepsilon Z_n(\langle \delta K_n^\varepsilon, Z_n\rangle - \psi_n)], \qquad K_0^\varepsilon = K_0. \tag{1.3}$$

The purpose of (1.3) is to obtain the estimate $\hat{K}$ which minimizes in

$$\begin{aligned}\min_{|\hat{K}|\leq 1} E\left|y_n - \int_0^1 \hat{K}(s)z(n\Delta-s)\,ds\right|^2 &= \min_{|\hat{K}|\leq 1} E\left|\int_0^1 [\hat{K}(s)-K(s)]z(-s)\,ds\right|^2 + \operatorname{var}\psi_n \\ &\equiv 2\min_{|\hat{K}|\leq 1} J(\hat{K}) + \operatorname{var}\psi_n.\end{aligned} \tag{1.4}$$

The asymptotic analysis is based on "weak convergence" techniques, and the parts of this theory which we use are stated in § 2. The limits of $\{K^\varepsilon(\cdot)\}$ are discussed in § 3, together with those for (1.3). It will be shown (under some additional conditions) that for (1.2), $K^\varepsilon(\cdot)\Rightarrow \bar{K}(\cdot)$ (an H-valued function on $[0,\infty)$ with value $\bar{K}(t)$ at time t, and point values $\bar{K}(t,u)$, $0\leq u\leq 1$, at time t), where

$$\frac{d\bar{K}(t,u)}{dt} = -\int_0^1 R(u-s)[\bar{K}(t,s)-K(s)]\,ds, \qquad \bar{K}(0,u)=K_0(u). \tag{1.5}$$

Equivalently,

$$\frac{d\bar{K}(t)}{dt} = -R(\bar{K}(t)-K).$$

In § 4, we study the asymptotic behavior of $K^\varepsilon(t_\varepsilon+\cdot)$ where $t_\varepsilon\to\infty$ as $\varepsilon\to 0$, in order to get a better understanding of the large time behavior.

The rate of convergence is discussed in §§ 5 and 6. Section 5 treats some preliminary problems concerning weak convergence of a sequence of processes to an H-valued Wiener process, and the full rate of convergence problem is treated in § 6. The rate of convergence, as dealt with here, concerns the behavior of the normalized sequence $\{(K_n^\varepsilon-\bar{K}(\varepsilon n))/\sqrt{\varepsilon}\}$, analogous to a usual procedure in the finite-dimensional case. We also comment on the relationship between (1.1) and its finite-dimensional "time-discretized" version.

Remarks and applications. There are few works on stochastic approximation in Hilbert (or in any abstract) space, and these seem to be entirely devoted to the case where ε is replaced by $\varepsilon_n>0$ and $\sum_n \varepsilon_n=\infty$, $\sum_n \varepsilon_n^2<\infty$ ([1]-[4]). Even with this change, these works do not allow us to treat (1.1) or (1.3), owing to the correlation among the $\{z_n, \psi_n\}$. Also, there are no results (known to the authors) on rates of convergence, constrained algorithms or on the properties as $\varepsilon n\to\infty$ as $\varepsilon\to 0$. The general ideas and techniques which are developed for our special case serve as a guide to the treatment of the more general nonlinear cases.

The applications of (1.1) or (1.3) in adaptive signal detection or identification are the same as those for the finite-dimensional case [5], [6]. We use the language of

"identification" here, but in the adaptive detection problem, one wants to iteratively adjust $K(\cdot)$ so as to have $\int_0^1 K(1-s)z(n\Delta-s)\,ds$ "best" fit a given sequence $\{y_n\}$, but the mathematical development is the same as for (1.1). There are many other applications which involve natural extensions of the ideas here. Consider the problem of an adaptive matched filter, where we wish to recursively adjust $K(\cdot)$ so as to maximize the signal to noise ratio in the scalar system output $\int_0^1 K(1-u)[s(u)+\xi(u)]\,du$, where $s(\cdot)=$ signal, $\xi(\cdot)=$ noise. Two formulations are: (1) Constrain $K(\cdot)$ such that $\int_0^1 K(1-u)s(u)\,du=1$, $|K|\leqq$ constant and min $E[\int_0^1 K(1-u)\xi(u)\,du]^2$; (2) Hold $|K|=1$ and minimize $E[\int_0^1 K(1-u)(s(u)+\xi(u))\,du]^2$. For either form there is an appropriate form of the "projected" algorithm (1.3) which can be used.

The time parameter u in (1.1) or (1.3) can be vector-valued, so that one can use an extension or (1.1) of (1.3) to, e.g., iteratively find $K_1(\cdot)$, $K_2(\cdot)$ which minimize in the Volterra form

$$E\left[y_n-\int_0^1 K_1(u)z(u)\,du-\int_0^1\int_0^1 K_2(u,v)z(u)-z(v)\,du\,dv\right]^2.$$

For this example, the covariance operator R is replaced by a two parameter covariance operator, but the general ideas are the same.

An additional motivation for the present work is that it allows treatment (in a relatively simple context) of problems arising in the modelling and approximation of distributed parameter systems. For example, the convergence and rate of convergence results seem to be useful for the problem of obtaining the correct stochastic PDE's which model distributed parameter systems where there are "nonwhite" noise processes (in analogy to the methods used in [7], [8], [9] to obtain the proper Itô equation approximation to systems with wide bandwidth inputs). Currently available work (e.g., [10], [11]) on distributed system approximation uses a completely Markovian structure, where the state of interest is Markovian. This will be pursued in a future work.

2. Weak convergence—preliminaries. All the material in the first part of this section is taken from Kurtz [12], although with slightly altered terminology. Let S denote a metric space with metric d. Define the space $D_S[0,\infty)$ of functions from $[0,\infty)$ to S where $\lim_{s\downarrow t} f(s)=f(t)$ and $\lim_{s\uparrow t} f(s)$ exists. The natural and commonly used topology (which we use) on $D_S[0,\infty)$ is the Skorokhod topology (see [12] or [13]). If (S,d) is complete and separable, the the Skorokhod topology can be metrized so that it is complete and separable also. If $f_n(\cdot)\to f(\cdot)$ in $D_S[0,\infty)$ under the Skorokhod topology, where $f(\cdot)$ is continuous, then

$$\sup_{t\leqq T} d(f_n(t),f(t))\to 0 \quad \text{for each } T<\infty. \tag{2.1}$$

Define $D_S(-\infty,\infty)$ in the same way as $D_S[0,\infty)$ was defined, except that $t\in(-\infty,\infty)$. The discussion below is restricted to $D_S[0,\infty)$, but exactly the same facts apply to $D_S(-\infty,\infty)$.

Let $\{P_n\}$ and P denote measures on $D_S[0,\infty)$—corresponding to the set $\{Y_n(\cdot)\}$ and $Y(\cdot)$ of S-valued random processes with paths in $D_S[0,\infty)$. We say that $P_n\Rightarrow P$ or, equivalently, that $Y_n(\cdot)\Rightarrow Y(\cdot)$ (converges weakly) if for each bounded and continuous real-valued function $F(\cdot)$ on $D_S[0,\infty)$,

$$\int F(y)\,dP_n(y)\to\int F(y)\,dp(y).$$

Equivalently, $EF(Y_n(\cdot))\to EF(Y(\cdot))$. If $Y_n(\cdot)\Rightarrow Y(\cdot)$ and $Y(\cdot)$ is a (nonrandom) continuous function, then $\sup_{t\leqq T} d(Y_n(t),Y(t))\to 0$ in probability for each $T>0$.

Tightness. The sequences $\{P_n\}$ or (equivalently) $\{Y_n(\cdot)\}$ above are said to be tight if for each $\delta > 0$, there is a compact $\Gamma_\delta \in D_S[0, \infty)$ such that[1]

$$\inf_n P\{Y_n(\cdot) \in \Gamma_\delta\} \geqq 1 - \delta.$$

If S is complete and separable, then Prokhorov's theorem states that $\{Y_n(\cdot)\}$ is tight if and only if each subsequence of $\{Y_n(\cdot)\}$ contains a further subsequence which converges weakly.

Criteria for tightness. Define $d_1(x, y) = \min[1, d(x, y)]$.

THEOREM 2.1 [12, Thm. 2.7]. *Let $\{Y_n(\cdot)\}$ have paths in $D_S[0, \infty)$, where (S, d) is complete and separable. Suppose that for each $t > 0$ and $\delta > 0$ there is a compact $\Gamma_{t,\delta}$ such that $\inf_n P\{Y_n(t) \in \Gamma_{t,\delta}\} \geqq 1 - \delta$. Let $\mathcal{F}_t^n = \sigma(Y_n(s), s \leqq t)$. Then if (2.2) holds for each $T > 0$, we have tightness of $\{Y_n(\cdot)\}$ in $D_S[0, \infty)$. The $\sup_{\tau \leqq T}$ is over all $\{\mathcal{F}_t^n, t \leqq T\}$ stopping times*

$$\lim_{\delta \downarrow 0} \overline{\lim_n} \sup_{\tau \leqq T} E d_1(Y_n(\tau + \delta), Y_n(\tau)) = 0. \tag{2.2}$$

The symbol τ will always denote such a stopping time. The criteria of Theorem 2.1 will be readily verified in our case.

Skorokhod representation. Let $Y_n(\cdot) \Rightarrow Y(\cdot)$ and let (S, d) be complete and separable. Since weak rather than w.p.1. convergence is of interest, the underlying probability space is unimportant, and we can choose it in any convenient manner as long as the distributions of each $Y_n(\cdot)$ and of $Y(\cdot)$ remains the same. By Skorokhod [14, Thm. 3.1.1], there is a probability space $(\tilde{\Omega}, \tilde{B}, \tilde{P})$ with processes $(\tilde{Y}_n(\cdot), \tilde{Y}(\cdot))$ defined on it such that $\tilde{Y}_n(\cdot) \to \tilde{Y}(\cdot)$ w.p.1. in the topology of $D_S[0, \infty)$ and for each Borel set $\Gamma \in D_s[0, \infty)$,

$$\tilde{P}\{\tilde{Y}_n(\cdot) \in \Gamma\} = P\{Y_n(\cdot) \in \Gamma\}, \qquad \tilde{P}\{\tilde{Y}(\cdot) \in \Gamma\} = P\{Y(\cdot) \in \Gamma\}.$$

This extremely useful technique of turning weak convergence into w.p.1. convergence via a particular choice of the probability space is often called *Skorokhod representation*, but we prefer the term *Skorokhod imbedding* for this. We will use it often and *without* the tilde affix for simplicity. The $\{\tilde{Y}_n(\cdot)\}$ will be referred to as the *imbedded* sequence.

Weak convergence in the weak L_2-topology. Let H_w denote the *unit ball* in $L_2[0, 1]$ with the weak topology. Let $\{l_1, \cdots\}$ denote a countable set whose finite linear combinations are dense in H, with $|l_i| = 1$. Then H_w can be metrized with the invariant metric

$$d(f) = d(f, 0) = \sum_1^\infty 2^{-n} |\langle f, l_n \rangle| / [1 + |\langle f, l_n \rangle|].$$

Tightness in $D_{H_w}[0, \infty)$. The closure of a set Γ in H_w is compact if $\sup_{x \in \Gamma} |x| < \infty$. Thus to use Theorem 2.1 on $D_{H_w}[0, \infty)$ it is enough that (recall that τ is a stopping time)

$$\sup_n E|Y_n(t)| < \infty \quad \text{for each } t, \tag{2.3}$$

$$\lim_{\delta \downarrow 0} \overline{\lim_n} \sup_{\tau \leqq T} E d_1(Y_n(\tau + \delta) - Y_n(\tau)) = 0 \quad \text{for each } T > 0. \tag{2.4}$$

Criterion (2.4) is equivalent to (for each $T > 0$ and $l \in H$)

$$\lim_{\delta \downarrow 0} \overline{\lim_n} \sup_{\tau \leqq T} E \min[1, |\langle l, Y_n(\tau + \delta) - Y_n(\tau) \rangle|] = 0. \tag{2.5}$$

The same criterion is used for tightness in $D_{H_w}(-\infty, \infty)$, except that $|\tau| \leqq T$ replaces $\tau \leqq T$.

[1] For notational convenience, we sometimes write $P\{Y_n(\cdot) \in \Gamma\}$ for $P_n\{\Gamma\}$.

If $Y_n(\cdot) \Rightarrow Y(\cdot)$ in $D_{H_w}(-\infty, \infty)$, then for the associated Skorokhod imbedded sequences $\{\tilde{Y}_n(\cdot), \tilde{Y}(\cdot)\}$, $\tilde{Y}_n(\cdot) \to \tilde{Y}(\cdot)$ w.p.1. in the Skorokhod topology. We will always drop the tilde affix. Doing this and letting $Y(\cdot)$ be continuous, there is a sequence $T_n \to \infty$ such that for each $l \in H$ (and under weak convergence and using the Skorokhod imbedding technique),

$$\sup_{|t| \leqq T_n} |\langle l, Y_n(t) - Y(t)\rangle| \to 0 \quad \text{w.p.1.} \tag{2.6}$$

as $n \to \infty$. Similarly for weak convergence in $D_S[0, \infty)$.

Tightness in $D_H[0, \infty)$. Let $\{f_i\}$ denote any orthonormal basis in H. A bounded set Γ in H is compact if for each $\eta > 0$ there is an $N_\eta < \infty$ such that

$$\sup_{x \in \Gamma} \sum_{N_\eta}^{\infty} |\langle x, f_i\rangle|^2 \leqq \eta. \tag{2.7}$$

Thus to use Theorem 2.1 in $D_H[0, \infty)$, it is enough if for each $t \geqq 0$ and $\eta > 0$, $\{|Y_n(t)|\}$ is tight and

$$\lim_N \sup_n P\left\{\sum_N^\infty |\langle Y_n(t), f_i\rangle|^2 \geqq \eta\right\} = 0, \tag{2.8}$$

$$\lim_\delta \overline{\lim_n} \sup_{\tau \leqq T} E \min[1, |Y_n(\tau + \delta) - Y_n(\tau)|] = 0 \quad \text{for each } T > 0. \tag{2.9}$$

The same criterion holds for $D_H(-\infty, \infty)$, except that $|\tau| \leqq T$ and $|t| < \infty$ are used.

A more convenient form of (2.8) is the following: Fix $t > 0$. For each $\eta > 0$ and $\rho > 0$ there is an $N_{\rho\eta} < \infty$ such that

$$P\left\{\sum_{N_{\rho\eta}}^{\infty} \langle Y_n(t), f_i\rangle^2 \geqq \eta\right\} \leqq \rho, \qquad n \geqq N_{\rho\eta}. \tag{2.10}$$

To see that (2.10) implies (2.8), assume (2.10) and fix $\eta > 0$, $\rho > 0$. Note that for each m,

$$\lim_N P\left\{\sum_N^\infty \langle Y_m(t), f_i\rangle^2 \geqq \eta\right\} = 0.$$

By this and (2.10), there are $N'_{\rho\eta} < \infty$ such that

$$P\left\{\sum_{N'_{\rho\eta}}^{\infty} \langle Y_n(t), f_i\rangle^2 \geqq 2\eta\right\} \leqq 2\rho \quad \text{for all } n.$$

This and the arbitrariness of ρ and η imply (2.8).

Finally, if $Y_n(\cdot) \Rightarrow Y(\cdot)$ in $D_H[0, \infty)$, where $Y(\cdot)$ is continuous, then (use the Skorokhod imbedding technique as done in connection with (2.6)) there are $T_n \to \infty$ such that

$$\sup_{t \leqq T_n} |Y_n(t) - Y(t)| \to 0 \quad \text{w.p. 1.} \tag{2.11}$$

Let X_n be a random variable taking values in H. We say that $\{X_n\}$ is tight in H if for each $\eta > 0$, there is a compact (strongly) set A_η: $P\{X_n \in A_\eta\} \geqq 1 - \eta$ for all n. Similarly for tightness in H_w, but then we use weakly compact sets A_η. Also, $X_n \Rightarrow X$ in H (resp., in H_w) if $EF(X_n) \to EF(X)$ for all strongly (resp., weakly) continuous bounded $F(\cdot)$.

3. Weak convergence of $\{K^\varepsilon(\cdot)\}$. First, in Theorem 3.1, we deal with algorithm (1.1); then we discuss the few changes which are required to deal with the projected

case (1.3). The proof of Theorem 3.1 divides into three parts. The first part simply defines a useful truncation device, which allows us to treat the $\{K_n^\varepsilon\}$ as though they were bounded. Then the tightness (in $D_H[0,\infty)$) of the interpolated truncated sequence is proved. Finally, we show that the limit of the truncated $K^\varepsilon(\cdot)$ satisfies a "truncated" form of (1.5). Owing to the nature of the truncation, this convergence is enough to give the desired result.

Notation. Let E_j denote conditioning on $\{Z_k, \psi_k, k<j, K_0\}$ or on $\{\psi_k, z(s), k<j, s\leqq k\Delta-\Delta, K_0\}$. Let E_t^ε denote conditioning on $\{Z_k, \psi_k, k<t/\varepsilon, K_0\}$ or on $\{\psi_k, k<t/\varepsilon, z(s), s\leqq t/\varepsilon-\Delta, K_0\}$, where for t/ε we always take the integer part.

The following assumptions will be used.

Assumption 3.1. $E_j\psi_k\to 0$ in the mean, as $k-j\to\infty$; $E[z(t+s)|z(u), u\leqq t]\to 0$ in the mean uniformly in t, as $s\to\infty$.

Assumption 3.2. $E[z(t+s)z(t+s-v)-R(v)|z(\sigma), \sigma\leqq t]\to 0$ in the mean, uniformly in t and in $v\in[0,1]$ as $s\to\infty$.

THEOREM 3.1. *Under the assumptions of* § 1 *and Assumptions* 3.1 *and* 3.2, $K^\varepsilon(\cdot)\Rightarrow \bar K(\cdot)$ *in* $D_H[0,\infty)$, *where* $\bar K(\cdot)$ *satisfies* (1.5).

Remark. When $\{e_n\}$ is complete, then K is the only stationary solution of (1.5) and $\bar K(t)\to K$ (strongly) as $t\to\infty$.

Proof. Part 1. Since the proof is simpler if the $\{K_n^\varepsilon\}$ are bounded, we use a truncation device, and first prove weak convergence for a "truncated" sequence. For each N, let $q_N(\cdot)$ be a Lipschitz continuous function (called the trunction function) from H to $[0,1]$ where $q_N(x)=1$ for $|x|\leqq N$ and $q_N(x)=0$ for $|x|\geqq N+1$. For each N, define the sequence $\{K_n^{\varepsilon,N}\}$ by

$$
\begin{aligned}
K_{n+1}^{\varepsilon,N} &= K_n^{\varepsilon,N}-(\varepsilon Z_n\langle \delta K_n^{\varepsilon,N}, Z_n\rangle-\varepsilon Z_n\psi_n)q_N(K_n^{\varepsilon,N}) \\
&= K_0-\varepsilon\sum_0^n Z_j\langle \delta K_j^{\varepsilon,N}, Z_j\rangle q_N(K_j^{\varepsilon,N})+\varepsilon\sum_0^n Z_j\psi_j q_N(K_j^{\varepsilon,N})
\end{aligned}
\tag{3.1}
$$

where $\delta K_n^{\varepsilon,N}=K_n^{\varepsilon,N}-K$. Clearly, $K_n^{\varepsilon,N}=K_n^\varepsilon$ until first escape from the N-sphere and $K_n^{\varepsilon,N}$ remains constant after first exiting the $N+1$-sphere. Define the piecewise constant interpolation $K^{\varepsilon,N}(\cdot)$ (the paths are in $D_H[0,\infty)$) as $K^\varepsilon(\cdot)$ was defined, but using $\{K_n^{\varepsilon,N}\}$. Define $q_{Nn}^\varepsilon=q_N(K_n^{\varepsilon,N})$.

To slightly simplify the problem, let K_0 be bounded uniformly in ω. In Part 2 of the proof, it is proved that $\{K^{\varepsilon,N}(\cdot), \varepsilon>0\}$ is tight in $D_H[0,\infty)$, and in Part 3 that the limits of this sequence satisfy (3.2). Assume these facts now. Let ε index a weakly convergent subsequence with limit $\bar K^N(\cdot)$. We now show that those facts imply the theorem.

We will show that $K^{\varepsilon,N}(\cdot)\Rightarrow\bar K^N(\cdot)$ in $D_H[0,\infty)$, where $\bar K^N(\cdot)$ satisfies

$$
\frac{d\bar K^N(t,u)}{dt}=-\int_0^1 R(u-s)[\bar K^N(t,s)-K(s)]q_N(\bar K^N(s))\,ds, \qquad \bar K^N(0)=K_0
\tag{3.2}
$$

for large N.

There is an N_1 such that $|\bar K(t)|\leqq N_1$ and $|\bar K^N(t)|\leqq N_1$ for all N, t. Thus for large N, $\bar K^N(\cdot)=\bar K(\cdot)$, and also $K^{\varepsilon,N}(\cdot)\Rightarrow\bar K(\cdot)$ in $D_H[0,\infty)$. Consequently, via the weak convergence and (2.11), for each $T>0$ and large N

$$
\sup_{t\leqq T}|K^{\varepsilon,N}(t)-\bar K(t)|\xrightarrow{\varepsilon}0 \quad \text{in distribution}
\tag{3.3}
$$

(use Skorokhod imbedding if $\bar K_0$ is random; otherwise it is not needed). But this

implies that for each $T>0$ and $\eta>0$

$$\lim_N \overline{\lim_\varepsilon} P\{\sup_{t\leqq T} |K^{\varepsilon,N}(t)-K^\varepsilon(t)|\geqq \eta\}=0,$$

which, together with (3.3), is equivalent to the theorem assertion. Thus we need only show that $K^{\varepsilon,N}(\cdot)\Rightarrow \bar{K}^N(\cdot)$ in $D_H[0,\infty)$ and that $\{K^{\varepsilon,N}(\cdot), \varepsilon>0\}$ is tight for each N. This is done in the following two parts.

Part 2. *Tightness of* $\{K^{\varepsilon,N}(\cdot)\}$ *in* $D_H[0,\infty)$. We apply Theorem 2.1. By the remarks at the end of § 2, in order to satisfy the first condition of Theorem 2.1, it is sufficient that for each $T>0$ and $\eta>0$,

$$\lim_M \sup_{\substack{\varepsilon,n\\ \varepsilon n\leqq T}} P\left\{\sum_{i=M}^{\infty} \langle K_n^{\varepsilon,N}, e_i\rangle^2 \geqq \eta\right\}=0. \tag{3.4}$$

Henceforth T is fixed and C_N and C denote constants (not depending on ε or n), whose values might change from usage to usage.

Define $I_{jL}=I_{\{|Z_j|\leqq L\}}$ and $I'_{jL}=1-I_{jL}$.

In the argument below, it is implicitly assumed that all $|K_n^{\varepsilon,N}|$ are bounded for each N. This is not necessarily true, since the last jump before leaving the $N+1$-ball (if this occurs) might not be bounded. But this causes no problems since for each $\rho>0$ and $N<\infty$, our assumptions on $\{\psi_n\}$ and $z(\cdot)$ imply that

$$P\{\sup_{n\leqq T/\varepsilon} |K_{n+1}^{\varepsilon,N}-K_n^{\varepsilon,N}|\geqq\rho\}\to 0$$

as $n\varepsilon\to 0$. So, for convenience, we assume boundedness of the $|K_n^{\varepsilon,N}|$.

First, we treat the first *sum* in (3.1), which we split as follows:

$$H_n^{\varepsilon,N}=\varepsilon\sum_0^n Z_j\langle Z_j, \delta K_j^{\varepsilon,N}\rangle q_{Nj}^\varepsilon I_{jL}+\varepsilon\sum_0^n Z_j\langle Z_j, \delta K_j^{\varepsilon,N}\rangle q_{Nj}^\varepsilon I'_{jL}\equiv H_{1n}^L+H_{2n}^L.$$

We have

$$\begin{aligned}\textstyle\sum_{i=M}^\infty E\langle H_{1n}^L, e_i\rangle^2 &\leqq E|H_{1n}^L|^2=\varepsilon^2\sum_{j,k=0}^n E\langle Z_j, Z_k\rangle\langle Z_k, \delta K_k^{\varepsilon,N}\rangle\langle Z_j, \delta K_j^{\varepsilon,N}\rangle q_{Nj}^\varepsilon q_{Nk}^\varepsilon I_{jL}I_{kL}\\ &\leqq C_N\varepsilon^2\sum_{j,k=0}^n E|Z_j|^2|Z_k|^2 I_{jL}I_{kL}\leqq\delta_L,\end{aligned} \tag{3.5}$$

where $\delta_L\to 0$ as $L\to\infty$ (since $E|Z_j|^{4+\alpha}<\infty$ for some $\alpha>0$).

Next, evaluating $E\langle e_i, H_{2n}^L\rangle^2$ yields (using $E\langle e_i, Z_j\rangle^2=\lambda_i$ and the Schwarz inequality to get the next to last inequality)

$$\begin{aligned}&E\left[\varepsilon\sum_{j=0}^n \langle e_i, Z_j\rangle\langle Z_j, \delta K_j^{\varepsilon,N}\rangle q_{Nj}^\varepsilon I'_{jL}\right]^2\\ &\leqq\varepsilon^2\sum_{j,k=0}^n E\langle e_i, Z_j\rangle\langle e_i, Z_k\rangle q_{Nj}^\varepsilon q_{Nk}^\varepsilon\langle Z_j, \delta K_j^{\varepsilon,N}\rangle\langle Z_k, \delta K_k^{\varepsilon,N}\rangle I'_{jL}I'_{kL}\\ &\leqq\varepsilon^2 C_N\sum_{j,l=0}^n E|\langle e_i, Z_j\rangle\langle e_i, Z_k\rangle|L^2\\ &\leqq(n+1)^2\varepsilon^2 C_N\lambda_i L^2\leqq 2T^2L^2C_N\lambda_i.\end{aligned} \tag{3.6}$$

Since $\sum\lambda_i<\infty$, (3.5) and (3.6) imply that the first condition of Theorem 2.1 holds for the interpolation of $H_n^{\varepsilon,N}$.

We now need to verify the first condition of Theorem 2.1 for the second sum of (3.1), namely, $\varepsilon \sum_0^n Z_j\psi_j$. But, owing to the independence of $\{\psi_j\}$ and $z(\cdot)$, this is straightforward and the proof is omitted.

To complete the proof of tightness, only (2.2) needs to be shown. We have (possibly modulo an end term in the sum)

$$|K^{\varepsilon,N}(\tau+\delta)-K^{\varepsilon,N}(\tau)| = |\varepsilon \sum_{\tau/\varepsilon}^{(\tau+\delta)/\varepsilon} Z_j(\langle Z_j\, \delta K_j^{\varepsilon,N}\rangle - \psi_j) q_{Nj}^{\varepsilon}|.$$

Criterion (2.2) certainly holds if the Z_j and ψ_j are bounded. The general proof follows from this and the fact that $E|Z_j|^4<\infty$, $E|\psi_j|^2<\infty$, since for each $T<\infty$

$$\varepsilon E \sum_0^{T/\varepsilon} (|Z_j|^2+|Z_j\psi_j|) I_{\{|Z_j|\geq L \text{ or } |\psi_j|\geq L\}} \to 0$$

for each $L<\infty$.

Thus $\{K^{\varepsilon,N}(\cdot)\}$ is tight in $D_H[0,\infty)$. Also, any weak limit is strongly continuous w.p.l. if $\{\psi_j, Z(\cdot)\}$ is bounded. But, by a truncation argument such as the one used just above, we still have strong continuity under our conditions.

Part 3. *The limit* $\bar{K}^N(\cdot)$. Fix a weakly convergent subsequence of $\{K^{\varepsilon,N}(\cdot)\}$ in $D_H[0,\ )$, *indexed also by* ε and with limit $\bar{K}^N(\cdot)$. By taking a subsequence if necessary, suppose (w.l.o.g.) that the processes defined by the continuous parameter interpolation of the two sums in (3.1) also converge weakly in $D_H[0,\infty)$, and also to strongly continuous limits. The form of the limit process will not depend on the particular chosen subsequence. In order to show that $\bar{K}^N(\cdot)$ satisfies (3.2), it is sufficient to show that for each $f\in H$, the weak convergence in (3.7), (3.8) holds for the processes $\{S_1^{\varepsilon,N}(\cdot), S_2^{\varepsilon,N}(\cdot)\}$ defined there. Recall that $\langle f, K^{\varepsilon,N}(t)\rangle = \langle f, K_0\rangle + S_1^{\varepsilon,N}(t) - S_2^{\varepsilon,N}(t)$ and write $\delta\bar{K}^N(t) = \bar{K}^N(t) - K(t)$.

$$S_1^{\varepsilon,N}(t) = \varepsilon \sum_{j=0}^{t/\varepsilon} \psi_j \langle f, Z_j\rangle q_{Nj}^{\varepsilon} \Rightarrow \text{zero process}, \tag{3.7}$$

$$S_2^{\varepsilon,N}(t) = \varepsilon \sum_{j=0}^{t/\varepsilon} \langle f, Z_j\rangle \langle \delta K_j^{\varepsilon,N}, Z_j\rangle q_{Nj}^{\varepsilon}$$

$$\Rightarrow \int_0^t du \int_0^1 \int_0^1 f(v)\, \delta\bar{K}^N(u,s) q_N(\bar{K}^N(u)) R(v-s)\, dv\, ds = S_f^N(t). \tag{3.8}$$

The sequence $\{S_1^{\varepsilon,N}(\cdot)\}$ is tight in $D_R[0,\infty)$. Then to get (3.7), we only need to show that $S_1^{\varepsilon,N}(t) \overset{P}{\to} 0$ for each t (Billingsley [13, Thm. 15.1]). But, by the hypotheses, $E(S_1^{\varepsilon,N}(t))^2 \to 0$ as $\varepsilon\to 0$. Thus (3.8) holds.

The weak convergence of $\{K^{\varepsilon,N}(\cdot)\}$ and the hypotheses on $z(\cdot)$ imply tightness of $\{S_2^{\varepsilon,N}(\cdot)\}$ in $D_R[0,\infty)$, and that the limits are strongly continuous. Thus, to obtain (3.8), we need only to show that

$$\varepsilon \sum_0^{t/\varepsilon} \langle f, Z_j\rangle \langle \delta K_j^{\varepsilon,N}, Z_j\rangle q(K_j^{\varepsilon,N}) \overset{D}{\to} S_f^N(t) \tag{3.9}$$

for each $t>0$. To complete the proof a Skorokhod imbedding procedure will be used. Choose and imbed a weakly convergent subsequence of $\{K^{\varepsilon,N}(\cdot), S_2^{\varepsilon,N}(\cdot)\}$. Owing to the nature of the imbedding the random variables Z_j which are used to represent $S_1^{\varepsilon,N}(\cdot)$ depends on ε, although their distributions do not, and we rewrite the left side

of (3.9) as

$$\varepsilon \sum_{0}^{t/\varepsilon} \langle f, Z_j^\varepsilon\rangle\langle \delta K_j^{\varepsilon,N}, Z_j^\varepsilon\rangle q(K_j^{\varepsilon,N}). \tag{3.9'}$$

By the Skorokhod imbedding

$$\sup_{t\leqq T} |K^{\varepsilon,N}(t) - \bar{K}^N(t)| \to 0 \quad \text{w.p. 1.} \tag{3.10}$$

Thus the limits of (3.9') remain the same if $\delta K_j^{\varepsilon,N}$ and $K_j^{\varepsilon,N}$ are replaced by $\delta \bar{K}^N(\varepsilon j)$ and $\bar{K}^N(\varepsilon j)$, respectively. So, we need only show

$$\varepsilon \sum_{0}^{t/\varepsilon} \langle f, Z_j^\varepsilon\rangle\langle \delta \bar{K}^N(\varepsilon j), Z_j^\varepsilon\rangle q(\bar{K}^N(\varepsilon j)) \Rightarrow S_f^N(t), \tag{3.9''}$$

where Z_j^ε is obtained from a process $z^\varepsilon(\cdot)$, just as Z_j was obtained from $z(\cdot)$, and $z^\varepsilon(\cdot)$ and $z(\cdot)$ have the same distributions.

If $\bar{K}(\varepsilon j)$ were replaced by a constant in (3.9''), then (3.9'') would hold by the law of large numbers for stationary processes [15]. We get the same result, owing to the strong continuity of $\bar{K}^N(\cdot)$. Q.E.D.

The projected algorithm (1.3). By the nature of the projection, there are $c_n^\varepsilon \geqq 0$ such that $\varepsilon c_n^\varepsilon \leqq 1$ and algorithm (1.3) can be rewritten in the form

$$\begin{aligned} K_{n+1}^\varepsilon &= [K_n^\varepsilon - \varepsilon Z_n(\langle Z_n, \delta K_n^\varepsilon\rangle - \psi_n)](1 - \varepsilon c_n^\varepsilon) \\ &= K_n^\varepsilon - \varepsilon Z_n(\langle Z_n, \delta K_n^\varepsilon\rangle - \psi_n) - \varepsilon c_n^\varepsilon K_n^\varepsilon + O_j^\varepsilon. \end{aligned} \tag{3.11}$$

In fact, we have (write $\Delta_n^\varepsilon = Z_n(\langle Z_n, \delta K_n^\varepsilon\rangle - \psi_n)$),

$$\varepsilon c_n^\varepsilon \leqq \varepsilon |\Delta_n^\varepsilon|/(1 + \varepsilon |\Delta_n^\varepsilon|). \tag{3.12}$$

Thus $\sum_0^{T/\varepsilon} E|O_j^\varepsilon| \overset{\varepsilon}{\to} 0$, and we can (and will) ignore the O_j^ε terms. For future use, note that by the conditions on the noise in § 1 and for any sequence $T_\varepsilon \to \infty$ and $\alpha > 0$,

$$P\{\sup_{n\leqq T_\varepsilon/\varepsilon} \varepsilon c_n^\varepsilon \geqq \alpha\} \leqq \sum_{n=0}^{T_\varepsilon/\varepsilon} P\{\varepsilon c_n^\varepsilon \geqq \alpha\} = 0(\varepsilon/\alpha^2) T_\varepsilon.$$

Thus, we can (and will) suppose (by altering the algorithm on a set of arbitrarily small probability) that there is a sequence $T_\varepsilon \to \infty$ such that $\varepsilon c_n^\varepsilon < \frac{1}{2}$ for $n \leqq T_\varepsilon/\varepsilon$.

THEOREM 3.2. (algorithm (1.3)). *Let* $|\bar{K}(0)| = |K_0| < 1$. *Under the assumptions of Theorem* 3.1, $K^\varepsilon(\cdot) \Rightarrow \bar{K}(\cdot)$ *in* $D_H[0,\infty)$, *where* $(\delta\bar{K}(t) = \bar{K}(t) - K)$

$$\frac{d\bar{K}(t)}{dt} = \bar{\pi}_1[-R\delta\bar{K}(t)], \qquad \bar{K}(0) = \pi_1(K_0), \tag{3.13}$$

and the operator[2] $\bar{\pi}_1$ *projects the dynamics onto the unit ball in H. Equivalently, there is a real valued function* $\bar{c}(\cdot) \geqq 0$ *such that*

$$\frac{d\bar{K}(t)}{dt} = -R\delta\bar{K}(t) - \bar{c}(t)\bar{K}(t), \tag{3.14}$$

where $\bar{c}(t) = 0$ *if* $|\bar{K}(t)| < 1$. $\bar{c}(\cdot)$ *is just large enough to keep* $\bar{K}(\cdot)$ *from leaving the unit ball.*

[2] That is, if $|\bar{K}(t)| < 1$, then $\pi_1 =$ identity. If $|\bar{K}(t)| = 1$ and $d/dt\langle \bar{K}(t), \bar{K}(t)\rangle > 0$, then π_1 projects the derivative on to the tangent plane to the unit ball at the point $\bar{K}(t)$, so that $\bar{K}(\cdot)$ cannot leave the unit ball.

Remarks on the proof. Write (dropping the O_j^ε terms)

$$K_{n+1}^\varepsilon = \prod_0^{n+1} (1-\varepsilon c_j^\varepsilon) K_0 - \sum_0^n \prod_{j+1}^n (1-\varepsilon c_k^\varepsilon) \varepsilon Z_j(\langle Z_j, \delta K_j^\varepsilon\rangle - \psi_j). \tag{3.15}$$

Fix $T>0$. Using $\varepsilon c_n^\varepsilon \leqq 1$ and the representation (3.15), the tightness proof in Theorem 3.1 carries over with only minor modifications. The rest of the details are also about the same as for Theorem 3.1, and we make only a few comments. Define the interpolated process $C^\varepsilon(t) = \varepsilon \sum_0^{t/\varepsilon} c_j^\varepsilon$. Then, by the conditions on $z(\cdot)$ and ψ_n, $\{C^\varepsilon(\cdot)\}$ is tight in $D_R[0,\infty)$ and all weak limits $\bar{C}(\cdot)$ are Lipschitz continuous. This follows from $(t>s)$

$$\bar{C}(t) - \bar{C}(s) \leqq \text{weak limit of} \sum_{s/\varepsilon}^{t/\varepsilon} \varepsilon[|Z_j|^2 + |Z_j\psi_j|], \tag{3.16}$$

and the fact that the weak limit of the function defined on the right of (3.16) is of the form $c_0 t$.

Fix a weakly convergent subsequence of $\{K^\varepsilon(\cdot), C^\varepsilon(\cdot)\}$ with limit $\bar{K}(\cdot), \bar{C}(\cdot)$, and write $\bar{C}(t) = \int_0^t \bar{c}(s)\, ds$. We have $|\bar{K}(t)| \leqq 1$. By a Skorokhod imbedding and the consequent uniform convergence (as in (3.7) for the problem of Theorem 3.1), $\bar{c}(t) = 0$ on any interval on which $|\bar{K}(t)| < 1$. It follows that $\bar{c}(\cdot)$ must be just large enough to keep $\bar{K}(\cdot)$ from leaving the unit sphere. With this $\bar{c}(\cdot)$, the solution to (3.14) is unique, hence the chosen subsequence is unimportant and the original sequence converges weakly to the solution to (3.14).

Discretization. For application, it is important to understand the behavior of a discretized algorithm. Let $a = (\varepsilon, \beta)$, where $\beta > 0$. Define $z^a(t) = \int_{l\beta-\beta}^{l\beta} z(s)\, ds/\beta$ on $t \in [l\beta, l\beta+\beta)$. Define Z_n^a as before. Then $E|Z_j^a|^{4+\alpha} \leqq E|Z_j|^{4+\alpha} < \infty$, and Assumptions 3.1 and 3.2 imply that, if $m(\cdot)$ is the modulus of continuity of $R(\cdot)$,

$$\overline{\lim_{s\uparrow\infty}}\, E|E[z^a(t+s)z(t+s-v)|z(u), u \leqq t] - R(v)| \leqq m(2\beta) \tag{3.17}$$

and the limit is uniform in t and $v \in [-1, 1]$. Define

$$K_{n+1}^a(u) = K_n^a(u) - z^a(n\Delta - u)\left[\int_0^1 K_n^a(s) z(n\Delta - s)\, ds - y_n\right], \quad K_0^a(u) = K_0(u). \tag{3.18}$$

This gives a discretized version in that, as u ranges over $[0, 1]$, $K_{n+1}^a(u)$ takes a finite number of values. Define $K^a(t, u)$ and $K^a(t)$ as before. Let $a = (\varepsilon, \beta) \to (0, 0)$. Theorem 3.3 shows that $K^a(\cdot) \Rightarrow \bar{K}(\cdot)$ satisfying (1.5).

THEOREM 3.3. *Under the assumptions of* § 1 *and Assumptions* 3.1 *and* 3.2 $K^a(\cdot) \Rightarrow \bar{K}(\cdot)$ *in* $D_H[0,\infty)$, *where* $\bar{K}(\cdot)$ *satisfies* (1.5).

Proof. The proof is very similar to the proof of Theorem 3.1, and we only make some remarks on the tightness (Part 2 of Theorem 3.1). Define $I_{jL} = I_{\{|Z_j| \leqq L\}}$ as before. Then the estimate for H_{1n}^L holds, by the uniform integrability of $|Z_j|^4$ and the integrability of $|Z_j^a|^4$. To get an estimate for H_{2n}^L, note that (cf. (3.6))

$$\sum_{i=M}^\infty \varepsilon^2 C_N \sum_{j,l=0}^n E|\langle e_i, Z_j^a\rangle\langle e_i, Z_k^a\rangle| L^2$$
$$\leqq n\varepsilon^2 C_N L^2 \sum_{j=0}^n \left(\sum_{i=M}^\infty E\langle e_i, Z_j\rangle^2 + |E|Z_j|^2 - E|Z_j^a|^2| \right).$$

A straightforward computation shows that

$$|E|Z_j|^2 - E|Z_j^a|^2| \leqq m(2\beta).$$

This and the argument in (3.6) gives the desired estimate. Q.E.D.

4. Asymptotic behavior, $\varepsilon \to 0$ and $n\varepsilon \to \infty$ for algorithm (1.3). The behavior of (1.3) as $n\varepsilon \to \infty$ and $\varepsilon \to 0$ cannot be ascertained directly from the weak convergence results of Theorems 3.1 and 3.2. In this section, we show that if $|K| > 1$, then $K_n^\varepsilon \Rightarrow \bar{K}_\infty$ in H, where $\bar{K}_\infty$ is the unique minimizer in (1.4). From this and the arguments of Theorems 3.1 and 3.2, we have $K^\varepsilon(t_\varepsilon + \cdot) \Rightarrow$ constant function $\bar{K}_\infty$, if $t_\varepsilon \to \infty$. So far, we have not been able to obtain the analogous result when $|K| < 1$ (mainly because we have not been able to prove tightness of $\{K_n^\varepsilon, \varepsilon > 0, n < \infty\}$ in H for that case).

Write $J(\hat{K}) = \frac{1}{2}\langle \hat{K} - K, R(\hat{K} - K)\rangle$. Just as in the finite-dimensional case, the strict convexity and continuity of $J(\cdot)$ and the convexity of the unit sphere imply that $J(\cdot)$ has a unique minimizer on the unit sphere and also that: if there is a $|\hat{K}| < 1$ such that $R(\hat{K} - K) = 0$, or a $(\hat{K}, \hat{c})$ such that $\hat{c} \geqq 0$, $|\hat{K}| = 1$ and $R(\hat{K} - K) + \hat{c}\hat{K} = 0$, then $\hat{K}$ is a global constrained minimum. Thus, there is a unique solution (on the boundary of the unit sphere) to (4.2) below. The $\bar{c}_\infty$ in (4.2b) is determined by $\langle \bar{K}_\infty, R(\bar{K}_\infty - \bar{K})\rangle + \bar{c}_\infty |\bar{K}_\infty|^2 = 0$, and using $|\bar{K}_\infty|^2 = 1$.

The Fréchet derivative of $J(\hat{K})$ is $R(\hat{K} - K)$. The steepest descent algorithm for the constrained minimization of $J(\cdot)$ takes the form $\dot{\hat{K}} = -R(\hat{K} - K)$ for $|\hat{K}| < 1$. When $|\hat{K}| = 1$, it is $\dot{\hat{K}} = -R(\hat{K} - K) - \hat{c}\hat{K}$, where $\hat{c}$ is just large enough to keep $\hat{K}$ from leaving the unit sphere. Thus the steepest descent path obeys (3.13), (3.14). These facts will be heavily used in the proof of Theorem 4.1.

THEOREM 4.1. *Assume algorithm* (1.3) *and the conditions of Theorem* 3.2, *but let* $|K| > 1$. *Then, if* $t_\varepsilon \to \infty$ *as* $\varepsilon \to 0$, $K^\varepsilon(t_\varepsilon) \to \bar{K}_\infty$, *the stationary solution to* (3.13) *or* (3.14), *where* $\bar{K}_\infty$ *minimizes in*

$$\min_{|\hat{K}| \leqq 1} J(\hat{K}) \tag{4.1}$$

and is the unique solution to

$$\bar{K}_\infty = (R + \bar{c}_\infty)^{-1} RK, \tag{4.2a}$$

$$\bar{c}_\infty = -\langle \bar{K}_\infty - K, R\bar{K}_\infty \rangle. \tag{4.2b}$$

Proof. Part 1. Let $t_\varepsilon \to \infty$. Define the interpolation $K^\varepsilon(t_\varepsilon + \cdot)$ as usual for $t_\varepsilon + t \geqq 0$, and set $K^\varepsilon(t) = K_0$ for $t_\varepsilon + t \leqq 0$. If there is a random variable $\bar{K}(0)$ such that $K^\varepsilon(t_\varepsilon) \Rightarrow \bar{K}(0)$ in H, then the argument of Theorem 3.1 implies that $K^\varepsilon(t_\varepsilon + \cdot) \Rightarrow \bar{K}(\cdot)$, in $D_H(-\infty, \infty)$, with initial condition $\bar{K}(0)$. In any case, since $|K_n^\varepsilon| \leqq 1$, $\{K_n^\varepsilon\}$ is tight in H_w. Let (choose a subsequence, if necessary) $K_n^\varepsilon \Rightarrow \bar{K}(0)$ in H_w. Then by the arguments of Theorems 3.1 and 3.2, $K^\varepsilon(t_\varepsilon + \cdot) \Rightarrow \bar{K}(\cdot)$ in $D_{H_w}(-\infty, \infty)$, and also $\bar{K}(\cdot)$ satisfies (3.13), (3.14). Actually, to prove (3.9) in Theorem 3.1, norm convergence of $\{K^{\varepsilon,N}(\cdot)\}$ was used. But convergence in $D_{H_w}[0, \infty)$ is enough. To see this drop the $q_j^{\varepsilon,N}$, which do not appear in the current case and note that the summands in (3.9) are uniformly (in j, ε) integrable. Then, it is enough to show that for each $\delta < 0$

$$P\left\{\left|\left\langle \delta K_j^\varepsilon, \sum_{i=N}^{\infty} z_i^j e_i \right\rangle\right| > \delta\right\} \to 0$$

as $N \to \infty$, uniformly in ε and j, where $z_i^j = (e_i, Z_j)$. But this follows from the norm boundedness of $\{\delta K_j^\varepsilon\}$ and

$$E\left\langle \sum_{i=N}^{\infty} z_i^j e_i, \sum_{i=N}^{\infty} z_i^j e_i \right\rangle = \sum_{N}^{\infty} \lambda_i \overset{N}{\to} 0.$$

In Part 3 below, it is prove that $\{K_n^\varepsilon, \varepsilon > 0, n \geqq 0\}$ is tight in H. This result implies that $\{K^\varepsilon(t_\varepsilon + \cdot)\}$ is tight in $D_H(-\infty, \infty)$, and also that the set $\{\bar{K}(t), |t| < \infty\}$ of all possible weak limits in H (over all subsequences ε and sequences $\{t_\varepsilon\}$) is tight in H.

Until Part 3, suppose that $\{K_n^\varepsilon$, small $\varepsilon > 0$, $n \geqq 0\}$ is tight in H. Fix $t_\varepsilon \to \infty$ and a subsequence (indexed by ε) such that $\{K^\varepsilon(t_\varepsilon + \cdot), C^\varepsilon(t_\varepsilon + \cdot)\}$ converges weakly in $D_H(-\infty, \infty) \times D_R(-\infty, \infty)$ to the pair $(\bar{K}(\cdot), \bar{C}(\cdot))$, and write $\bar{c}(t) = \dot{\bar{C}}(t)$. This pair satisfies (3.13), (3.14) on $(-\infty, \infty)$.

By the weak convergence and the continuity of the limit processes, there are $\alpha_\varepsilon \to 0$ and $T_\varepsilon \to \infty$ such that

$$\begin{aligned} &\lim_\varepsilon P\{\sup_{|t| \leqq T_\varepsilon} |K^\varepsilon(t_\varepsilon + t) - \bar{K}(t)| \geqq \alpha_\varepsilon\} = 0, \\ &\lim_\varepsilon P\{\sup_{|t| \leqq T_\varepsilon} |C^\varepsilon(t_\varepsilon + t) - \bar{C}(t)| \geqq \alpha_\varepsilon\} = 0. \end{aligned} \tag{4.3}$$

Equation (4.3) (and the fact that $c_n^\varepsilon = 0$ if $|K_{n+1}^\varepsilon| < 1$) implies that $\bar{c}(t) = 0$ on any interval on which $|\bar{K}(t)| < 1$. Also, as in Theorem 3.2, if $|\bar{K}(t)| = 1$ on an interval, then $\bar{c}(t)$ must be just large enough to keep $\bar{K}(t)$ from leaving the unit sphere on that interval. These properties imply that the solution to (3.14) is a steepest descent path for the constrained minimization of $J(\cdot)$. We now proceed to show that $\bar{K}(t) = \bar{K}_\infty$, the unique stationary point. The analysis uses only the assertion that $\{\bar{K}(t), |t| < \infty\}$ *is in a strongly compact set* (to be proved below), and the steepest descent property of (3.13), (3.14).

For $|\bar{K}(t)| < 1$ (hence $\bar{c}(t) = 0$),

$$\dot{J}(\bar{K}(t)) = -\langle R(\bar{K}(t) - K), R(\bar{K}(t) - K)\rangle \leqq 0, \tag{4.4}$$

which is bounded away from zero, uniformly in $|\bar{K}(t)| < 1$. This implies that $\bar{K}(\cdot)$ eventually stays on the surface of the unit sphere, and that the *total time* during which $|\bar{K}(t)| < 1$ is *bounded uniformly in the initial condition* $\bar{K}(0)$. Define $\bar{k}_i(\cdot)$ and k_i by the expansions

$$\bar{K}(t) = \sum_1^\infty \bar{k}_i(t) e_i, \qquad K = \sum_1^\infty k_i e_i.$$

If $\bar{K}(t)$ is on the boundary $|\bar{K}(t)| = 1$ on some interval, then (on that interval) obviously

$$0 = |\dot{\bar{K}}(t)|^2 = -2\langle \bar{K}(t), R(\bar{K}(t) - K) + \bar{c}(t)\bar{K}(t)\rangle,$$

which yields the value

$$\bar{c}(t) = -\langle R\bar{K}(t), \bar{K}(t) - K\rangle = -\langle \bar{K}(t), R(\bar{K}(t) - K)\rangle. \tag{4.5}$$

By a similar direct calculation on the boundary and use of (4.5)

$$\dot{J}(\bar{K}(t)) = -\langle R(\bar{K}(t) - K), R(\bar{K}(t) - K)\rangle + \langle R(\bar{K}(t) - K), \bar{K}(t)\rangle^2 \leqq 0. \tag{4.6a}$$

In terms of components,

$$\dot{J}(\bar{K}(t)) = -\sum \lambda_i^2 (\bar{k}_i(t) - k_i)^2 + \left(\sum_i \lambda_i (\bar{k}_i(t) - k_i)\bar{k}_i(t)\right)^2 \leqq 0. \tag{4.6b}$$

Expression (4.6) is $\leqq 0$ because it is the derivative along a steepest descent path. It is also < 0 if $\bar{K}(t) \neq \bar{K}_\infty$ and it is strongly continuous on each strongly compact set.

Since $\{\bar{K}(t), |t| < \infty\}$ is assumed to be in a strongly compact set (see below for proof) the comments in the last paragraph imply that for each $\delta > 0$, there is a (strong) neighborhood N_δ of $\bar{K}_\infty$ such that for $\bar{K}(t) \notin N_\delta$, $\dot{J}(\bar{K}) \leqq -\delta$. This implies that the path $\bar{K}(\cdot)$ can only spend a *finite total time out of* N_δ, and that this total time is bounded uniformly in the chosen convergent subsequence. The strong compactness and the decreasing property of $\langle \bar{K} - K, R(\bar{K} - K)\rangle$ along paths of (3.13), (3.14) imply that

eventually (the required time T_δ, say, being independent of the chosen subsequence) $\bar{K}(\cdot)$ will remain in N_δ. The above argument actually implies that $\bar{K}(t) = \bar{K}_\infty$, by the following reasoning: Let $K^\varepsilon(t_\varepsilon + \cdot) \Rightarrow \bar{K}(\cdot)$. Then $\bar{K}(t) \in N_\delta$ for $t \geqq T_\delta$. Also for any $T > 0$, $\{K^\varepsilon(t_\varepsilon - T_\delta - T + \cdot) \Rightarrow \bar{K}_{T_\delta}(\cdot) = \bar{K}(\cdot - T_\delta - T)$ and $\bar{K}(t) = \bar{K}_{T_\delta}(T_\delta + T + t) \in N_\delta$ for all $t \geqq -T$ by the above argument.

It only remains to prove the tightness of $\{K_n^\varepsilon$, small $\varepsilon > 0$, $n \geqq 0\}$ and the strong compactness assumption made above (4.4).

For arbitrary $T > 0$, let $\bar{K}(\cdot)$ and $\bar{K}_T(\cdot)$ be limits (in $D_{H_w}(-\infty, \infty)$) of $\{K^\varepsilon(t_\varepsilon + \cdot)\}$ and $\{K^\varepsilon(t_\varepsilon - T + \cdot)\}$, respectively. Write

$$\bar{K}(t) = \sum \bar{k}_i(t) e_i, \quad \bar{K}_T(t) = \sum \bar{k}_{T,i}(t) e_i, \quad K = \sum k_i e_i.$$

By (3.14), $|\bar{k}_{T,i}(t)| \leqq 1$ and

$$|\bar{k}_i(t)| = |\bar{k}_{T,i}(t+T)| \leqq |k_i| + |(1-k_i)\, e^{-\lambda_i T}|,$$

since $c(t) \geqq 0$. Since T is arbitrary, $|\bar{k}_i(t)| \leqq |k_i|$ for all i and t. This implies the strong compactness.

Part 2. In preparation for Part 3, we now get an estimate for $\prod_{t/\varepsilon}^{(t+s)/\varepsilon} (1 - \varepsilon c_n^\varepsilon)$ as $\varepsilon \to 0$, $t \to \infty$. Let $\{\sigma_\varepsilon\}$ be a sequence which might tend to ∞. As noted above Theorem 3.2, we can suppose, for $T_\varepsilon \to \infty$ slowly enough, that $\varepsilon c_j^\varepsilon \leqq 1/2$ for $\varepsilon j \in [\sigma_\varepsilon - T_\varepsilon, \sigma_\varepsilon + T_\varepsilon]$. By the hypotheses, the weak convergence of $C^\varepsilon(\cdot)$ to $\bar{C}(\cdot)$ in $D_R(-\infty, \infty)$, and the weak convergence $\sum_{\sigma_\varepsilon/\varepsilon}^{(\sigma_\varepsilon+t)/\varepsilon} |\varepsilon c_j^\varepsilon|^2 \Rightarrow 0$, there are $T_\varepsilon \to \infty$ such that

$$\sup_{0 \leqq t \leqq T_\varepsilon} \left| \prod_{\sigma_\varepsilon/\varepsilon}^{(\sigma_\varepsilon+t)/\varepsilon} (1 - \varepsilon c_j^\varepsilon) / \exp - \int_0^t \bar{c}(u)\, du \right| \Rightarrow 1. \tag{4.7}$$

A similar result holds when $\prod_{(\sigma_\varepsilon - t)/\varepsilon}^{\sigma_\varepsilon/\varepsilon}$ and $\exp - \int_{-t}^0 \bar{c}(u)\, du$ are used.

We now estimate $\bar{c}(\cdot)$. Order $\{e_i\}$ such that the λ_i are nonincreasing. Since $|K|^2 = \sum k_i^2 > 1$, we can find $\beta > 0$ and $m < \infty$ such that $\sum_1^m k_i^2 \geqq 1 + \beta$. Define $|x|_m^2 = \sum_1^m \langle e_i, x \rangle^2$ and set

$$K_m = \{k_1, \cdots, k_m\}', \quad \bar{K}_m(t) = \{\bar{k}_1(t), \cdots, \bar{k}_m(t)\}', \quad \Lambda_m = \operatorname{diag}(\lambda_1, \cdots, \lambda_m).$$

Then, by (3.13), (3.14),

$$\dot{\bar{K}}_m = -\Lambda_m \bar{K}_m - \bar{c}\bar{K}_m + \Lambda_m K_m,$$

$$\bar{K}_m(t) - \bar{K}_m(s) = -\Lambda_m \int_s^t (\bar{K}_m(u) - K_m)\, du - \int_s^t \bar{c}(u) \bar{K}_m(u)\, du,$$

where $|\bar{K}_m(t)| \leqq 1$, $|K_m|^2 \geqq 1 + \beta$. Thus

$$2 \geqq |\bar{K}_m(t) - \bar{K}_m(s)| \geqq \lambda_m \left| \int_0^t [\bar{K}_m(u) - K_m]\, du \right| - \int_0^t \bar{c}(u)\, du.$$

This implies $\int_s^t c(s)\, ds \geqq \alpha_1 (t-s) - 2$, for $\alpha_1 = \lambda_m \beta$, a bound which is independent of the chosen convergent subsequence.

Combining this last result and (4.7) yields that for some $T_\varepsilon \to \infty$ (and also for $-T_\varepsilon \leqq t \leqq 0$),

$$\overline{\lim_\varepsilon} \sup_{t \leqq T_\varepsilon} \frac{\prod_{\sigma_\varepsilon/\varepsilon}^{(\sigma_\varepsilon+t)/\varepsilon} (1 - \varepsilon c_j^\varepsilon)}{(\exp - \alpha_1 t) e^2} \leqq 1. \tag{4.8}$$

Part 3. We now show tightness in H of $\{K_n^\varepsilon$, small $\varepsilon > 0$, $t < \infty\}$. For each $T > 0$, $\{K^\varepsilon(t)$, $t \leqq T$, $\varepsilon > 0\}$ is tight in H. *Hence, we need only show that for each sequence* $s_\varepsilon \to \infty$, *the set* $\{K^\varepsilon(s_\varepsilon)\}$ *is tight in* H. Let $S_\varepsilon \to \infty$ such that $S_\varepsilon \leqq T_\varepsilon$ (the T_ε used in Part

2) and $s_\varepsilon \geqq S_\varepsilon$. We have (possibly moduly an "end" term in each sum and product)

$$K^\varepsilon(s_\varepsilon) = \prod_{(s_\varepsilon - S_\varepsilon)/\varepsilon}^{s_\varepsilon/\varepsilon} (1-\varepsilon c_i^\varepsilon) K^\varepsilon(s_\varepsilon - S_\varepsilon) + \sum_{j=(s_\varepsilon - S_\varepsilon)/\varepsilon}^{s_\varepsilon/\varepsilon} \prod_{i=j+1}^{s_\varepsilon/\varepsilon} (1-\varepsilon c_i^\varepsilon) Z_j(\langle Z_j, \delta K_j^\varepsilon\rangle - \psi_j)$$

$$= \Delta_1^\varepsilon + \Delta_2^\varepsilon. \tag{4.9}$$

By (4.8), $|\Delta_1^\varepsilon| \Rightarrow 0$. The estimate (4.8) and the eigenfunction expansions used in Theorem 3.1 yield that for each $\delta > 0$, there is an $M < \infty$ such that

$$\overline{\sup_\varepsilon} P\left\{\sum_M^\infty \langle e_i, \Delta_2^\varepsilon\rangle^2 \geqq \delta\right\} \leqq \delta.$$

The last two sentences and the argument used to show that (2.10) implies (2.8) yield the tightness of $\{K^\varepsilon(s_\varepsilon)\}$ in H. Q.E.D.

5. Weak convergence to an H-valued Wiener process. In this section we develop some technical results which will be used in the analysis of the normalized processes introduced in § 6. That analysis requires the weak convergence in $D_H[0, \infty)$ of a certain sequence $\{W^\varepsilon(\cdot), \hat{W}^\varepsilon(\cdot)\}$ to H-valued Wiener processes $W(\cdot)$, $\hat{W}(\cdot)$, resp., and the proofs are given here.

DEFINITIONS [16], [17]. The covariance Q of an H-valued random variable Y is an operator from H to H defined by

$$Qf = EY\langle f, Y\rangle \equiv E[Y \circ Y]f.$$

Let $\tilde{R}(\cdot, \cdot)$ be a continuous function and define the operator $\tilde{R}$ on H by $\tilde{R}f(t) = \int_0^1 \tilde{R}(t, s)f(s)\, ds$, and let $\{\tilde{e}_i, \tilde{\lambda}_i\}$ denote its eigenfunctions and eigenvalues. Then by Mercer's theorem [17], [18] (the convergence is uniform on $[0, 1]^2$)

$$\tilde{R}(t, s) = \lim_N \sum_1^N \tilde{\lambda}_i \tilde{e}_i(t)\tilde{e}_i(s). \tag{5.1}$$

A process $W(\cdot)$ is a zero mean (stationary increment) H-valued Wiener process if there are mutually independent real valued, zero mean, Wiener processes $\{W_i(\cdot)\}$ with covariances $t\rho_i$ such that $\sum \rho_i < \infty$ and

$$\sum_1^\infty W_i(t)q_i = W(t) \tag{5.2}$$

where $\{q_i\}$ is orthonormal [16], [17]. The covariance operator of $W(t)$ is defined by $E\langle W(t), f\rangle\langle W(t), g\rangle = t\langle g, Qf\rangle = t\sum_i \rho_i\langle q_i, f\rangle\langle q_i, g\rangle$.

Convergence with $\{\psi_n\}$ i.i.d. Define $\sigma_\psi^2 = E\psi_n^2$.

THEOREM 5.1. *Under the assumptions of* § 1 *and $\{\psi_n\}$ i.i.d., the sequence of interpolations defined by $W^\varepsilon(t) = \sqrt{\varepsilon}\sum_0^{t/\varepsilon} \psi_n Z_n$ converges weakly in $D_H[0, \infty)$ to the zero mean Wiener process $W(\cdot)$ with covariance operator tQ, where*

$$\langle Qf, g\rangle = \sigma_\psi^2 \int_0^1 \int_0^1 R(u-s)f(u)g(s)\, du\, ds.$$

Proof. Tightness of $\{W^\varepsilon(\cdot)\}$. We use Theorem 2.1. To verify the first condition of that theorem, evaluate (let $t/\varepsilon =$ integer)

$$E\langle e_i, W^\varepsilon(t)\rangle^2 = t\sigma_\psi^2 \int_0^1 \int_0^1 e_i(s)e_i(u)R(u-s)\, ds\, du = \lambda_i t\sigma_\psi^2.$$

Since $\sum \lambda_i < \infty$, the first condition of Theorem 2.1 holds. To check (2.2) or (2.9), evaluate

(use the i.i.d. of $\{\psi_n\}$ and its independence of $z(\cdot)$)

$$E\left|\sqrt{\varepsilon}\sum_{\tau/\varepsilon}^{(\tau+\delta)/\varepsilon}\psi_n Z_n\right|^2 = \sigma_\psi^2 \varepsilon E\sum_{\tau/\varepsilon}^{(\tau+\delta)/\varepsilon}\langle Z_n, Z_n\rangle.$$

This and the assumption $E|Z_n|^4<\infty$ yield (2.9). (See an analogous argument at the end of Part 2 of the proof of Theorem 3.1.) Thus $\{W^\varepsilon(\cdot)\}$ is tight in $D_H[0,\infty)$.

Weak convergence. Now fix a weakly convergent subsequence, indexed by ε and with limit $W(\cdot)$. By the tightness and the definition (5.2) of an H-valued Wiener process, it is sufficient to show that for each m, $\{\langle e_i, W(\cdot)\rangle, i\leqq m\}=\{W_i(\cdot), i\leqq m\}$ are mutually independent zero mean Wiener processes, with cov $W_i(t)=t\lambda_i\sigma_\psi^2$. Then $W(\cdot)=\sum W_i(\cdot)e_i$ is an H-valued Wiener process with covariance operator Qt.

Define $W_i^\varepsilon(t)=\langle e_i, W^\varepsilon(t)\rangle$. By the weak convergence of $W^\varepsilon(\cdot)\Rightarrow W(\cdot)$ in $D_H[0,\infty)$, we have the weak convergence $\{W_i^\varepsilon(\cdot), i\leqq m\}\Rightarrow\{W_i(\cdot), i\leqq m\}$ in $D_{R^m}[0,\infty)$. The weak convergence of $\{W_i(\cdot), i\leqq m\}$ is also a standard problem in weak convergence of Euclidean space valued processes. The perturbed test function method of [9], [20] can readily be used to obtain that $\{W_i(\cdot), i\leqq m\}$ is the desired Wiener process with covariance $t\cdot\operatorname{diag}(\lambda_1,\cdots,\lambda_m)\sigma_\psi^2$, under our hypotheses. Since the limit distributions do not depend on the chosen subsequence, $W^\varepsilon(\cdot)$ converges weakly in $D_H[0,\infty)$ to a Wiener process $\sum_i e_i W_i(\cdot)$, where $\{W_i(\cdot)\}$ are zero mean mutally independent Wiener processes with $EW_i^2(t)=\lambda_i\sigma_\psi^2 t$. We now obviously have the covariance operator defined by (see Mercer's theorem, equation (5.1))

$$E\langle f, W(t)\rangle\langle g, W(t)\rangle = \sigma_\psi^2 t\sum_i \lambda_i\int_0^1\int_0^1 e_i(s)e_i(u)f(s)g(u)\,ds\,du$$

$$= \sigma_\psi^2 t\int_0^1\int_0^1 R(u-s)f(s)g(u)\,ds\,du.$$

Non i.i.d. $\{\psi_n\}$. We now drop the i.i.d. assumption on $\{\psi_n\}$. In order to keep the exposition simple, we use Assumption 5.1, which is much stronger than needed. Let $\{z_n\}$ denote a real valued process, and set $\mathscr{F}_n^m=\sigma\{z_i, n\leqq i\leqq m\}$. The process is called ϕ-mixing [13] with rate $\{\phi_n\}$ if for each m and $A\in\mathscr{F}_{-\infty}^n$ and $B\in\mathscr{F}_{n+m}^\infty$,

$$|P\{A\cap B\}-P\{A\}P\{B\}|\leqq\phi_m P\{A\}, \text{ for all } n.$$

An ergodic finite state Markov chain is ϕ-mixing and $\phi_n\to 0$ geometrically. A similar definition of ϕ-mixing applies to a continuous parameter process $z(\cdot)$.

Assumption 5.1. $\{\psi_n\}$ is bounded, stationary and ϕ-mixing with $\sum\phi_i^{1/2}<\infty$. $z(\cdot)$ is either Gaussian and stationary and with $R(s)\to 0$ exponentially, or it is bounded, stationary and ϕ-mixing with rate $\int_0^\infty\phi^{1/2}(u)\,du<\infty$.

We will use the fact that if $\{z_n\}$ is ϕ-mixing and bounded by, say, L, then [19] $|E[z_{n+m}-Ez_{n+m}|\mathscr{F}_{-\infty}^n]|\leqq 2L\phi_m$ (and similarly for a continuous parameter ϕ-mixing process $z(\cdot)$).

Define $R_\psi(j)=E\psi_n\psi_{n+j}$. Note that if $|\psi_n|\leqq A$, then under ϕ-mixing $R_\psi(j)\leqq 2A^2\phi_j$.

THEOREM 5.2. *Theorem* 5.1 *holds if Assumption* 5.1 *replaces the i.i.d. condition on* $\{\psi_n\}$ *and the covariance operator* Q *is replaced by* $\bar{Q}$, *where*

$$\langle\bar{Q}f, g\rangle=\int_0^1\int_0^1\bar{R}(u-s)f(u)g(s)\,du\,ds$$

and

$$\bar{R}(s)=\sum_{-\infty}^{\infty}R_\psi(j)R(j\Delta+s),\qquad |s|\leqq 1.$$

Proof. Tightness of $\{W^\varepsilon(\cdot)\}$. Theorem 2.1 will be used. For each t, let $W^\varepsilon(t, s)$, $s \in [0, 1]$, denote the point values of the H-valued random variable $W^\varepsilon(t)$. For each *fixed* ε *and* t (t is a parameter here—not a time variable of the covariance function), define the covariance function (time parameter s),

$$\bar{R}^{\varepsilon,t}(s, u) = EW^\varepsilon(t, u)W^\varepsilon(t, s)$$

$$= \varepsilon \sum_{j,k=0}^{t/\varepsilon\,\prime\prime} R_\psi(j-k)R(j\Delta - k\Delta + s - u) = \bar{R}^{\varepsilon,t}(s-u, 0).$$

By Assumption 5.1,

$$\bar{R}^{\varepsilon,t}(s, 0) \to t\bar{R}(s) \quad \text{as } \varepsilon \to 0, \tag{5.3}$$

uniformly in $|s| \leqq 1$ and on each bounded t interval. Thus $\bar{R}(\cdot)$ is a continuous covariance function, since it is a uniform limit of a sequence of continuous covariance functions.

Let $\{\bar{e}_i, \bar{\lambda}_i, i \geqq 1\}$ and $\{\bar{e}_i^{\varepsilon,t}, \bar{\lambda}_i^{\varepsilon,t}, i \geqq 1\}$ denote the eigenfunctions and values of the operators (on H) $\bar{R}$ and $\bar{R}^{\varepsilon,t}$, resp., (corresponding to the covariances $\bar{R}(\cdot)$ and $\bar{R}^{\varepsilon,t}(\cdot,\cdot)$, resp.). We first prove an approximation and convergence result for these eigenfunctions and values. By Mercer's theorem [17], [18],

$$\bar{R}^{\varepsilon,t}(s, s) = \bar{R}^{\varepsilon,t}(0, 0) = \sum_1^\infty \bar{\lambda}_i^{\varepsilon,t}(\bar{e}_i^{\varepsilon,t}(s))^2. \tag{5.4}$$

Also,

$$\bar{\lambda}_i^{\varepsilon,t} \to t\bar{\lambda}_i, \quad \bar{e}_i^{\varepsilon,t} \to \bar{e}_i \quad \text{in } H, \text{ as } \varepsilon \to 0. \tag{5.5}$$

Equation (5.5) follows from the general method of construction of the eigenfunctions and values of a covariance operator, and is proved roughly as follows. Wong [18, p. 83-85] gives a detailed outline of a method for the construction of the eigenfunctions and values of a covariance operator. By the method of construction there, $\bar{\lambda}_1^{\varepsilon,t} \to t\bar{\lambda}_1$ as $\varepsilon \to 0$. Suppose that $\bar{\lambda}_j^{\varepsilon,t} \to t\bar{\lambda}_j$, $j \leqq i$ and $\bar{e}_j^{\varepsilon,t} \to \bar{e}_j$, $j < i$ (true for $i = 1$). The operator $\bar{R}$ is strongly compact on H, and all function convergences below are in H. Use the compactness of $\bar{R}$ to get a function ϕ_i (choose a subsequence, if necessary, but the argument below implies that the original sequence itself converges) $\bar{R}\bar{e}_i^{\varepsilon,t} \to \phi_i$. Then

$$(t\bar{R} - \bar{R}^{\varepsilon,t})\bar{e}_i^{\varepsilon,t} \to 0,$$

$$\lim_\varepsilon (t\bar{R} - \bar{R}^{\varepsilon,t})\bar{e}_i^{\varepsilon,t} = \lim_\varepsilon (t\bar{R}\bar{e}_i^{\varepsilon,t} - \bar{\lambda}_i^{\varepsilon,t}\bar{e}_i^{\varepsilon,t}) = \lim_\varepsilon (t\phi_i - t\bar{\lambda}_i\bar{e}_i^{\varepsilon,t}).$$

Thus $\phi_i = \lim_\varepsilon \bar{\lambda}_i\bar{e}_i^{\varepsilon,t} = \bar{\lambda}_i\bar{e}_i$. These calculations imply that $\bar{e}_i^{\varepsilon,t} \to \bar{e}_i$. This, in turn, implies (by the method of construction of the eigenvalues outlined in [18]) that $\bar{\lambda}_{i+1}^{\varepsilon,t} \to t\bar{\lambda}_i$. Thus (5.5) follows by an induction argument.

Equations (5.3) to (5.5) imply that

$$\bar{R}^{\varepsilon,t}(0, 0) = \sum_i \bar{\lambda}_i^{\varepsilon,t} \to t\sum_i \bar{\lambda}_i = t\bar{R}(0). \tag{5.6}$$

Also,

$$E\sum_1^\infty \langle \bar{e}_i^{\varepsilon,t}, W^\varepsilon(t)\rangle^2 = E\sum_1^\infty \langle \bar{e}_i, W^\varepsilon(t)\rangle^2 = E|W^\varepsilon(t)|^2 \to t\bar{R}(0) \tag{5.7}$$

and

$$\lim_\varepsilon E\langle \bar{e}_i^{\varepsilon,t}, W^\varepsilon(t)\rangle^2 = \lim_\varepsilon \bar{\lambda}_i^{\varepsilon,t} = \lim_\varepsilon E\langle \bar{e}_i, W^\varepsilon(t)\rangle^2 = t\bar{\lambda}_i. \tag{5.8}$$

We are now prepared to verify (2.10). By (5.3) to (5.8), for each $\eta > 0$, there are $N_\eta < \infty$ and $\varepsilon_\eta > 0$ such that for $\varepsilon \leqq \varepsilon_\eta$,

$$\left| \sum_{1}^{N_\eta - 1} [E\langle \bar{e}_i^{\varepsilon,t}, W^\varepsilon(t)\rangle^2 - E\langle \bar{e}_i, W^\varepsilon(t)\rangle^2] \right| \leqq \eta,$$

$$\sum_{N_\eta}^{\infty} E\langle \bar{e}_i, W^\varepsilon(t)\rangle^2 \leqq \eta. \tag{5.9}$$

(5.9) and the arbitrariness of η implies (2.10).

To verify the second condition of Theorem 2.1 (e.g., in the form (2.9)), let $\tau \leqq T$ be a stopping time and evaluate

$$E\left| \sqrt{\varepsilon} \sum_{\tau/\varepsilon}^{(\tau+\delta)/\varepsilon} \psi_j Z_j \right|^2 \leqq 2\varepsilon E \sum_{j=\tau/\varepsilon}^{(\tau+\delta)/\varepsilon} \sum_{k \geqq j}^{(\tau+\delta)/\varepsilon} E_\tau^\varepsilon \psi_j \psi_k \langle Z_j, Z_k\rangle.$$

By Assumption 5.1, the right-hand side goes to zero as $\delta \to 0$ and $\varepsilon \to 0$, uniformly in $\tau \leqq T$ due to either: (a) the boundedness and mixing of $\{\psi_n\}$ and boundedness of $z(\cdot)$, or (b) $E|z(t)|^4 < \infty$ and boundedness and mixing of $\{\psi_n\}$. Thus $\{W^\varepsilon(\cdot)\}$ is tight in $D_H[0, \infty)$.

Weak convergence. Fix a weakly convergent subsequence, indexed also by ε, and with limit $W(\cdot)$.

Since $\{W^\varepsilon(\cdot)\}$ is tight and converges weakly in $D_H[0, \infty)$, to complete the proof, we need only show that for each m,

$$\{\langle W^\varepsilon(\cdot), \bar{e}_i\rangle, i \leqq m\} \Rightarrow \{W_i(\cdot), i \leqq m\},$$

an R^m-valued Wiener process with mean zero, mutually independent components and covariance

$$EW_i^2(t) = t\langle \bar{e}_i, \bar{R}\bar{e}_i\rangle.$$

This is a problem of weak convergence in a finite-dimensional Euclidean space and, as in Theorem 5.1, the perturbed test function method [9], [20], [21] yields this result under our hypotheses, and we omit the details. Q.E.D.

Weak convergence of $\{\hat{W}_k^\varepsilon(\cdot)\}$. To obtain the result in § 6, we need to know that the sequence determined by the centering of the interpolated process

$$\sqrt{\varepsilon} \sum_{1}^{t/\varepsilon} Z_j \langle Z_j, \bar{K}(\varepsilon j)\rangle$$

converges to a Wiener process. We now proceed to set this up.

Fix $k \in H$. Define

$$\xi_n^k(s) = \int_0^1 R(s-u)k(u)\,du - z(n\Delta - s)\int_0^1 z(n\Delta - u)k(u)\,du,$$

$$\xi_n^k = Rk - Z_n\langle Z_n, k\rangle.$$

For each ε, t, define the H-valued random variable $\hat{W}_k^\varepsilon(t)$ with point values $\hat{W}_k^\varepsilon(t, s)$ (for $s \in [0, 1]$) by

$$\hat{W}_k^\varepsilon(t, s) = \sqrt{\varepsilon} \sum_{0}^{t/\varepsilon} \xi_n^k(s).$$

The H-valued process $\hat{W}_n^\varepsilon(\cdot)$ is constant on each $[n\varepsilon, n\varepsilon + \varepsilon)$.

We will use the following assumption.

Assumption 5.2. The sum

$$\sum_{n=-\infty}^{\infty} E\xi_n^k(s)\xi_0^k(u) = \hat{R}_k(s, u)$$

converges absolutely and uniformly in (s, u) in bounded intervals, and each summand is continuous. The convergence is uniform in k in any strongly compact set.

Remark. The absolute and uniform convergence follows from the conditions of Assumption 5.1 and so does the continuity, if $R(\cdot)$ is continuous and $z(\cdot)$ Gaussian.

THEOREM 5.3. *Under Assumptions* 5.1 *and* 5.2, *and the conditions of* § 1, $\hat{W}_k^\varepsilon(\cdot) \Rightarrow \hat{W}_k(\cdot)$ *in* $D_H[0, \infty)$, *where* $\hat{W}_k(\cdot)$ *is a zero mean Wiener process with covariance operator* $t\hat{R}_k$ *defined by*

$$\hat{R}_k f(s) = \int_0^1 \hat{R}_k(s, u) f(u)\, du.$$

If $\{W^\varepsilon(\cdot), \hat{W}_k^\varepsilon(\cdot)\} \Rightarrow (W(\cdot), \hat{W}_k(\cdot))$ *in* $D_H^2[0, \infty)$, *then* $W(\cdot)$ *is independent of* $\hat{W}_k(\cdot)$, ($W^\varepsilon(\cdot)$ *is defined in Theorem* 5.2).

COROLLARY. *Let* $F(\cdot)$ *be an H-valued strongly continuous function defined on* $[0, \infty)$. *Define*

$$\rho_n^\varepsilon(s) = \int_0^1 R(s-u)F(\varepsilon n, u)\, du - z(n\Delta - s) \int_0^1 z(n\Delta - u)F(\varepsilon n, u)\, du$$

and

$$\hat{W}^\varepsilon(t) = \sqrt{\varepsilon} \sum_0^{t/\varepsilon} \rho_n^\varepsilon.$$

Then $\hat{W}^\varepsilon(\cdot) \Rightarrow \hat{W}(\cdot)$ *in* $D_H[0, \infty)$, *where* $\hat{W}(\cdot)$ *is a (nonstationary) zero mean Wiener process with*

$$E\langle \hat{W}(t), f\rangle\langle \hat{W}(t), g\rangle = \int_0^t dv \int_0^1 \int_0^1 f(u)g(s)\hat{R}_{F(v)}(s, u)\, ds\, du.$$

The last sentence of the theorem holds if $\hat{W}^\varepsilon(\cdot)$ *and* $\hat{W}(\cdot)$ *replace* $\hat{W}_k^\varepsilon(\cdot)$ *and* $\hat{W}_k(\cdot)$, *resp.*

Comments on the proof. The proof closely follows that of Theorem 5.2. For each ε, t, we can view $\hat{W}_k^\varepsilon(\cdot)$ as a nonstationary process on $[0, 1]$. As $\varepsilon \to 0$, the covariance $\hat{R}_k^{\varepsilon,t}(\cdot, \cdot)$ of this process converges uniformly and smoothly to $t\hat{R}_k(\cdot, \cdot)$. The eigenfunction expansions used in Theorem 5.2 are also used here in the same way—except that $\hat{R}_k^{\varepsilon,t}(\cdot, \cdot)$ and $\hat{R}_k(\cdot, \cdot)$ are nonstationary. The eigenfunction and eigenvalue approximations also carry over; approximate the eigenfunctions and values of $\hat{R}_k^{\varepsilon,t}(\cdot, \cdot)$ by those of $t\hat{R}_k(\cdot, \cdot)$. Once this method is used to prove tightness of $\{\hat{W}_k^{\varepsilon,t}(\cdot)\}$ in $D_H[0, \infty)$, the limit $\hat{W}_k(\cdot)$ is identified as it would be in Theorem 5.2: we simply work with $\{\langle \hat{e}_{k,i}, \hat{W}_k(\cdot)\rangle, i \leqq m\}$, where $\{\hat{e}_{k,i}\}$ are the eigenfunctions of $\hat{R}_k$.

If $(\hat{W}_k^\varepsilon(\cdot), W^\varepsilon(\cdot)) \Rightarrow (\hat{W}_k(\cdot), W(\cdot))$, then $\hat{W}_k(\cdot)$ and $W(\cdot)$ are mutually independent, since they are Gaussian and

$$E\langle \hat{W}_k(t), f\rangle\langle W(t), g\rangle \equiv 0 \quad \text{for all } f, g \in H,$$

since $\{\psi_n\}$ *is independent of* $z(\cdot)$, and $\{\psi_n\}$ does not appear in $\{\hat{W}_k^\varepsilon(\cdot)\}$. The details are omitted.

The corollary is proved in the same way.

6. Convergence of a normalized error process. Define $U_n^\varepsilon = (K_n^\varepsilon - \bar{K}(\varepsilon n))/\sqrt{\varepsilon}$, where K_n^ε is defined by (1.1) and $\bar{K}(\cdot)$ is the limit in Theorem 3.1. Let $U^\varepsilon(\cdot)$ denote the H-valued process with paths in $D_H[0,\infty)$ and values $U^\varepsilon(t) = U_n^\varepsilon$ on $[n\varepsilon, n\varepsilon+\varepsilon)$. Under the conditions of Theorem 3.1, we have $\dot{\bar{K}} = -R\delta\bar{K}$ and for each $T<\infty$

$$\bar{K}(\varepsilon n+\varepsilon) = \bar{K}(\varepsilon n) - \varepsilon R\delta\bar{K}(\varepsilon n) + \alpha_n^\varepsilon, |\alpha_n^\varepsilon| = O(\varepsilon^2), \qquad \varepsilon n \leqq T.$$

Then, using

$$K_{n+1}^\varepsilon = K_n^\varepsilon - \varepsilon Z_n\langle Z_n, \delta K_n^\varepsilon\rangle + \varepsilon Z_n\psi_n,$$

we have

$$U_{n+1}^\varepsilon = U_n^\varepsilon + \sqrt{\varepsilon}\, Z_n\psi_n + \sqrt{\varepsilon}\,\xi_n^\varepsilon - \varepsilon Z_n\langle Z_n, U_n^\varepsilon\rangle + \beta_n^\varepsilon, |\beta_n^\varepsilon| = O(\varepsilon^{3/2}), \quad U_0 = 0, \quad \varepsilon n \leqq T, \tag{6.1}$$

where

$$\xi_n^\varepsilon(s) = R(\bar{K}(\varepsilon n) - K) - Z_n\langle Z_n, \bar{K}(\varepsilon n) - K\rangle. \tag{6.2}$$

THEOREM 6.1. *Under the conditions of* § 1 *and Assumptions* 5.1 *and* 5.2, $U^\varepsilon(\cdot)$ *converges weakly in* $D_H[0,\infty)$ *to* $\bar{U}(\cdot)$ *satisfying*

$$\bar{U}(t) = -\int_0^t dv\, R\bar{U}(v) + W(t) + \hat{W}(t), \tag{6.3}$$

where $W(\cdot)$ *and* $\hat{W}(\cdot)$ *are the (mutually independent) Wiener processes defined in Theorems* 5.2 *and the corollary to Theorem* 5.3, *where* $F = \bar{K} - K$ *is to be used.*

Comments on the proof. Almost all the details have been worked out in Theorems 3.1, 5.2 and 5.3. As in Theorem 3.1, it is convenient to start with the "truncated" algorithm (6.4):

$$U_{n+1}^{\varepsilon,N} = U_n^{\varepsilon,N} + \sqrt{\varepsilon} Z_n\psi_n + \sqrt{\varepsilon}\xi_n - \varepsilon Z_n\langle Z_n, U_n^{\varepsilon,N}\rangle q_N(U_n^{\varepsilon,N}), \qquad U_0^\varepsilon = 0. \tag{6.4}$$

By the method of Theorem 3.1, the sequence of processes $\{Y^\varepsilon(\cdot)\}$

$$\varepsilon \sum_0^{t/\varepsilon} Z_n\langle Z_n, U_n^{\varepsilon,N}\rangle q_N(U_n^{\varepsilon,N}) \equiv Y^\varepsilon(t) \tag{6.5}$$

is tight in $D_H[0,\infty)$, and all limits are (strongly) continuous. By Theorems 5.2 and 5.3 (and its Corollary), the $W^\varepsilon(\cdot)$ and $\hat{W}^\varepsilon(\cdot)$ (replace the $F(\cdot)$ in the Corollary by $\bar{K}(\cdot) - K$) converge in $D_H[0,\infty)$ to H-valued Wiener (hence strongly continuous) processes. Thus $\{U^{\varepsilon,N}(\cdot)\}$ (the piecewise constant interpolation of $\{U_n^{\varepsilon,N}\}$) is tight in $D_H[0,\infty)$. Fix a convergent subsequence, indexed by ε, and with limit $\bar{U}^N(\cdot)$. The limit is (strongly) continuous. By the method in Theorem 3.1, the sequence (6.5) converges to

$$\int_0^t dv\, R\bar{U}^N(v) q(\bar{U}^N(v)) \text{ in } D_H[0,\infty).$$

Thus

$$\bar{U}^N(t) = -\int_0^t dv\, R\bar{U}^N(v) q(\bar{U}^N(v)) + W(t) + \hat{W}(t), \tag{6.6}$$

which has a unique solution and equals $\bar{U}(t)$ up until the time of first escape of $\bar{U}^N(\cdot)$ from the N-sphere in H. Since for each $T>0$, $P\{\bar{U}^N(t) = \bar{U}(t), t \leqq T\} \to 1$ as $N \to \infty$, the theorem is proved. Q.E.D.

REFERENCES

[1] J.-P. BERTRAN, *Optimisation stochastique dans un espace de Hilbert: Méthode de gradient*, C.R. Acad. Sci. Paris, Ser. A, 276 (19 Feb., 1973), pp. 613-616.

[2] ———, *Optimisation stochastique dans un espace de Hilbert et analyse fonctionnelle*, Thesis, Univ. of Nancy, 1975, Doctor des Sciences Mathématiques.

[3] G. I. SALOV, *On a stochastic approximation theorem in a Hilbert space and its applications*, Theory Prob. Appl., 24 (1979), pp. 413-419.

[4] P. RÉVÉSZ, *Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes* I, Studia Scientiarum Mathematicarum Hungarica, 8 (1973), pp. 391-398.

[5] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22 (1977), p. 551-575.

[6] B. WIDROW, J. M. MCCOOL, M. G. LARIMORE AND C. R. JOHNSON, *Stationary and nonstationary learning characteristics of the* LMS *adaptive filter*, Proc. IEEE, 64 (1976), pp. 1151-1162.

[7] T. G. KURTZ, *Semigroups of conditioned shifts and approximations of Markov processes*, Ann. Prob., 4 (1975), pp. 618-642.

[8] H. J. KUSHNER, *Jump-diffusion approximations for ordinary differential equations with wideband random right hand sides*, this Journal, 17 (1979), pp. 729-744.

[9] ———, *A martingale method for the convergence of a sequence of processes to a jump-diffusion process*, Z. Wahr., 53 (1980), pp. 207-219.

[10] W. F. FLEMING AND M. VIOT, *Some measure-valued Markov processes in population genetics theory*, Indiana Univ. Math. J., 28 (1979), pp. 817-844.

[11] P. KOTELENEZ, *Law of large numbers and control limit theorem for chemical reactions with diffusion*, Rept. 81, Dec. '82, Forschungsschwerpunkt Dynamische Systeme, Universität Bremen.

[12] T. G. KURTZ, *Approximation of Population Processes*, CBMS Regional Conference Series in Applied Mathematics 36, Society for Industrial and Applied Mathematics, Philadelphia, 1981.

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

[14] A. V. SKOROKHOD, *Limit theorems for stochastic processes*, Theory Prob. Appl., 1 (1956), pp. 262-290.

[15] H. CRAMÉR AND M. R. LEADBETTER, *Stationary and Related Stochastic Processes*, John Wiley, New York, 1967.

[16] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences, Springer, Berlin, 1978.

[17] A. V. BALAKRISHNAN, *Applied Functional Analysis*, 2nd edition, Springer, Berlin, 1981.

[18] E. WONG, *Stochastic Processes in Information and Dynamical Systems*, John Wiley, New York, 1971.

[19] G. C. PAPANICOLAOU AND W. KOHLER, *Asymptotic theory of mixing ordinary differential equations*, Comm. Pure Appl. Math., 27 (1974), pp. 641-668.

[20] H. J. KUSHNER AND HAI HUANG, *On the weak convergence of a sequence of general stochastic difference equations to a diffusion*, SIAM J. Appl. Math., 40 (1981), pp. 528-541.

[21] H. J. KUSHNER, *Approximation and Weak Convergence Methods for Random Processes, with Applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

008

ON SINGULAR EXTREMALS IN THE TIME MINIMAL CONTROL PROBLEM IN $\mathbb{R}^3$*

BERNARD BONNARD†

Abstract. In the time minimal control problem of a single input system in $\mathbb{R}^3$: $dv/dt = X(v) + uY(v)$, where X, Y are analytic, the singular controls are defined by a feedback of the form $u(v) = \Delta'(v)/\Delta(v)$. The purpose of this article is to classify the local behaviors of singular trajectories near the points v such that $\Delta(v) = 0$ for a generic system with X, Y respectively quadratic and constant.

Key words. time optimal control, singular extremals, polynomial systems

Introduction. In the investigation of the time minimal control problem for single input analytic systems in $\mathbb{R}^3$, a logical but difficult continuation of the work in [9], we have to take into account the existence of singular trajectories. These trajectories are solutions of a differential equation, analytic outside an analytic set. The behavior of the trajectories near this set is crucial in the analysis of the time minimal control problem and has to be studied. In this article this behavior is classified for a specific class of systems which contain the important Euler equation of the rigid body control problem.

The analysis of the behaviors of singular trajectories is only a step in the time optimal synthesis. Two other problems (at least), not discussed in this article, have to be considered. The first one is the optimality problem of singular trajectories. This problem, extensively studied in the literature, can be divided into two parts: necessary optimality conditions, see for instance [4], [5], and sufficient conditions [7], [8], [10]. The second problem is the classification of the behaviors of nonsingular trajectories near the switching surface [6]. All these additional problems shall be discussed in a forthcoming article dealing with the time minimal synthesis in the Euler equation.

1. Singular extremals. In this section we give some definitions and basic results on singular extremals in the time minimal control problem for single input systems.

DEFINITION 1.1. A singular extremal is an absolutely continuous curve $(x(t), p(t))$ on $\mathbb{R}^n \times \mathbb{R}^n - \{0\}$ which satisfies for almost all $t \geqq 0$ the equations

$$\frac{dx(t)}{dt} = X(x(t)) + u(t)\,Y(x(t)), \tag{1}$$

$$\frac{dp(t)}{dt} = -\left({}^t\frac{\partial X}{\partial x}(x(t)) + u(t)\,{}^t\frac{\partial Y}{\partial x}(x(t))\right)p(t), \tag{2}$$

$$(p(t),\, Y(x(t))) = 0, \tag{3}$$

where X, Y are analytic, $x(t)$ is the singular trajectory, $p(t)$ the adjoint vector and $u(t)$ a singular control.

DEFINITION 1.2. Let $(x(t), p(t))$ a singular extremal and set $\lambda(t) = (p(t), Y(x(t)))$. The order of the extremal is the first integer k such that $d^k\lambda(t)/dt^k$ can be written as $a(x(t), p(t)) + u(t)b(x(t), p(t))$ with $t \mapsto b(x(t), p(t))$ not identically zero.

* Received by the editors March 16, 1984, and in revised form August 1, 1984. This work was supported in part by the Office of Naval Research under JSEP contract N00014-75-C-0648.

† Division of Applied Sciences, Harvard University, Cambridge, Massachusetts 02138.

Present address, ENSIEG, Laboratoire d'Automatique de Grenoble, Domaine Universitaire, B.P. 46, 38402 Saint-Martin-d'Hères, France.

DEFINITION 1.3. Let $Z: \mathbb{R}^n \mapsto \mathbb{R}^n$, $f: \mathbb{R}^n \mapsto \mathbb{R}$ be analytic maps. By solution of the equation $dy/dt = Z(y)/f(y)$ we mean an absolutely continuous curve $y(t)$, a solution almost everywhere of the above equation, such that $f(y(0)) \neq 0$ and with no arc $y(t)$, $t \in [t_1, t_2]$, $t_2 > t_1$, contained in the set $f = 0$.

DEFINITION 1.4. Let $Z_1, Z_2: \mathbb{R}^n \mapsto \mathbb{R}^n$ analytic. The Lie bracket $[Z_1, Z_2]$ is the map defined by

$$[Z_1, Z_2](x) = \frac{\partial Z_2}{\partial x}(x) Z_1(x) - \frac{\partial Z_1}{\partial x}(x) Z_2(x).$$

An easy computation gives the following result.

LEMMA 1.5. *Let $(x(t), p(t))$ be a singular extremal. Then for almost all $t \geqq 0$ the equations below are satisfied*:

$$\lambda(t) = (p(t),\ Y(x(t))) = 0,$$

$$\frac{d\lambda}{dt}(t) = (p(t), [X,\ Y](x(t))) = 0,$$

$$\frac{d^2\lambda}{dt^2}(t) = (p(t), [X, [X,\ Y]](x(t)) + u(t)[Y, [X,\ Y]](x(t))) = 0.$$

PROPOSITION 1.6. *Let $(x(t), p(t))$ be a singular extremal of order 2. Then a singular control is defined by $u(t) = u(x(t), p(t))$ with $u(x, p)$ given by*

$$u(x, p) = \frac{(p, [X, [Y, X]](x))}{(p, [Y, [X,\ Y]](x))}. \tag{4}$$

The singular extremals of order 2 are the solutions of the system (1), (2), *where u is given by* (4) *whose initial conditions satisfy the relation*

$$(p(0),\ Y(x(0))) = (p(0), [X,\ Y](x(0))) = 0. \tag{5}$$

Proof. Apply Lemma 1.5.

COROLLARY 1.7. *Consider the case of a system* (1) *in $\mathbb{R}^3$. Let us denote respectively $\Delta(v)$ and $\Delta'(v)$ the determinant of the vectors $\{Y(v), [X,\ Y](v), [Y, [X,\ Y]](v)\}$ and $\{Y(v), [X,\ Y](v), [X, [Y,\ X]](v)\}$. Then $(v(t), p(t))$ is a singular extremal of order 2 if and only if $v(t)$, $t \geqq 0$, is a solution of the equation in $\mathbb{R}^3$*

$$\frac{dv}{dt}(t) = X(v(t)) + \frac{\Delta'(v(t))}{\Delta(v(t))} Y(v(t)) \tag{6}$$

and $p(t)$ is a solution of (2) *satisfying* (5).

Proof. Since $p(t) \neq 0\ \forall t$, the relation $\lambda(t) = d\lambda(t)/dt = d^2\lambda(t)/dt^2 = 0$ implies that $-\Delta'(v) + u\Delta(v) = 0$. Therefore singular controls of order 2 can be expressed as the feedback $u(v) = \Delta'(v)/\Delta(v)$.

Remark 1.8. For generic systems (1) there are then as many singular extremals of order 2 as solutions of an analytic differential equation with initial conditions in a subset of codimension 2. Therefore, these extremals have to be considered in each time minimal control problem. On the other hand, for reasons not discussed here one may conjecture that for generic systems (1) there is no singular extremal of order >2, the study of singular extremals being therefore equivalent to the analysis of a unique differential equation given by Proposition 1.6. This equation is analytic outside an analytic set. A priori this set may cause a junction between extremals, complicated behaviors or a blowing-up phenomenon. Therefore the behaviors of the solutions near

this set has to be studied. Such equations also occur in the disturbance decoupling problem and beyond the restrictive study of this article a theory of these equations has to be done.

2. The local behaviour of singular trajectories for quadratic systems in $\mathbb{R}^3$. Now we will only consider systems (1) in $\mathbb{R}^3$ where X is a map whose components are quadratic forms and Y is a constant vector. We will classify in this section the local behaviors of singular trajectories of order 2, solutions of the equation (6), near the set $\Delta = 0$, for an open dense set S of systems (in the topology induced by the coefficients) and conclude by noticing that there exist no singular extremals of order >2.

Notation 2.1. We note e_1, e_2, e_3 the canonical basis of $\mathbb{R}^3$. To save indices note (x, y, z) the coordinates of a vector $v \in \mathbb{R}^3$, (X_1, X_2, X_3) the ones of X with $X_1(v) = a_1x^2 + a_2y^2 + a_3z^2 + a_4xy + a_5xz + a_6yz$, X_2 and X_3 being respectively defined by turning in X_1a_i into b_i and c_i. Observe that the map $[X, Y][v]$ is linear and set $[X, Y](v) = -Av$. Denote ad A, the classical coadjoint of A. If A^{-1} exists we have $(\det A)A^{-1} = \operatorname{ad} A$. Set $w = (\operatorname{ad} A)Y$ and notice that $[X, Y](w) = -(\det A)Y$. Define the two lines L_1, L_2 by $L_1 = \rho Y$, $L_2 = \rho w$.

We will now compute (6) in an adapted basis and define S.

PROPOSITION 2.2. *If Y and w are linearly independent then $\Delta = 0$ is the plane generated by Y and w; $\Delta = 0$ is equal to $\mathbb{R}^3$ if and only if Y and w are collinear.*

Assume $Y = e_1$. Then we have $w = e_3$ if and only if $b_5 = c_5 = 0$ and $2(b_1c_4 - b_4c_1) = 1$. In this case we have

(a) $\Delta(v) = -y$;

(b) *the restriction of Δ' to $\Delta = 0$ is the polynomial*

$$P(x, z) = b_1x^3 - 2dx^2z - b_3xz^2;$$

with $d = 2[b_1(b_6c_1 - b_1c_6) + 2c_1(b_3c_1 - b_1c_3)]$;

(c) $\Delta'(v) = P(x, z) + a_3yz^2 +$ *other monomials having degree* $\geqq 2$ *in* y.

Proof. Let $Y \neq 0$; then we can assume that $Y = e_1$. Computations give

$$[X, e_1](v) = -((2a_1x + a_4y + a_5z), (2b_1x + b_4y + b_5z), (2c_1x + c_4y + c_5z));$$

$$w = ((b_4c_5 - b_5c_4), 2(b_5c_1 - b_1c_5), 2(b_1c_4 - b_4c_1));$$

$$\Delta(v) = 2y(b_4c_1 - b_1c_4) + 2z(b_5c_1 - b_1c_5).$$

Therefore $\Delta(e_1) = \Delta(w) = 0$; Δ is zero if and only if w is collinear to e_1. Clearly $[X, e_1](e_3)$ is collinear to e_1 if and only if $b_5 = c_5 = 0$ and $w = e_3$ implies that $2(b_1c_4 - b_4c_1) = 1$. The computation of $\Delta'(v)$ is straightforward with $Y = e_1$ and $w = e_3$ and yields (b) and (c).

Notation 2.3. Assume that Y and w are linearly independent, i.e., L_1 and L_2 distinct. Then the solutions of $\Delta' = \Delta = 0$ are the line L_2 and eventually two lines noted L_3, L_4. If we assume that $Y = e_1$ and $w = e_3$ then L_3, L_4 are given by ρv, $v = e_1 + ze_3$, z being real solutions of the second order equation $P(1, z) = 0$.

DEFINITION 2.4. We define S as the set of pairs (X, Y) such that the lines L_1, L_2, L_3, L_4 (if the two last exist) are distinct. Notice that by Proposition 2.2, S is an open dense set.

DEFINITION 2.5. Let $x \in \mathbb{R}^n$, $x \neq 0$. A conic neighborhood of the line ρx is the set of points $y \neq 0$ such that one of the two points $\pm y/|y|$ are in a given polydisk (on the sphere) of center $x/|x|$.

THEOREM 2.6. *Consider a pair (X, Y) in S. The only points where singular trajectories can cross the plane $\Delta = 0$ at finite distance are on the lines L_2, L_3, L_4 (if L_3, L_4 exist). This is done transversally, with continuous controls and in the following manner.*

(a) *Let $v \neq 0$ on the line L_2. Then there exists near v an analytic manifold V_v of dimension 2, tranverse to L_2, and invariant for the singular trajectories. If $X(v)$ points towards $\Delta > 0$ ($\Delta < 0$ resp.) each singular trajectory in V_v and $\Delta < 0$ ($\Delta > 0$ resp.) cross the plane $\Delta = 0$ at v, the singular controls and the adjoint vectors being all distinct.*

(b) *Let $v \neq 0$ on the line L_3 or L_4. Then there there exists exactly one singular trajectory crossing $\Delta = 0$ at v.*

We have the following blowing-up phenomenon. If $X(Y)$ points towards $\Delta > 0$ ($\Delta < 0$ resp.) there exists a conic neighborhood U of L_1 such that each singular trajectory with initial condition in U and $\Delta < 0$ ($\Delta > 0$ resp.) hits in finite time and with an infinite control the plane $\Delta = 0$ at infinite distance and on a line parallel to L_1.

There exists no singular extremal of order > 2.

The last part of this section is devoted to explaining and proving the above theorem.

A trick to cope with the peculiarity of (6) is to introduce the new time variable

$$d\tau = \frac{dt}{\Delta(v(t))}. \tag{7}$$

By setting $v(\tau) = v(t)$, (6) is then turned into the ordinary differential equation

$$dv(\tau)/d\tau = \Delta(v(\tau))X(v(\tau)) + \Delta'(v(\tau))Y. \tag{8}$$

Reversing the orientation of the solutions of (8) if $\Delta < 0$, we obtain the behavior of the singular trajectories.

Before investigating the properties of the solutions of (8) near $\Delta = 0$, which are indeed very special, we have to recall a few results. First we have the following proposition.

PROPOSITION 2.7 (see [2]). *Consider a real analytic system in $\mathbb{R}^n$. $dx/db = Mx + o(|x|)$ where $M \sim \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$, $0 > \lambda_1 \geqq \cdots \geqq \lambda_n$. Then there exists a local real analytic change of coordinates $y = x + o(|x|)$ which reduces the system to the form $dy_i/db = \lambda_i y_i + g_i(y_1, \cdots, y_{i-1})$, where g_i is a polynomial containing solely terms in $y_1^{m_1}, \cdots, y_{i-1}^{m_{i-1}}$ such that there exists a resonant relation $\lambda_i = m_1\lambda_1 + \cdots + m_{i-1}\lambda_{i-1}$.*

This is just a weak version of Dulac's theorem, but sufficient for our purpose. Notice that by integrating the second system in cascade we obtain $y_i(b) = e^{\lambda_i b}K_i(b)$ where K_i is a polynomial. The solution $x(b)$, $b \geqq 0$, $x(0)$ small enough, can then be computed via the inversion of the change of coordinates. If $\lambda_i > 0\ \forall i$, the result is still valid by turning b into $-b$. And more generally we can evaluate for each analytic system the solution near a nondegenerate singular point in the stable and unstable manifold.

Below we give a few definitions and basic results about homogeneous systems necessary in the study of (8).

Homogeneous systems 2.8. Consider the system in $\mathbb{R}^n$

$$\frac{dx}{d\tau} = H(x) \tag{9}$$

where all the components of H are homogeneous polynomials of the same degree k.

Note $x(\tau, x_0)$ the solution such that $x(0, x_0) = x_0$. Then $x(\tau, \rho x_0) = \rho x(\rho^{k-1}\tau, x_0)\ \forall \rho \in \mathbb{R}$. Then $y(b) = x(\tau)/|x(\tau)|$, with $db = |x(\tau)|^{k-1}\, d\tau$, are the solutions of a differential equation on the sphere called the projected system.

The singular points of this equation are the points $y/|y|$, $y \in \mathbb{R}^n$, $y \neq 0$, such that $H(y)$ is colinear to y. If $H(y) = 0$ the line ρy is a set of singular points of (9), otherwise

the line is supporting a solution of (9) which is a ray to or from the origin if $(H(y), y) < 0$ or $(H(y), y) > 0$, respectively.

Let $y \neq 0$, such that $H(y)$ is colinear to y. Note $\{e_1, \cdots, e_n\}$ the canonical basis of $\mathbb{R}^n$. We may suppose that $y = e_n$. The system (9) can be written as

$$\begin{aligned} \frac{d\bar{x}}{d\tau} &= x_n^{k-1}\bar{M}\bar{x} + \bar{R}(x), \\ \frac{dx_n}{d\tau} &= \lambda_n x_n^k + R_n(x), \end{aligned} \tag{10}$$

where $\bar{x}$ is the (column) vector of $\mathbb{R}^{n-1}(x_1, \cdots, x_{n-1})$, $\bar{M}$ is a $(n-1)\times(n-1)$ matrix, $\bar{R}$ is a polynomic mapping whose components are of degree $\leqq k-2$ in x_n. R_n is a polynomial having degree $\leqq k-1$ in x_n, $\lambda_n = 0$ if and only if v is a singular point.

Assume $\bar{M} \sim \operatorname{diag}(\lambda_1, \cdots, \lambda_{n-1})$ with $\lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_p < 0 < \lambda_{p+1} \leqq \cdots \leqq \lambda_{n-1}$. One may suppose that in (10) $\bar{M}$ is diagonal. By setting

$$u_i = \frac{x_i}{x_n}, \quad i \neq n, \qquad db = x_n^{k-1}\, d\tau, \tag{11}$$

the system (10) is transformed into

$$\frac{du_i}{db} = (\lambda_i - \lambda_n)u_i + K_i(u), \tag{12}$$

$$\frac{dx_n}{db} = x_n(\lambda_n + K_n(u)), \tag{13}$$

where $K_i(u) = o(|u|)$ for $i \neq n$ and $K_n(0) = 0$. Clearly (12) is the projected system expressed in the chart $u_i = x_i/x_n$.

The numbers λ_i have in fact an intrinsic significance, $\lambda_i - \lambda_n$, $i \neq n$, which shall be called the transverse eigenvalues and λ_n the eigenvalue of the line ρy.

Suppose that $\lambda_n = 0$, i.e., $y = e_n$ is a singular point. For $x_n > 0$, set $v = x_n y$ and

$$\begin{aligned} V_v^s &= \{x_0 \text{ s.t. } x(\tau, x_0) \mapsto v \text{ as } \tau \mapsto +\infty\}, \\ V_v^u &= \{x_0 \text{ s.t. } x(\tau, x_0) \mapsto v \text{ as } \tau \mapsto -\infty\}. \end{aligned}$$

By the analytic version of [3, p. 243, 6.2] the invariant sets V_v^s and V_v^u are near v analytic manifolds of dimension p and $n-p-1$ respectively called the stable and unstable manifold of v. Moreover the linear parts of the restriction of (10) to V_v^s, V_v^u are respectively defined by the matrices $x_n^{k-1}\operatorname{diag}(\lambda_1, \cdots, \lambda_p)$ and $x_n^{k-1}\operatorname{diag}(\lambda_{p+1}, \cdots, \lambda_{n-1})$. Since the system is homogeneous, we have $V_{\rho v}^s = \rho V_v^s$ and $V_{\rho v}^u = \rho V_v^u \; \forall \rho > 0$.

We can summarize in a few words all these results. The behavior of the solutions near a line which is a set of singular points or two rays solutions is described completely by a set of numbers λ_i. These numbers can be read in the system written as (10). Each nonzero singular point generates two families of homothetic integral manifolds. The behavior of the solutions near this point is entirely described by the projected system. Near a ray, the behavior of the solutions is also given by the projected system except the loss of information in the projection when the solutions are at infinite distance.

We can now carry through the analysis of the solutions of (8) near $\Delta = 0$.

Hypothesis 2.9. It is not a restriction to suppose that $Y = e_1$ and $w = e_3$. Δ and Δ' are given by Proposition 2.2, in particular $\Delta = 0$ is the plane $y = 0$.

LEMMA 2.10. *The equation* (8) *is a homogeneous polynomial equation of degree* 3. *The plane* $y=0$ *is an invariant in which the restriction of the system is given by the vector field* $P(x, z)e_1$. *The singular points of* (8) *in* $y=0$ *are the points on the lines* L_2, L_3, L_4 (*if the two last exist*). *In such a point* e_1 *is eigenvector of the linearized system. The line* $L_1=\rho e_1$ *is supporting a ray solution of* (8) *in* $y=0$.

These assertions are direct consequence of the definition of (8) and Proposition 2.2. From this lemma we can deduce at once that singular trajectories can cross $y=0$ at finite distance only at points on the line L_2, L_3, L_4. Since the line L_1 in $y=0$ is supporting a ray solution of (8), singular trajectories could eventually hit $y=0$ near L_1 at infinite distance. Let us study the behaviour of the solutions of (8) near these lines.

Behavior near the line L_2 2.11. The line L_2 is the set $x=y=0$. The system (8) can be written as $dv/d\tau = z^2Mv+R(v)$, where R is a polynomic mapping whose components are of degree $\leqq 1$ in z and M is computed using Proposition 2.2:

$$M=\begin{bmatrix} -b_3 & 0 & 0 \\ 0 & -b_3 & 0 \\ 0 & -c_3 & 0 \end{bmatrix}.$$

Hence $M\sim \operatorname{diag}(-b_3, -b_3, 0)$. Since $(X, Y)\in S$, $b_3\neq 0$, therefore X is transverse to $y=0$ on L_2. We can assume that $b_3>0$. Let $v\in L_2$ and set $V_v=$ stable manifold of v, then by Homogeneous systems 2.8 V_v is near v a two-dimensional analytic manifold, with tangent space at v given by the equation $-c_3y+b_3z=0$. All trajectories of (8) in V_v are hitting the plane $y=0$ transversally and with different slopes because if $v=(0, 0, z_0)$ the matrix defining the linear part at v of the restriction of (8) to V_v is $\sim -z_0^2 \operatorname{diag}(b_3, b_3)$.

Since the plane $y=0$ is hit transversally and $Y=e_1$ is contained in this plane, the singular controls are finite at v. In fact these controls can be evaluated using Dulac's theorem. We can suppose that $c_3=0$, i.e., $e_2=$ eigenvector of M. Express the system (8) and the singular controls Δ'/Δ in the variables introduced in 2.8 $u_1=x/z$, $u_2=y/z$, z and $db=z^2\,d\tau$. Then $\Delta'(x, y, z)=z^3Q(u_1, u_2)$ where Q is a polynomial of degree 3 without constant term, and $\Delta(x, y, z)=-zu_2$. Using Proposition 2.7 we can express for $u_i(0)$ small enough, $b\geqq 0$, $u_i(b)$ as series $\sum_{n\geqq 1} A_{ni}\,e^{-nb_3b}$, $A_{12}\neq 0$ if and only if $u_2(0)\neq 0$, i.e., $y(0)\neq 0$, each term of the series being uniformly convergent for $b\geqq 0$. Therefore $Q(u_1, u_2)$ can be written as $\sum_{n\geqq 1} B_n\,e^{-nb_3b}$ and $\Delta'/\Delta\mapsto -z_0^2B_1/A_{12}$ as $b\mapsto+\infty$.

The behavior near L_2 of the adjoint vector along a singular trajectory going towards L_2 can be evaluated in the following manner. Notice that we can express (2) in the u_i, b variables as an equation

$$\frac{dp}{db}=u_2B(u_i)p, \tag{14}$$

where $B(u_i)$ is a matrix whose coefficients are polynomials of degree one. Set $C(b)=u_2(b)B(u_i(b))$. The solution of (14) can be expressed using Chen's expansion theorem [1] as $p(b)=D(b)p(0)$, with $D(b)=I+\int_0^b C(s)\,ds+\cdots$. Using Proposition 2.7, $D(b)$ can be computed along solution $u_i(b)$, $b\geqq 0$, $u_i(0)$ small enough and in particular $D(+\infty)=\lim D(b)$ as $b\mapsto+\infty$ can be written as $I+u_2(0)E(u_i(0))$ where E is a power series. Set $p=(p_1, p_2, p_3)$. The initial conditions of $p(0)$ of the adjoint vector can be computed as a function of $u_i(0)$ using the relations $(p, Y)=(p, [X, Y])=0$. By Proposition 2.2 we get: $p_1(0)=0$,

$$p_2(0)=-p_3(0)\frac{(2c_1u_1(0)+c_4u_2(0))}{(2b_1u_1(0)+b_4u_2(0))} \qquad \text{if } 2b_1u_1(0)+b_4u_2(0)\neq 0,$$

$p_1(0) = p_3(0) = 0$ otherwise. Note that only initial conditions $u_i(0)$ on the same lines through 0 yield the same $p(0)$. As $[Y,[X, Y]] = -2(a_1, b_1, c_1)$ and $2(b_1c_4 - b_4c_1) = 1$, clearly $(p(0),[Y,[X, Y]]) \neq 0$ if and only if $u_2(0) \neq 0$, i.e., $y(0) \neq 0$. The value of the adjoint variable when the plane $y = 0$ is crossed is $p(+\infty) = D(+\infty)p(0)$. For $u_i(0)$ small enough $(p(+\infty), [Y,[X, Y]]) \sim (p(0),[Y,[X, Y]])$ which is $\neq 0$ if $y(0) \neq 0$.

This shows that the singular extremals remain in the set $(p,[Y,[X, Y]]) \neq 0$ when the plane $y = 0$ is crossed by the singular trajectories through L_2. Singular extremals being therefore solutions of an analytic differential equation, no junction between singular extremals is possible. In fact this result was more and less obvious because L_2 is the set where Y and $[X, Y]$ are colinear. Therefore the relations $(p, Y) = (p,[X, Y]) = 0$ and $(p,[Y,[X, Y]]) \neq 0$ are compatible on L_2. However, the labored proof above is done on purpose to show how to evaluate the adjoint vector. More generally in our problem all is computable.

Behavior near the linears L_3, L_4 2.12. By Proposition 2.2 and 2.3, the lines L_3, L_4 are ρv_ε with $v_\varepsilon = (1, 0, z_\varepsilon)$, $z_\varepsilon = -(d + \varepsilon\sqrt{d^2 + b_1b_3})/b_3$, $d^2 \geqq = b_1b_3$, $\varepsilon = \pm 1$. Let us denote $(\bar{x}, \bar{y}, \bar{z})$ the coordinates of $\bar{v} \in \mathbb{R}^3$ in the basis e_1, e_2, v_ε, i.e., $x = \bar{y} + \bar{z}$, $y = \bar{y}$ and $z = z_\varepsilon\bar{z}$. To see the behavior of the solutions of (8) near L_3, L_4 we have, by 2.8, to write the system as $d\bar{v}/d\tau = \bar{z}^2 M\bar{v} + R(\bar{v})$, where R is a polynomic mapping whose components are of degree $\leqq 1$ in $\bar{z}$. Clearly M has the form

$$\begin{bmatrix} \lambda_1 & * & 0 \\ 0 & \lambda_2 & 0 \\ 0 & * & 0 \end{bmatrix},$$

the eigenvalues of M being λ_1, λ_2, 0. A straightforward computation yields

$$\lambda_1 = 3b_1 - 4\, dz_\varepsilon - b_3 z_\varepsilon^2,$$

$$\lambda_2 = -b_1 - b_3 z_\varepsilon^2.$$

LEMMA. *For each pair* (X, Y) *in* S *we have* $\lambda_1\lambda_2 < 0$.

Proof. Set $u = b_1b_3$, $\alpha_1 = -\lambda_1 b_3$, $\alpha_2 = -\lambda_2 b_3$. Then sign $(\lambda_1\lambda_2) =$ sign $(\alpha_1\alpha_2)$ and we have $\alpha_2 = u + (d + \varepsilon\sqrt{d^2 + u})^2$, $\alpha_1 = -4\delta + \alpha_2$ with $\delta = u + d(d + \varepsilon\sqrt{d^2 + u})$. Notice that if we change d into $-d$ and ε into $-\varepsilon$ we get the same values, then we can assume $\varepsilon = +1$. Clearly $\alpha_2 = 0 \Leftrightarrow u = -d^2$ or $u = 0$ and $d < 0$. Moreover, $\alpha_1 = 0 \Leftrightarrow \alpha_2 = 0$. Fix d, the variation of α_i as a function of u can be studied. Note $\alpha_i' = d\alpha_i/du$, we have $\alpha_2' = (d + 2\sqrt{d^2 + u})/\sqrt{d^2 + u}$, $\alpha_1' = -\alpha_2'$ and $\alpha_2' = 0 \Leftrightarrow u = -3d^2/4$ and $d < 0$. Therefore we have two cases. If $d > 0$, $\alpha_2 > 0$, $\alpha_1 < 0$ for $u > -d^2$. If $d < 0$, $\alpha_2 < 0$, $\alpha_1 > 0$ for $u \in]-d^2, 0[$ and $\alpha_2 > 0$, $\alpha_1 < 0$ for $u > 0$. In conclusion if $b_1b_3 \neq -d^2, 0$, we have $\lambda_1\lambda_2 < 0$. Otherwise the system is not in S.

Since $\lambda_1\lambda_2 < 0$, by 2.8 there is for each $v \in L_3$, L_4 only one singular trajectory crossing $y = 0$ at v. This is done transversally; the singular control then remains finite. In fact it can be evaluated in the following manner. Use [3, 5.1, p. 235] to linearize by a nonlinear change of coordinates the invariant manifold of (8) transverse to $y = 0$ at v. Then use Proposition 2.7 to compute the solution of (8) in this manifold.

Behaviour near the line L_1 2.13. The line L_1 is the set $y = z = 0$. By Proposition 2.2 the system (8) can be written

$$\frac{dx}{d\tau} = b_1x^3 + R_1(v), \qquad \frac{d\bar{v}}{d\tau} = x^2\bar{M}\bar{v} + R_2(v)$$

where

$$\bar{v} = (y, z), \qquad \bar{M} = \begin{bmatrix} -b_1 & 0 \\ -c_1 & 0 \end{bmatrix},$$

R_1 is a polynomial having degree $\leqq 2$ in x and R_2 a polynomic mapping whose components are of degree $\leqq 1$ in x. Note that since $(X, Y) \in S$, $b_1 \neq 0$. One may assume that $b_1 > 0$. $\bar{M} \sim \operatorname{diag}(-b_1, 0)$ and the transverse eigenvalues are $-2b_1$, $-b_1$. Therefore the line L_1 produces a stable node for the projected system of (8).

The behavior near L_1 of the trajectories of (8) is the following. One may assume that $c_1 = 0$. The system written in the variables $u_1 = y/x$, $u_2 = z/x$, x^{-1} and $db = x^2\, d\tau$ is by 2.8, (12) and (13):

$$\frac{du_1}{db} = -2b_1 u_1 + o(|u|),$$

$$\frac{du_2}{db} = -b_1 u_2 + o(|u|),$$

$$\frac{dx^{-1}}{db} = -b_1 x^{-1} + o(|(u, x^{-1})|).$$

The point $u_1 = u_2 = x^{-1} = 0$ is a singular point and the matrix of the linearized system is $\operatorname{diag}(-2b_1, -b_1, -b_1)$. We can apply Proposition 2.7 to evaluate the solutions. There are three resonant relations and the solutions can be expressed as: $x^{-1}(b) = A\, e^{-b_1 b} + O(e^{-2b_1 b})$, $u_2(b) = B\, e^{-b_1 b} + O(e^{-2b_1 b})$, $u_1(b) = e^{-2b_1 b} P(b) + o(e^{-2b_1 b})$, where P is a polynomial. Back to the original coordinates we get, that as $b \mapsto +\infty$, $x(b) \mapsto \infty$, $y(b) \mapsto 0$ and $z(b) \mapsto B/A$. The set of initial conditions such that $x(b) \mapsto +\infty$ (or $-\infty$) and $z(b) \mapsto \rho$ fixed as $b \mapsto +\infty$ is an invariant analytic manifold of dimension 2 for $x^{-1}(0)$, $u_1(0)$, $u_2(0)$ small enough and we have a family of such manifolds when ρ varies which are homothetic (in the original coordinates).

Finally notice that clearly when $x(b) \mapsto \infty$ as $b \mapsto +\infty$, the real time variable $t(b)$ tends to a finite limit.

Remarks 2.14. A straightforward computation shows that to produce such a behavior in the neighborhood of the line ρY, X has to be only a homogeneous polynomial of degree $\geqq 2$.

An interesting question is to study the contact between the two families of invariant sets generated by the lines L_1 and L_2 respectively.

Proposition 2.15. *For each pair* (X, Y) *in* S *there exists no singular extremals of order* >2.

Proof. By Lemma 1.5 a singular trajectory of order >2 has to stay in $\Delta = \Delta' = 0$, i.e., on L_2, L_3, or L_4. This is not possible because at each point of $L_i X$ is transverse and Y tangent to the plane $y = 0$. Therefore a control cannot force the system to track one of the lines L_2, L_3, or L_4.

3. Conclusion. The remaining cases, i.e., pairs (X, Y) not in S, can be understood as collisions between the lines L_i when (X, Y) varies. In particular if we fix X, they are bifurcations between different types of global behaviors of singular trajectories when Y varies.

We have focused our work on a special class of systems connected with the rigid body control problem but the same techniques can be applied to classify the behaviors of singular trajectories near the singular set for generic control systems in $\mathbb{R}^3$.

This article is only a piece of a puzzle call the time minimal control problem in $\mathbb{R}^3$. Additional results concerning the optimality problem of singular extremals and the global behaviors of singular trajectories in Euler's equation shall appear before long.

Finally we must emphasize the fact that singular trajectories of the time minimal control problem are due to singularities of the input-output mapping and are then an important invariant of a control system. These trajectories are somehow connected with the controllability properties of the system and have to be taken into account in each control problem.

REFERENCES

[1] K. T. CHEN, *Expansion of solutions of differential systems*, Arch. Rat. Mech. Anal., 13 (1963), pp. 348–363.

[2] H. DULAC, *Solutions d'un système d'équations differentielles dans le voisinage de valeurs singulières*, Bull. Soc. Math. de France, 40 (1912), pp. 324–392.

[3] P. HARTMAN, *Ordinary Differential Equations*, Birkhauser, Boston, 1982.

[4] H. HERMES, *Local controllability and sufficient conditions in singular problems*, J. Differential Equation, 20 (1976), pp. 213–232.

[5] A. J. KRENER, *The high order maximal principle and its application to singular extremals*, this Journal, 15 (1977), pp. 256–293.

[6] I. KUPKA, *Generic properties of extremals in optimal control problems*, in Differential Geometric Control Theory, Progress in Mathematics Vol. 27, Birkhauser, Boston, 1983, pp. 310–315.

[7] H. G. MOYER, *Sufficient conditions for a strong minimum in singular control problems*, this Journal, 11 (1973), pp. 620–636.

[8] A. V. SARYCEV, *The index of the second variation of a control system*, Math. USSR Sbo., 41 (1982), 3, pp. 383–401.

[9] H. J. SUSSMANN, *Time optimal control in the plane*, in Proc. Conference on Linear and Nonlinear Systems, Bielefeld and Rome, 1981.

[10] ———, *Lie brackets and local controllability: a sufficient condition for scalar-input systems*, this Journal, 21 (1983), pp. 686–713.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 5, September 1985

009

ADDENDUM: STABLE AND REGULAR REACHABILITY OF RELAXED HEREDITARY DIFFERENTIAL SYSTEMS*

FRITZ COLONIUS†

Abstract. This paper characterizes regular reachability of relaxed hereditary differential systems as a positive controllability property of an associated linear system in the Sobolev space $W^{1,2}$. Thus the results of F. Colonius [SIAM J. Control Optim., 20 (1982), pp. 675-694] are improved. Regular reachability is relevant as a regularity condition in the proof of a maximum principle for fixed final state optimal control problems.

The paper [2] considered optimal control problems for relaxed hereditary differential systems of the form

$$x_{t_0} = \phi^0, \qquad \dot{x}(t) = f(x_t, v(t), t) \quad \text{a.a. } t \in T := [t_0, t_1]. \tag{1}$$

For problems with a fixed final state $x_{t_1} = \phi \in W^{1,2}([-r, 0], \mathbb{R}^n)$, a maximum principle was proved, provided that the optimal solution (x^0, v^0) satisfies the following regularity condition for some $\delta > 0$:

$$\dot{\phi}(t - t_1) \in \operatorname{int}_\delta \operatorname{co} f(x_t^0, \Omega(t), t) \quad \text{a.a. } t \in T_1 := [t_1 - r, t_1] \tag{2}$$

(the abbreviations T and T_1 introduced above will be used throughout this note).

The paper [3] was devoted to the problem of understanding the regularity condition (2). However, no complete characterization in terms of controllability properties of a linearized system could be given. The present note solves this problem. This gives deeper insight into the relations between optimal control and structure theory of hereditary differential systems and provides the missing link between the special situation of [2] and the general Banach space setting of [4].

In the following, we assume that ϕ is continuously differentiable and that f and the set Ω do not depend on time t. This is in order not to overburden this note with technical details.

Furthermore, we suppose that the assumptions (1.1), (1.2), and (1.4) of [3] are satisfied, and define for an interval $I \subset T$ and $p = 2$ or $p = \infty$

$$U_p(I) := \{u \in L_p(I; \mathbb{R}^n) : u(t) \in P(t) := \mathbb{R}_+(\operatorname{co} f(x_t^0, \Omega) - f(x_t^0, v^0(t)) \text{ for a.a. } t \in I\};$$

here $\mathbb{R}_+$ is the set of all nonnegative reals.

Along with (1) we consider the following *linearized system* ($p = 2$ or $p = \infty$):

$$x_{t_0} = 0, \qquad \dot{x}(t) = \mathscr{D}_1 f(x_t^0, v^0(t)) x_t + u(t) \quad \text{a.a. } t \in T, \tag{3}$$

$$u \in \mathscr{U}_p(T). \tag{4}$$

The present analysis differs in two essential points from the previous one in [3]. (a) We consider the final states x_{t_1} of the linearized system in $W^{1,2}([-r, 0], \mathbb{R}^n)$, instead of $W^{1,\infty}([-r, 0], \mathbb{R}^n)$. (b) In [3, Lemma 1.5], the control values $u(t)$ of the linearized system were required to lie in $\operatorname{co} f(x_t^0, \Omega) - f(x_t^0, v^0(t))$, while (4) above allows the control values to lie in the closed convex cone (with vertex at 0) generated by this set. Thus the admittance of L_2-controls instead of L_∞-controls significantly changes the situation as we will see in a moment.

* This Journal, 20 (1982), pp. 675-694. Addendum received by the editors March 12, 1984, and in revised form October 15, 1984.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. Part of this research was performed during a visit to the Institut für Mathematik der Universität Graz supported by Deutsche Forschungsgemeinschaft.

THEOREM 1. *The following two conditions are equivalent*:

(i) $\{(\dot{x})_{t_1}$: x *solves* (3) *for some* $u \in \mathcal{U}_2(T)\} = L_2([-r, 0], \mathbb{R}^n)$.

(ii) $P(t) = \mathbb{R}^n$ *for a.a.* $t \in T_1$.

Furthermore if for a.a. $t \in T_1$ *the cone* $P(t)$ *is a subspace of* $\mathbb{R}^n$, *the following two conditions are equivalent*:

(iii) $\mathcal{U}_\infty(T_1) = \{\lambda u$: $\lambda \in \mathbb{R}_+$, $u \in L_\infty(T_1, \mathbb{R}^m)$ *with* $u(t) \in \operatorname{co} f(x_t^0, \Omega) - f(x_t^0, v^0(t))$ *a.e.*$\}$.

(iv) *For some* $\delta > 0$

$$0 \in \operatorname{relint}_\delta \operatorname{co} f(x_t^0, \Omega) - f(x_t^0, v^0(t)) \quad \text{for a.a. } t \in T_1.$$

Here relint Q of $Q \subset \mathbb{R}^n$ denotes the interior of Q with respect to the smallest linear subspace containing Q.

The proof of these results will be postponed until after Theorem 2. First we discuss their significance.

It is immediate from the proof of Theorem 1 that the set at the left-hand side of the equation in (i) does not change, if $u|[t_0, t_1 - t]$ is required to lie in L_∞. The cone $\{\lambda u$: $\lambda \in \mathbb{R}_+$, $u \in L_\infty(T, \mathbb{R}^m)$ with $u(t) \in \operatorname{co} f(x_t^0, \Omega) - f(x_t^0, v^0(t))$ a.e.$\}$ corresponds to the *cone of admissible directions* for the control constraint in L_∞ in the fixed final state optimal control problem. By [4, Example 1.1], the L_2-closure of $\mathcal{U}_\infty(T_1)$ coincides with $\mathcal{U}_2(T_1)$. Hence, taken together, the regularity condition (2) (being equivalent to (ii), (iv)) means by (i) and (iii) that the L_2-closure of the cone of admissible directions in L_∞ is mapped onto $L_2([-r, 0], \mathbb{R}^n)$ under the linearized control-to-final-state-velocity-map. Thus the assumptions of [4, Thm. 1.2] can be verified and Lagrange multipliers in $W^{1,\infty}([-r, 0], \mathbb{R}^n)$ can be identified with functions in $W^{1,2}([-r, 0], \mathbb{R}^n)$ (observe that the finite dimensional part $x(t_1 - r) = \phi(-r)$ does not pose any problem here).

Furthermore, Theorem 1 shows very clearly, where the uniformity condition (that is the δ-bound) in (2) comes in: It guarantees that the cone $\mathcal{U}_\infty$ defined by pointwise restrictions is not bigger than the cone of admissible directions (on the relevant interval T_1).

Thus Theorem 1 clarifies the relation between the regularity condition (2) and the required controllability condition for a linearized system and embeds the special situation of [2] into the general Banach space setting of [4].

Remark 1. The role of the uniformity condition as interpreted above shows that the regularity condition might be weakened somewhat, since we are only interested in the L_2-closure of the cone of admissible directions: If $\operatorname{co} f(x_t^0, \Omega) - f(x_t^0, v^0(t))$ "shrinks fast enough" around zero for some point $\bar{t} \in T_1$ and (ii) is satisfied, $\mathcal{U}_2(T_1)$ will still coincide with this L_2-closure and the surjectivity condition is satisfied.

For a proof of the result above, we consider the following slightly more general problem of controllability under positivity constraints for the control:

$$x_{t_0} = 0, \qquad \dot{x}(t) = L(t)x_t + B(t)u(t) \quad \text{a.a. } t \in T, \tag{5}$$

$$u(t) \in P(t) \qquad \text{a.a. } t \in T \tag{6}$$

where $L: T \to \mathcal{L}(C([-r, 0], \mathbb{R}^n), \mathbb{R}^n)$ with $t \mapsto L(t)\phi$ measurable for all $\phi \in C([-r, 0], \mathbb{R}^n)$ and ess sup $\|L(t)\| < \infty$, $B \in L_\infty(T, \mathbb{R}^{n \times m})$; and $P(t) \subset \mathbb{R}^m$ is a closed convex cone with vertex at zero and $t \mapsto P(t)$ measurable (see [1, p. 68]) ($\mathcal{L}$ denotes the space of bounded linear maps).

We say that (5), (6) is completely controllable to $W^{1,2}([-r, 0], \mathbb{R}^n)$ if the reachable set $\mathscr{R}$ defined by

$$\mathscr{R} := \{x_{t_1}: x \text{ solves (5) for some control } u \in L_2(T, \mathbb{R}^m) \text{ satisfying (6)}\},$$

coincides with this space.

Define the multiplication operator $B: L_2(T_1, \mathbb{R}^m) \to L_2(T_1, \mathbb{R}^n)$ by

$$(\tilde{B}u)(t) := B(t)u(t) \quad \text{a.a. } t \in T_1,$$

and define the closed convex cone $\tilde{P} \subset L_2(T_1, \mathbb{R}^m)$ by

$$\tilde{P} := \{u \in L_2(T_1, \mathbb{R}^m): u(t) \in P(t) \text{ a.e.}\}.$$

LEMMA. *Suppose that the generalized inverse $B(t)^+$ of $B(t)$ has essentially bounded norm on T_1. Then $\tilde{B}\tilde{P} = L_2(T_1, \mathbb{R}^n)$ iff $B(t)P(t) = \mathbb{R}^n$ a.e. on T_1.*

Proof. One direction is trivial. Conversely, suppose that $B(t)P(t) = \mathbb{R}^n$ a.e. on T_1. By [5, Lemma 3] our hypothesis means that $\tilde{B}$ has a closed range. Consider the set

$$\{u \in L_2(T_1, \mathbb{R}^m): u(t) \in P_1(t) \text{ a.e.}\}$$

where $P_1(t)$ is the projection of $P(t)$ to $[\operatorname{Ker} B(t)]^\perp$. This set is a closed linear subspace which is mapped onto a closed linear subspace X of $L_2(T_1, \mathbb{R}^n)$, since $\tilde{B}|\{u \in L_2(T_1, \mathbb{R}^m): u(t) \in [\ker B(t)]^\perp \text{ a.e.}\}$ is a homeomorphism onto the image of $\tilde{B}$. Thus the space $X = \tilde{B}\tilde{P}$ is a closed linear subspace of L_2 which naturally is also dense. This proves the assertion.

We obtain the following result.

THEOREM 2. *Suppose that the generalized inverse $B(t)^+$ of $B(t)$ has an essentially bounded norm on T_1. Then system (5), (6) is completely controllable to $W^{1,2}([-r, 0], \mathbb{R}^n)$ iff the following two conditions are satisfied:*

(i) $\mathscr{R}^f := \{x(t_1 - r): x \text{ solves (5) for some control } u \in \tilde{P}\} = \mathbb{R}^n$;

(ii) $B(t)P(t) = \mathbb{R}^n$ *a.e. on* T_1.

Proof. By the lemma above conditions (i) and (ii) imply $\mathscr{R} = W^{1,2}([-r, 0], \mathbb{R}^n)$. Conversely, condition (i) follows trivially. Suppose that condition (ii) is violated, i.e. there is a subset $T_2 \subset T_1$ of positive Lebesgue measure such that

$$0 \in \partial B(t)P(t), \qquad t \in T_2;$$

here ∂ denotes the boundary.

Equivalently,

$$0 \in \partial\{B(t)P(t) \cap \mathscr{E}\}$$

where $\mathscr{E} := \{y \in \mathbb{R}^n: |y| = 1\}$.

Define for $t \in T_2$

$$\Gamma(t) := \{y \in \mathbb{R}^n: y \in \mathscr{E} \text{ and } \langle y, B(t)p\rangle \leqq 0 \text{ for all } p \in P(t)\},$$

Then Γ has compact values and is measurable. Hence there is a measurable selection γ of Γ satisfying

$$|\gamma(t)| = 1 \quad \text{and} \quad \langle \gamma(t), B(t)p\rangle \leqq 0$$

for a.a. $t \in T_2$ and all $p \in P(t)$.

By Lusin's theorem there exists a closed subset T_3 of T_2 of positive Lebesgue measure such that $\gamma|T_3$ and the components of $L(\cdot)$ considered as maps on T_3 with values in the dual of $W^{1,2}([-r, 0], \mathbb{R}^n)$ are continuous.

Let α be an arbitrary element of $L_2(T_3, \mathbb{R})$ and define $\zeta \in L_2(T_1, \mathbb{R}^n)$ by

$$\zeta(t) = \begin{cases} \alpha(t)\gamma(t) & \text{for } t \in T_3, \\ 0 & \text{otherwise.} \end{cases}$$

Define $x^\alpha \in W^{1,2}(T, \mathbb{R}^n)$ on $[t_0, t_1 - r]$ by

$$x^\alpha(t) := 0 \tag{7}$$

and on T_1 as the unique solution of

$$\dot{x}^\alpha(t) = L(t)x_t^\alpha + \zeta(t). \tag{8}$$

Since we assume that $\mathcal{R} = W^{1,2}([-r, 0], \mathbb{R}^n)$, there is an admissible u^α such that the corresponding trajectory y^α satisfies $y_{t_1}^\alpha = x_{t_1}^\alpha$. Hence on T_1

$$\dot{x}^\alpha(t) = \dot{y}^\alpha(t) = L(t)y_t^\alpha + B(t)u^\alpha(t)$$

and

$$0 = L(t)(x_t^\alpha - y_t^\alpha) + \zeta(t) - B(t)u^\alpha(t).$$

Taking inner products with $\gamma(t)$ in $\mathbb{R}^n$ yields for $t \in T_3$

$$0 = \langle \gamma(t), L(t)(x_t^\alpha - y_t^\alpha)\rangle + \alpha(t)\langle \gamma(t), \gamma(t)\rangle - \langle \gamma(t), B(t)u^\alpha(t)\rangle$$

or

$$\alpha(t) = -\langle \gamma(t), L(t)(x_t^\alpha - y_t^\alpha)\rangle + \langle \gamma(t), B(t)u^\alpha(t)\rangle.$$

The first term at the right-hand side is continuous, and the second one is negative. This contradicts the choice of α as an arbitrary element in L_2 and proves (ii).

Remark 2. The assertion of Theorem 2 is not valid for controllability to $W^{1,\infty}([-r, 0], \mathbb{R}^n)$ with L_∞-controls. In fact, controllability to $W^{1,\infty}$ only implies that the interior of $B(t)P(t)$ is nonempty for a.a. $t \in T_1$. This follows from [3, Thm. 3.3 and Example 3.1], and is a remarkable difference between controllability to $W^{1,\infty}$ and $W^{1,2}$. Taking up the line of argument in [3, Remark 3.3], Theorem 2 shows that one *cannot* prepare the reaching of an arbitrary $W^{1,2}$-function ϕ by a special way of reaching $\phi(-r)$ at $t_1 - r$. The reason is that in each neighbourhood of ϕ there is an element ξ such that $\dot{\phi} - \dot{\xi}$ is unbounded.

Remark 3. In some sense the result of Theorem 2 is negative: Controllability to $W^{1,2}$ can only be achieved if not only the well-known and strong rank condition on $B(t)$ is satisfied, but also the "positivity cone" $P(t)$ is the whole space $\mathbb{R}^m$ on the final interval. However, the rank condition is irrelevant for the *relaxed* optimal control problem (if the control appears nonlinearly). Here $B(t)$ is the identity matrix and generically, $\operatorname{int} P(t) = \operatorname{int} \operatorname{co} f(x_t^0, v^0(t)) \neq \emptyset$. Furthermore, the condition $P(t) = \mathbb{R}^m$ on T_1 means for the optimal control problem that the control constraint is not active on the final interval. Theorem 2 explains why this strong assumption has to be made.

Proof of Theorem 1. The equivalence of (i) and (ii) is immediate from the proof of Theorem 2. Furthermore, it is clear that (iv) implies (iii). Conversely, suppose that (ii) is violated. Due to our continuity assumption on $\dot{\phi}(t - t_1) = f(x_t^0, v^0(t))$, $t \in T_1$, this means that there is $\bar{t} \in T_1$ with

$$f(x_{\bar{t}}^0, v^0(\bar{t})) \in \partial \operatorname{co} f(x_{\bar{t}}^0, \Omega)$$

in the linear subspace spanned by $P(\bar{t})$. Thus there exists $\bar{u} \neq 0$ in the linear subspace spanned by $P(\bar{t})$ such that

$$\langle \bar{u}, p\rangle \leqq 0 \quad \text{for all } p \in P(\bar{t}).$$

Then there exists a function $\bar{u}(\cdot)$ in $\mathcal{U}_\infty(T_1)$ with

$$\bar{u}(t) = \bar{u} \quad \text{for a.a. } t \text{ in a neighbourhood of } \bar{t}.$$

However, there is no $\lambda \geqq 0$ such that

$$\frac{1}{\lambda}\bar{u} \in \operatorname{co} f(x_t^0, \Omega) - f(x_t^0, v^0(t)) \text{ a.e.}$$

This proves the equivalence of (iii) and (iv).

REFERENCES

[1] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics 580, Springer-Verlag, Berlin, 1977.

[2] F. COLONIUS, *The maximum principle for relaxed hereditary differential systems with function space end condition*, this Journal, 20 (1982), pp. 695-712.

[3] ———, *Stable and regular reachability for relaxed hereditary differential systems*, this Journal, 20 (1982), pp. 675-694.

[4] ———, *A note on the existence of Lagrange multipliers*, Appl. Math. Optim., 10 (1983), pp. 187-191.

[5] S. KURCYUSZ AND A. W. OLBROT, *On the closure in $W^{1,q}$ of the attainable subspace of linear time lag systems*, J. Differential Equations, 24 (1977), pp. 29-50.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 6, November 1985

REALIZATION THEORY FOR MULTIVARIATE STATIONARY GAUSSIAN PROCESSES*

ANDERS LINDQUIST† AND GIORGIO PICCI‡

Abstract. This paper collects in one place a comprehensive theory of stochastic realization for continuous-time stationary Gaussian vector processes which in various pieces has appeared in a number of our earlier papers. It begins with an abstract state space theory, based on the concept of *splitting subspace.* These results are then carried over to the spectral domain and described in terms of Hardy functions. Finally, differential-equations type stochastic realizations are constructed. The theory is coordinate-free, and it accommodates infinite-dimensional representations, minimality and other systems-theoretical concepts being defined by subspace inclusion rather than by dimension. We have strived for conceptual completeness rather than generality, and the same framework can be used for other types of stochastic realization problems.

CONTENTS

1. Introduction. The following inverse problem is of central importance in stochastic systems theory. Given a stationary Gaussian vector process $\{y(t);\, t\in\mathbb{R}\}$, find a vector-valued stationary Gaussian Markov process $\{x(t);\, t\in\mathbb{R}\}$ of smallest possible dimension so that

$$y(t) = Cx(t) \tag{1.1}$$

for some matrix C, and determine a stochastic differential equation for x. This is the *stochastic realization problem* and the representation is called a *minimal stochastic realization.*

This problem, first formulated by Kalman [21] in 1965, has generated a rather extensive literature. Most notable among the early contributions are the papers by Anderson [2] and Faurre [11], the main focus of which is the realization of spectral factors and the Yakubovich-Kalman-Popov lemma. The more recent work by Ruckebusch [39], Lindquist and Picci [25], and Pavon [36] is geared toward the characterization of Markovian representations in terms of the information carried by the given process. During the last decade, the bulk of the papers on stochastic realization theory have been concerned with geometric state space construction in Hilbert space. Here the forerunners are Akaike [1] and Picci [37], whereas the most comprehensive contributions are due to Lindquist and Picci [26]-[32] and Ruckebusch [40]-[44]. A more extensive bibliography can be found in our survey paper [24].

* Received by the editors May 22, 1984, and in revised form February 12, 1985.

† Department of Mathematics, Royal Institute of Technology, S-100 44 Stockholm, Sweden. The work of this author was partially supported by the National Science Foundation under grant ECS-8215660.

‡ Istituto di Elettrotecnica ed Elettronica, via Gradenigo 6A, 35131 Padova, Italy.

There are both conceptual and practical reasons why this problem is important. On the conceptual side, a theory of stochastic realization should give a firm foundation of the idea of *state* and *state space models.* Clearly this is of central importance in setting stochastic systems theory on a sound mathematical basis. The purpose of this paper is to present such a theory in which the idea of state is defined through a fundamental property of conditional independence (splitting), a natural generalization of the property of state in the deterministic theory. This point of view provides a general framework for stochastic modeling in which problems of stochastic systems theory can be set.

Important areas for potential application of this theory include identification, stochastic model reduction, and stochastic control, and there is preliminary evidence that the basic ideas presented here will prove to be fruitful. Moreover, there are already problems in estimation theory which have been successfully tackled by such an approach. Some cases in point are smoothing [49], interpolation [51], and, in general, problems with a noncausal information flow. Possible extensions of the theory presented here to the nonlinear (non-Gaussian) case will provide solution to even wider areas of important applications. For example, realization theory of finite-state processes would provide powerful technics to solve important problems in communication theory.

Stochastic realization theory is *not* a generalization of deterministic input–output realization theory. Characteristic of the stochastic problem is the fact that there are many different (minimal) causality structures which describe the same external behavior, the basic problem being to classify all of them. Note that a similar problem is encountered in J. C. Willems' deterministic realization theory [52] for "signals", a theory which has many points of contact with ours.

This invited paper collects in one place a reasonably self-contained treatment of the geometric theory of stochastic realization which in various pieces has appeared in a number of our previous papers [26]–[32], some of which are published in volumes of limited availability. We have strived for conceptual completeness rather than generality. Consequently, many of the results presented here have generalizations in various directions, some straightforward and others more nontrivial. The basic conceptual framework, however, is the same.

The need for a geometric theory of stochastic realization is illustrated by the problem formulation above. As it stands, the problem may not be meaningful unless the given process has a rational spectral density and hence a finite-dimensional representation is possible. In the general case, a representation of type (1.1) exists only under certain technical conditions (which we do not want to introduce at the beginning). Moreover, the concept of *minimality* needs a natural dimension-free formulation which also covers the infinite-dimensional situation. Finally, a geometric theory is *coordinate-free* and hence allows us to factor out, in the first analysis, the properties of the realizations which depend only on the choice of coordinates and may unduly complicate the picture.

To this end, let us reformulate the above problem in terms of Hilbert space geometry. Let $\{y(t);\, t\in\mathbb{R}\}$ be a stationary Gaussian stochastic vector process which is mean-square continuous and centered. Consider the space $\hat{H}$ of all finite linear combinations of the random variables $\{y_k(t);\, t\in\mathbb{R},\, k=1,2,\cdots,m\}$. Endowed with the inner product $\langle \xi, \eta\rangle := E\{\xi\eta\}$, where $E\{\cdot\}$ denotes mathematical expectation, $\hat{H}$ is a pre-Hilbert space. Let H be the Hilbert space obtained by taking the closure of $\hat{H}$; this is known as the Gaussian space of y [35]. A standard argument [38, p. 15] shows that there is a group $\{U_t;\, t\in\mathbb{R}\}$ of unitary operators on H such that $U_t y_k(s) = y_k(s+t)$ for all $s, t\in\mathbb{R}$ and $k=1,2,\cdots,m$. Since y is mean-square continuous, the group

$\{U_t; t \in \mathbb{R}\}$ is strongly continuous. We shall use the notation $E^X\lambda$ to denote the orthogonal projection of $\lambda \in H$ onto a subspace[1] X of H. This notation is motivated by the fact that $E^X\lambda$ coincides with the conditional expectation $E\{\lambda \mid \mathscr{X}\}$ where $\mathscr{X}$ is the σ-field generated by the random variables in X [9].

Consider the class of subspaces X of H with the properties

(i) $y_k(0) \in X$ for $k = 1, 2, \cdots, m$;

(ii) X is *Markovian* in the sense that

$$\langle \lambda - E^X\lambda, \mu - E^X\mu \rangle = 0 \quad \text{for } \lambda \in X^-, \mu \in X^+$$

where X^- and X^+ are the closed linear hulls of $\{U_tX; t \leqq 0\}$ and $\{U_tX; t \geqq 0\}$ respectively;

(iii) X is *minimal* in the sense that if X_1 is a subspace of X and X_1 satisfies (i) and (ii), then $X_1 = X$.

The term *Markovian* is motivated by the fact, as we shall see below (Proposition 2.1), that (ii) is equivalent to each of the two conditions

$$E^{X^-}\lambda = E^X\lambda \quad \text{for } \lambda \in X^+, \tag{1.2a}$$

$$E^{X^+}\lambda = E^X\lambda \quad \text{for } \lambda \in X^-. \tag{1.2b}$$

For reasons to be reported in § 3 (Proposition 3.1), a subspace X satisfying (i) and (ii) will be called a *Markovian splitting subspace.*

What is then the connection between such subspaces and the stochastic realization problem stated above? Let us *for the moment* assume that X has finite dimension n, and let $\{x_1, x_2, \cdots, x_n\}$ be a basis in X. Then, in view of property (i), there is an $m \times n$ matrix $C = \{c_{ij}\}$ such that $y_i(0) = \sum_{j=1}^n c_{ij}x_j$ for $i = 1, 2, \cdots, m$. Consequently,

$$y(t) = Cx(t) \tag{1.3}$$

where $\{x(t); t \in \mathbb{R}\}$ is the n-dimensional stationary stochastic process defined by setting $x_k(t) := U_tx_k$ for $k = 1, 2, \cdots, n$. Under suitable geometric conditions on X (to be introduced in § 3) this process is purely nondeterministic [38]; *for the sake of this example,* we shall assume that this is the case. Since

$$U_tX = \text{span}\,\{x_1(t), x_2(t), \cdots, x_n(t)\}, \tag{1.4}$$

condition (1.2a), shifted by U_t, is equivalent to

$$E\{x_k(s) \mid \mathscr{X}_t^-\} = E\{x_k(s) \mid \mathscr{X}_t\} \quad \text{for } s \geqq t, \quad k = 1, 2, \cdots, n \tag{1.5}$$

where $\mathscr{X}_t^-$ and $\mathscr{X}_t$ are the σ-fields generated by $\{x_k(\tau); \tau \leqq t, k = 1, 2, \cdots, m\}$ and $\{x_k(t); k = 1, 2, \cdots, n\}$ respectively. Consequently x is a vector *Markov process.* Finally, as we shall see below, condition (iii) insures that the dimension n is as small as possible. The condition $X \subset H$ is not implied by the original problem formulation, but it is not unnatural since the process y is the only thing given. Such realizations are called *internal* [25]. However, several of the applications mentioned above require that we consider the noninternal situation when H is imbedded in a larger Hilbert space. Although many of our results remain valid in the noninternal setting and others can be generalized [43], [44], we shall restrict ourselves here to a simple prototype problem.

It is well known that a vector Markov process *of the type described above* has a representation

$$x(t) = \int_{-\infty}^{t} e^{A(t-\sigma)}B\,du(\sigma) \tag{1.6}$$

[1] In this paper a *subspace* is assumed to be closed.

where A and B are matrices, u is a vector-valued orthogonal increment process with components in H, and the integral is defined in quadratic mean [11]. Together with (1.3) this yields a *forward* stochastic realization

$$\begin{aligned} dx &= Ax\,dt + B\,du, \\ y &= Cx. \end{aligned} \tag{1.7}$$

The forward property is characterized by $X \subset H^-(du)$, where $H^-(du)$ is the subspace generated by the components of the past increments $\{u(t)-u(s);\ t, s \leqq 0\}$. By symmetry and (1.2b), there is also a representation

$$x(t) = -\int_t^\infty e^{\bar{A}(\sigma - t)} \bar{B}\, d\bar{u}(\sigma) \tag{1.8}$$

which corresponds to a *backward* stochastic realization

$$\begin{aligned} dx &= -\bar{A}x\,dt + \bar{B}\,d\bar{u}, \\ y &= Cx. \end{aligned} \tag{1.9}$$

This realization is backward because $X \subset H^+(d\bar{u})$, the subspace generated by the components of the future increments $\{\bar{u}(t)-\bar{u}(s);\ t, s \geqq 0\}$. Characterizing Markovian representations in terms of pairs of realizations, one evolving forward and one backward, is one of the key ideas in [25] and in the present work. It is well known and easy to show that the transfer functions

$$W(s) = C(sI - A)^{-1}B, \tag{1.10a}$$

$$\bar{W}(s) = C(sI + \bar{A})^{-1}\bar{B} \tag{1.10b}$$

are rational spectral factors of y, W having all its poles in the left and $\bar{W}$ all its poles in the right half plane.

It follows from finite-dimensional stochastic realization theory [2], [11], [12] that (1.7) is a minimal stochastic realization if and only if (a) it is *reachable*, i.e. $[B, AB, A^2B, \cdots]$ is full rank, (b) it is *observable*, i.e. $[C', (CA)', (CA^2)', \cdots]$ is full rank (where prime denotes transpose), and (c) W has minimal degree (among spectral factors). Likewise (1.9) is minimal if and only if (a)′ it is *controllable*, i.e. $[\bar{B}, \bar{A}\bar{B}, \bar{A}^2\bar{B}, \cdots]$ is full rank, (b)′ it is *constructible*, i.e. $[C', (C\bar{A})', (C\bar{A}^2)', \cdots]$ is full rank[2], and (c)′ $\bar{W}$ has minimal degree. As can be easily checked, $x(0)$ being a basis in X automatically takes care of conditions (a) and (a)′, and hence they will not occur in the geometric theory. Conditions (b), (c), (b)′ and (c)′ will be given natural geometric and function theoretic characterizations below which hold also in the infinite-dimensional case. We shall see, for example, that minimality is equivalent not only to (b)+(c) or to (b)′+(c)′ but also to (b)+(b)′.

This paper divides naturally into three parts. The first part, consisting of §§ 3–5, is devoted to a characterization of the class of Markovian splitting subspaces and an analysis of their systems-theoretical properties. Section 2 is a preliminary in which we define the concept of *perpendicular intersection*, introduced in [29].

In the second part, consisting of §§ 6 and 7, the geometry is described in terms of Hardy spaces, and the Markovian splitting subspaces are characterized by pairs $(W, \bar{W})$ of spectral factors. This part of the theory has some connections with Lax-Phillips scattering theory [23].

[2] Using the terms *controllable* and *constructible* instead of *reachable* and *observable* when referring to a system evolving backwards is in agreement with accepted terminology in systems theory [22].

Finally, in §§ 8–10, we assign to each Markovian splitting subspace X two stochastic realizations, a forward one with transfer function W and a backward one with transfer function $\bar{W}$, having their systems-theoretical properties prescribed by X. Moreover, we study the relationships between realizations corresponding to different splitting subspaces.

2. Perpendicular intersection. Let A, B and X be subspaces of a real Hilbert space H. We shall say that A and B are *conditionally orthogonal given* X if

$$\langle \alpha - E^X\alpha, \beta - E^X\beta\rangle = 0 \quad \text{for } \alpha \in A, \quad \beta \in B. \tag{2.1}$$

This will be denoted $A \perp B|X$. When X is the trivial subspace, i.e. $X = 0$, this reduces to the usual orthogonality $A \perp B$. Conditional orthogonality is orthogonality after subtracting the components in X. We write $A \vee B$ to denote the vector sum, i.e. the closure of $\{\alpha + \beta \,|\, \alpha \in A, \beta \in B\}$ and $A \oplus B$ to denote orthogonal direct sum; $A \ominus B$ is the subspace $C \subset A$ such that $B \oplus C = A$; $B^\perp$ is the orthogonal complement of B in H, i.e. $B^\perp = H \ominus B$. Finally, $E^A B = \{E^A\beta \,|\, \beta \in B\}$. This space may not be closed, and we shall write $\bar{E}^A B$ to denote the closure.

PROPOSITION 2.1. *The following statements are equivalent.*

(i) $A \perp B|X$.
(ii) $B \perp A|X$.
(iii) $(A \vee X) \perp B|X$.
(iv) $E^{A\vee X}\beta = E^X\beta$ *for* $\beta \in B$.
(v) $(A \vee X) \ominus X \perp B$.
(vi) $E^A\beta = E^A E^X\beta$ *for* $\beta \in B$.

Proof. The equivalence between (i), (ii) and (iii) follows directly from the definition. Since $(\beta - E^X\beta) \perp X$, relation (2.1) may be written $\langle \alpha, \beta - E^X\beta\rangle = 0$. Therefore, (iii) is equivalent to $(\beta - E^X\beta) \perp A \vee X$, i.e. $E^{A\vee X}(\beta - E^X\beta) = 0$, which is precisely (iv). Moreover, (i) is equivalent to $(\beta - E^X\beta) \perp A$, i.e. $E^A(\beta - E^X\beta) = 0$, which is the same as (vi). Finally, set $Z := (A \vee X) \ominus X$; then $A \vee X = X \oplus Z$, i.e. $E^{A\vee X}\beta = E^X\beta + E^Z\beta$. Hence (iv) is equivalent to $E^Z\beta = 0$ for $\beta \in B$, i.e. $Z \perp B$. This is (v). □

PROPOSITION 2.2. *Let* $A \perp B|X$. *Then*

$$A \cap B \subset X. \tag{2.2}$$

Proof. Let $\lambda \in A \cap B$. Then $\lambda \perp \lambda|X$, i.e. $\lambda \in X$. □

PROPOSITION 2.3. *Let A and B be subspaces of H. Then*

$$A \perp B|\bar{E}^A B. \tag{2.3}$$

Moreover, any $X \subset A$ *such that* $A \perp B|X$ *contains* $\bar{E}^A B$.

To prove this we need the following decomposition.

LEMMA 2.1. *Let A and B be subspaces of H. Then*

$$A = \bar{E}^A B \oplus (A \cap B^\perp). \tag{2.4}$$

Proof. Let $\alpha \in A$ and $\beta \in B$. Then $\langle \alpha, E^A\beta\rangle = \langle \alpha, \beta\rangle$. Consequently, if $\alpha \perp \bar{E}^A B$, then $\alpha \in B^\perp$. □

Proof of Proposition 2.3. If $X \subset A$, $A \perp B|X$ is equivalent to $A \ominus X \perp B$ (Proposition 2.1). In particular, this is satisfied by $X = \bar{E}^A B$ (Lemma 2.1). In general, $A \ominus X \subset A \cap B^\perp$, i.e., $X \supset \bar{E}^A B$ (Lemma 2.1). □

Suppose $A \perp B|X$. Then it follows trivially from the definition that, if $A_1 \subset A$ and $B_1 \subset B$, then $A_1 \perp B_1|X$. A more interesting question is how far A and B can be *expanded* while remaining conditionally orthogonal given X.

THEOREM 2.1. *Let A_0 and B_0 be subspaces such that $A_0 \vee B_0 = H$, and suppose that $A_0 \perp B_0|X$. Let $A \supset A_0$ and $B \supset B_0$. Then $A \perp B|X$ if and only if*

$$\begin{aligned} A &\subset A_0 \vee X, \\ B &\subset B_0 \vee X. \end{aligned} \tag{2.5}$$

If the upper bounds are attained, i.e. $A = A_0 \vee X$ and $B = B_0 \vee X$, then $X = A \cap B$.

Proof. By Proposition 2.1, $(A_0 \vee X) \perp (B_0 \vee X)|X$. Therefore, $A \perp B|X$ whenever (2.5) holds. Conversely, assume that $A \perp B|X$. Then $(A \vee X) \perp B_0|X$, and therefore $Z \perp B_0|X$ when $Z := (A \vee X) \ominus (A_0 \vee X)$, i.e. $Z \perp (B_0 \vee X) \ominus X$ (Proposition 2.1). But, by definition, $Z \perp (A_0 \vee X)$ and therefore $Z \perp A_0 \vee B_0 = H$, i.e. $Z = 0$. Consequently $A \vee X = A_0 \vee X$, i.e. the first of relations (2.5) must hold. The second follows by symmetry. If $A = A_0 \vee X$ and $B = B_0 \vee X$, then $X \subset A \cap B$. But by Proposition 2.2, $A \cap B \subset X$. Hence $X = A \cap B$.

The following proposition describes the geometry of the maximal spaces in Theorem 2.1.

PROPOSITION 2.4. *The following conditions are equivalent.*

(i) $A \perp B|A \cap B$.
(ii) $E^A B = A \cap B$.
(iii) $E^B A = A \cap B$.
(iv) $E^A B = E^B A$.

Proof. First, suppose that (i) holds. Then, by Proposition 2.1 (iv), $E^A B = E^{A \cap B} B = A \cap B$, which is (ii). Condition (iii) follows by symmetry, exchanging A and B. Hence (iv) follows. Conversely, (ii) or (iii) and Proposition 2.3 imply (i). (Note that (ii) implies that $E^A B$ is closed and therefore $\bar{E}^A B = E^A B$.) Finally, if (iv) holds, $E^A B$, and hence $\bar{E}^A B$, is contained in $A \cap B$. But, by Propositions 2.2 and 2.3, $A \cap B \subset \bar{E}^A B$. Hence $A \cap B = \bar{E}^A B$. Consequently (i) holds (Proposition 2.3).

We shall say that two subspaces A and B satisfying the conditions of Proposition 2.4 *intersect perpendicularly.* As we have seen, perpendicular intersection corresponds to maximal A and B in Theorem 2.1. The upper bound is also attained in the inclusion $A \cap B \subset X$ of Proposition 2.2. Note that, for any pair (A, B) of perpendicularly intersecting subspaces, $E^A B$ is closed.

THEOREM 2.2. *Let A and B be subspaces such that $A \vee B = H$. Then the following conditions are equivalent.*

(i) *A and B intersect perpendicularly.*
(ii) $B^\perp \subset A$.
(iii) $H = A^\perp \oplus (A \cap B) \oplus B^\perp$.
(iv) *E^A and E^B commute.*

Proof. Set $X := A \cap B$. If (i) holds, $X = \bar{E}^A B$, and hence $A \ominus X \perp B$ (Lemma 2.1). But, since $X \subset B$ and $A \vee B = H$, $(A \ominus X) \oplus B = H$, and therefore $A \ominus X = B^\perp$, i.e. $A = X \oplus B$. Hence both (ii) and (iii) follow. Each of the conditions (ii) and (iii) implies the existence of a subspace X with the property $H = A^\perp \oplus X \oplus B^\perp$, so that if $\lambda \in H$,

$$E^A E^B \lambda = E^X E^B \lambda + E^{B^\perp} E^B \lambda = E^X \lambda \tag{2.6}$$

and

$$E^B E^A \lambda = E^X E^A \lambda + E^{A^\perp} E^A \lambda = E^X \lambda \tag{2.7}$$

and therefore (iv) follows. It just remains to prove that (iv) implies (i). But, $E^A E^B H = E^B E^A H$ yields $E^A B = E^B A$, i.e. A and B intersect perpendicularly (Proposition 2.4). □

3. The geometry of splitting subspaces. Let H be a real separable Hilbert space, let $\{U_t; t \in \mathbb{R}\}$ be a strongly continuous group of unitary operators on H, and let H^- and H^+ be subspaces enjoying the invariance properties

$$U_t H^- \subset H^- \quad \text{for } t \leqq 0, \tag{3.1a}$$

$$U_t H^+ \subset H^+ \quad \text{for } t \geqq 0 \tag{3.1b}$$

and together spanning H, i.e. $H^- \vee H^+ = H$.

Although these are the only assumptions needed for the geometric theory of §§ 3–5, the situation we have in mind is the one delineated in the Introduction: H is the Gaussian space of an m-dimensional stationary Gaussian vector process, which is mean-square continuous and centered, and $\{U_t; t \in \mathbb{R}\}$ is the group of shifts: $U_t y_k(s) = y_k(s+t)$. Moreover,

$$\begin{aligned} H^- &:= \overline{\text{span}}\,\{y_k(t); t \leqq 0, k = 1, 2, \cdots, m\}, \\ H^+ &:= \overline{\text{span}}\,\{y_k(t); t \geqq 0, k = 1, 2, \cdots, m\} \end{aligned} \tag{3.2}$$

where $\overline{\text{span}}\,\{\cdot\}$ denotes closed linear hull. Hence we shall refer to H^- and H^+ as the *past space* and the *future space* respectively.

We shall say that X is a *splitting subspace* if H^- and H^+ are conditionally orthogonal given X, i.e. $H^- \perp H^+ | X$. According to Proposition 2.1, this is equivalent to each of the two conditions

$$\begin{aligned} E^{H^- \vee X} \lambda &= E^X \lambda \quad \text{for } \lambda \in H^-, \\ E^{H^+ \vee X} \lambda &= E^X \lambda \quad \text{for } \lambda \in H^+. \end{aligned} \tag{3.3}$$

Consequently, a splitting subspace X can be thought of as a "memory" or a "sufficient statistic" containing all information about the past needed in predicting the future, or, equivalently, all the information about the future required to estimate the past. *Splitting subspace* is a concept originally introduced by McKean [34] in a somewhat more restricted sense. A splitting subspace is said to be *minimal* if it contains no proper subspace which is also a splitting subspace. The spaces H, H^- and H^+ are splitting subspaces, but in general they are not minimal.

A subspace X is said to be *Markovian* if the subspaces X^- and X^+ generated by $\{U_t X; t \leqq 0\}$ and $\{U_t X; t \geqq 0\}$ are conditionally orthogonal given X, i.e. $X^- \perp X^+ | X$. This is condition (ii) in § 1, and, as mentioned there, it is equivalent to each of the conditions (1.2) (Proposition 2.1).

We shall now reformulate the geometric problem of § 1, justifying the name *Markovian splitting subspace* introduced there.

PROPOSITION 3.1. *The subspace X satisfies the conditions*

(i) $y_k(0) \in X$, $k = 1, 2, \cdots, m$,

(ii) *X is Markovian*

if and only if X is a Markovian splitting subspace.

Proof. (if): Since X is a splitting subspace, it follows from Proposition 2.2 that $H^- \cap H^+ \subset X$. But $y_k(0) \in H^- \cap H^+$ for $k = 1, 2, \cdots, m$, and therefore (i) follows. Condition (ii) is part of the assumption. (only if): Condition (i) implies that $H^- \subset X^-$ and $H^+ \subset X^+$. Hence the splitting property of X follows from the Markovian property. □

The following characterization of the class of splitting subspaces will be of central importance in what follows.

THEOREM 3.1. [28], [29]. *A subspace X is a splitting subspace if and only if*

$$X = S \cap \bar{S} \tag{3.4}$$

for some pair $(S, \bar{S})$ of perpendicularly intersecting subspaces such that $S \supset H^-$ and $\bar{S} \supset H^+$. The correspondence $X \leftrightarrow (S, \bar{S})$ is one-one, S and $\bar{S}$ being given by

$$\begin{aligned} S &= H^- \vee X, \\ \bar{S} &= H^+ \vee X. \end{aligned} \tag{3.5}$$

Proof. (if): Suppose that S and $\bar{S}$ intersect perpendicularly. Then $S \perp \bar{S}|X$ where $X = S \cap \bar{S}$ (Proposition 2.4). But, since $S \supset H^-$ and $\bar{S} \supset H^+$, this implies that $H^- \perp H^+|X$, i.e. X is a splitting subspace. (only if): Suppose $H^- \perp H^+|X$. Let S and $\bar{S}$ be defined by (3.5). Then, by Theorem 2.1, $S \perp \bar{S}|X$ and $X = S \cap \bar{S}$. This implies that S and $\bar{S}$ intersect perpendicularly (Proposition 2.4).

(one-one): Suppose that S and $\bar{S}$ are perpendicularly intersecting subspaces such that $S \supset H^-$ and $\bar{S} \supset H^+$. Then $X = S \cap \bar{S}$ is a splitting subspace, i.e. $H^- \perp H^+|X$. We need to show that $S = H^- \vee X$ and $\bar{S} = H^+ \vee X$. But S contains H^- and X, and $\bar{S}$ contains H^+ and X; hence $S \supset H^- \vee X$ and $\bar{S} \supset H^+ \vee X$. On the other hand, $S \perp \bar{S}|X$ (Proposition 2.4), and therefore $S \subset H^- \vee X$ and $\bar{S} \subset H^+ \vee X$ (Theorem 2.1), establishing the required equalities. □

COROLLARY 3.1. [28]. *In Theorem* 3.1, (3.4) *can be exchanged for $X = E^S\bar{S}$ or $X = E^{\bar{S}}S$.*

Proof. Follows immediately from Proposition 2.4. □

We shall write $X \sim (S, \bar{S})$ to exhibit the unique pair $(S, \bar{S})$ corresponding to X. The geometry of Theorem 3.1 can be illustrated as in Fig. 1. It also illustrates

COROLLARY 3.2. [28]. *A subspace X is a splitting subspace if and only if there are subspaces $S \supset H^-$ and $\bar{S} \supset H^+$ such that*

$$H = S^\perp \oplus X \oplus \bar{S}^\perp. \tag{3.6}$$

The pair $(S, \bar{S})$ is the same as in Theorem 3.1, *i.e. $X \sim (S, \bar{S})$.*

Proof. (if): Relation (3.6) implies that $\bar{S}^\perp \subset S$, and therefore S and $\bar{S}$ intersect perpendicularly (Theorem 2.2). Also, by Theorem 2.2 (iii), $X = S \cap \bar{S}$. Then the rest follows from Theorem 3.1.

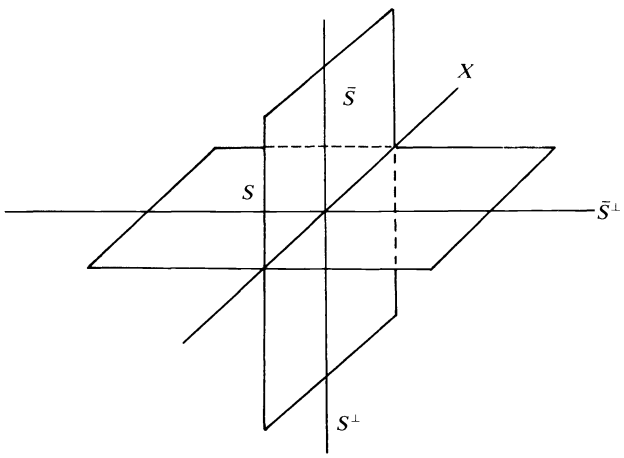


FIG. 1

(only if): In view of Theorem 3.1, it only remains to show that, if S and $\bar{S}$ intersect perpendicularly, (3.6) holds with $X = S \cap \bar{S}$. But this follows from Theorem 2.2. □

Equation (3.6) is analogous to the decomposition in terms of incoming and outgoing subspaces in Lax-Phillips scattering theory [23]. Adding the invariance conditions of Theorem 3.2 below, $\bar{S}^{\perp}$ corresponds to the incoming and $S^{\perp}$ to the outgoing subspace. The parallels will be more apparent in § 7, as we turn to Hardy space representation.

The splitting subspace $X \sim (S, \bar{S})$ is said to be *proper* if both $S^{\perp}$ and $\bar{S}^{\perp}$ are full range.[3] Since $S^{\perp}$ and $\bar{S}^{\perp}$ are the pieces of H in (3.6) which we discard, properness is to a certain extent an indication that the splitting subspace X offers nontrivial data reduction; H, H^- and H^+ are *not* proper.

THEOREM 3.2. [28]. *Let $X \sim (S, \bar{S})$ be a splitting subspace. Then X is Markovian if and only if*

$$U_t S \subset S \quad \text{for } t \leqq 0, \tag{3.7a}$$

$$U_t \bar{S} \subset \bar{S} \quad \text{for } t \geqq 0. \tag{3.7b}$$

Proof. (if): Since $X \subset S$, (3.7a) implies that $U_t X \subset S$ for $t \leqq 0$, i.e. $X^- \subset S$. In the same way, (3.7b) implies that $X^+ \subset \bar{S}$. Therefore, since $S \perp \bar{S} | X$, we have $X^- \perp X^+ | X$. (only if): Suppose that $X \sim (S, \bar{S})$ is a Markovian splitting subspace. Then $y_k(0) \in X$ for $k = 1, 2, \cdots, m$ (Proposition 3.1), and therefore $X^- \supset H^-$ and $X^+ \supset H^+$. Moreover, $X^- \perp X^+ | X$, and consequently $X = X^- \cap X^+$ (Theorem 2.1). Hence X^- and X^+ intersect perpendicularly (Proposition 2.4). Then, by Theorem 3.1, $X \sim (X^-, X^+)$ is a splitting subspace. But then, in view of the one-one correspondence $X \leftrightarrow (S, \bar{S})$, we must have $S = X^-$ and $\bar{S} = X^+$, which clearly have the required invariance properties. □

From Theorems 3.1 and 3.2 we see that $H \sim (H, H)$, $H^- \sim (H^-, H)$ and $H^+ \sim (H, H^+)$ are Markovian splitting subspaces, but they are not in general minimal.

Given an arbitrary splitting subspace $X \sim (S, \bar{S})$, how do we find a minimal one contained in it?

LEMMA 3.1. *Let $X \sim (S, \bar{S})$ and $X_0 \sim (S_0, \bar{S}_0)$ be splitting subspaces. Then $X_0 \subset X$ if and only if $S_0 \subset S$ and $\bar{S}_0 \subset \bar{S}$.*

Proof. The if-part follows from (3.4) and the only-if part from (3.5). □

To obtain a minimal splitting subspace, then, we would need to reduce S and $\bar{S}$ as far as possible, while preserving the splitting geometry of Theorem 3.1. By Theorem 2.2, S and $\bar{S}$ intersect perpendicularly if and only if $\bar{S}^{\perp} \subset S$ or, equivalently, $S^{\perp} \subset \bar{S}$. Therefore, in order that $S \supset H^-$, $\bar{S} \supset H^+$, and S and $\bar{S}$ intersect perpendicularly, we must have

$$S \supset H^- \vee \bar{S}^{\perp}, \tag{3.8a}$$

$$\bar{S} \supset H^+ \vee S^{\perp}. \tag{3.8b}$$

We must therefore reduce S and $\bar{S}$ without violating these conditions. The following theorem describes one procedure to do this.

THEOREM 3.3. [29]. *Let $X \sim (S, \bar{S})$ be a splitting subspace. Set $\bar{S}_0 := H^+ \vee S^{\perp}$ and $S_0 := H^- \vee \bar{S}_0^{\perp}$. Then $X_0 \sim (S_0, \bar{S}_0)$ is a minimal splitting subspace such that $X_0 \subset X$. If X is Markovian, then so is X_0.*

Proof. By definition, $S_0 \supset H^-$, $\bar{S}_0 \supset H^+$, and $\bar{S}_0^{\perp} \subset S_0$, i.e. S_0 and $\bar{S}_0$ intersect perpendicularly (Theorem 2.2). Hence $X_0 \sim (S_0, \bar{S}_0)$ is a splitting subspace (Theorem

[3] A subspace M of H is *full range* if the closed linear hull of the shifted spaces $\{U_t M ; t \in \mathbb{R}\}$ is all of H.

3.1). Also, $S^{\perp} \subset \bar{S}_0$, i.e. $\bar{S}_0^{\perp} \subset S$, and $H^- \subset S$, and therefore $S_0 \subset S$. Moreover, (3.8b) may be written $\bar{S}_0 \subset \bar{S}$. Consequently $X_0 \subset X$ (Lemma 3.1).

Next, we show that X_0 is minimal. To this end, suppose that $X_1 \sim (S_1, \bar{S}_1)$ is a splitting subspace such that $X_1 \subset X_0$. Then, by Lemma 3.1, $S_1 \subset S_0$ and $\bar{S}_1 \subset \bar{S}_0$. However, from the splitting geometry (3.8) we have

$$S_1 \supset H^- \vee \bar{S}_1^{\perp}, \tag{3.9a}$$

$$\bar{S}_1 \supset H^+ \vee S_1^{\perp}. \tag{3.9b}$$

Since $\bar{S}_1 \subset \bar{S}_0$, $\bar{S}_1^{\perp} \supset \bar{S}_0^{\perp}$, and therefore (3.9a) yields $S_1 \supset S_0$. Hence $S_1 = S_0$. Furthermore, $X_1 \subset X_0 \subset X$ implies that $S_1 \subset S$ (Lemma 3.1), i.e. $S_1^{\perp} \supset S^{\perp}$, which together with (3.9b) yields $\bar{S}_1 \supset \bar{S}_0$. Thus $\bar{S}_1 = \bar{S}_0$. Consequently $X_1 = X_0$, establishing the minimality of X_0.

It remains to shown that if X is Markovian then so is X_0. In view of Theorem 3.2, this amounts to showing that

$$U_t S_0 \subset S_0 \quad \text{for } t \leqq 0, \tag{3.10a}$$

$$U_t \bar{S}_0 \subset \bar{S}_0 \quad \text{for } t \geqq 0 \tag{3.10b}$$

follows from (3.7). It is well known and easy to show that if a subspace M is invariant under an operator T, i.e. $TM \subset M$, then the orthogonal complement $M^{\perp}$ is invariant under the adjoint T^*, i.e. $T^* M^{\perp} \subset M^{\perp}$. Then, noting that $U_t^* = U_{-t}$, we see that (3.7a) can be written $U_t S^{\perp} \subset S^{\perp}$ for $t \geqq 0$, which together with (3.1b) yields (3.10b). In the same way, (3.10b) and (3.1a) yields (3.10a). □

From this we see, as we could expect, that for minimality we must have equality in (3.8a) and in (3.8b).

COROLLARY 3.3. [28]. *A splitting subspace $X \sim (S, \bar{S})$ is minimal if and only if*

$$\bar{S} = H^+ \vee S^{\perp}, \tag{3.11a}$$

$$S = H^- \vee \bar{S}^{\perp}. \tag{3.11b}$$

Given S, (3.11a) is the smallest subspace $\bar{S}$ containing H^+ and intersecting S perpendicularly. Likewise, given $\bar{S}$, (3.11b) is the smallest subspace containing H^- and intersecting $\bar{S}^{\perp}$ perpendicularly. It follows from Theorem 3.3 that these minimality conditions remain the same if we restrict our analysis to Markovian splitting subspaces. Therefore the properties "minimal" and "Markovian" can be studied separately.

COROLLARY 3.4. [29]. *A Markovian splitting subspace which contains no Markovian splitting subspace as a proper subspace is a minimal splitting subspace.*

The existence of minimal splitting subspaces, finally, is insured by Theorem 3.3.

COROLLARY 3.5. *Each (Markovian) splitting subspace contains a minimal (Markovian) splitting subspace.*

Applying Theorem 3.3 to the Markovian splitting subspaces $H^- \sim (H^-, H)$ and $H^+ \sim (H, H^+)$, we obtain[4] the minimal Markovian splitting subspaces $H^{+/-} \sim (H^-, H^+ \vee (H^-)^{\perp})$ and $H^{-/+} \sim (H^- \vee (H^+)^{\perp}, H^+)$. Introducing

$$N^- := H^- \cap (H^+)^{\perp}, \tag{3.12a}$$

$$N^+ := H^+ \cap (H^-)^{\perp}, \tag{3.12b}$$

we may write[4] $H^{-/+} \sim (H^-, (N^-)^{\perp})$ and $H^{+/-} \sim ((N^+)^{\perp}, H^+)$. Moreover, by Corollary

[4] Recall that, for any subspaces A and B, $(A \vee B)^{\perp} = A^{\perp} \cap B^{\perp}$.

3.2, $H^- = H^{+/-} \oplus N^-$ and $H^+ = H^{-/+} \oplus N^+$, i.e., in view of Lemma 2.1, we have

$$H^{+/-} = \bar{E}^{H^-} H^+, \tag{3.13a}$$

$$H^{-/+} = \bar{E}^{H^+} H^-. \tag{3.13b}$$

Consequently $H^{+/-}$ and $H^{-/+}$ are the forward and backward *predictor spaces.* From Proposition 2.3 we see that $H^{+/-}$ is the only minimal splitting subspace contained in H^-, and $H^{-/+}$ is the only one contained in H^+.

Hence we have identified two minimal splitting subspaces, but where do we look for the others? To answer this question, first note that, since $H^- = H^{+/-} \oplus N^-$ and $H^+ = H^{-/+} \oplus N^+$,

$$H = N^- \oplus H^\square \oplus N^+ \tag{3.14}$$

where $H^\square$ is the *frame space*

$$H^\square = H^{+/-} \vee H^{-/+}. \tag{3.15}$$

Since $(N^+)^\perp \supset H^-$ and $(N^-)^\perp \supset H^+$, $H^\square$ is a splitting subspace (Corollary 3.2), which is Markovian (Theorem 3.2), but in general nonminimal.

THEOREM 3.4. [26]. *The frame space $H^\square$ is the closed linear hull of all minimal splitting subspaces. If X is a minimal splitting subspace, then*

$$H^- \cap H^+ \subset X \subset H^\square. \tag{3.16}$$

Proof. Let $X \sim (S, \bar{S})$ be a minimal splitting subspace. Then it satisfies (3.11). But, $S \supset H^-$ and $\bar{S} \supset H^+$, or, equivalently, $S^\perp \subset (H^-)^\perp$ and $\bar{S}^\perp \subset (H^+)^\perp$, which together with (3.11) yields $\bar{S} \subset (N^-)^\perp$ and $S \subset (N^+)^\perp$. Hence $X \subset H^\square$. In view of (3.15), the minimal splitting subspaces span $H^\square$. The relation $H^- \cap H^+ \subset X$ follows from Proposition 2.2. □

Consequently, as far as minimal splitting subspace construction is concerned, only the frame space $H^\square$ is of interest; the spaces N^- and N^+ in the decomposition (3.14) may be discarded. This observation is of importance in many applications, such as, for example, smoothing [4]. The point here is that, whenever y has a rational spectral density, $H^\square$ is finite-dimensional while of course H is not.

In the event that the past space H^- and future space H^+ intersect perpendicularly, $H^\square = H^- \cap H^+$, and hence, by Theorem 3.4, there is a unique minimal splitting subspace. In the finite-dimensional case, this happens if and only if y has a rational spectral density the numerator polynomial of which is constant.

The special role played by the minimal splitting subspaces $H^{+/-}$ and $H^{-/+}$ is further underlined by the following result. In §§ 4 and 7 we shall identify $H^{+/-}$ and $H^{-/+}$ as the minimum and maximum elements in a certain lattice of splitting subspaces.

THEOREM 3.5. [30]. *Let $X \sim (S, \bar{S})$ be a splitting subspace. Then $\bar{E}^{H^-} X = H^{+/-}$ if and only if $X \perp N^-$ and $\bar{E}^{H^+} X = H^{-/+}$ if and only if $X \perp N^+$.*

Proof. Applying the projection E^{H^-} to $X = E^S \bar{S}$ (Corollary 3.1) and noting that $H^- \subset S$, we obtain $\bar{E}^{H^-} X = \bar{E}^{H^-} \bar{S}$. But $\bar{S} \supset H^+$, and hence $\bar{E}^{H^-} X \supset H^{+/-}$. Conversely, suppose that $\xi \in X$. Then, since $H^- = H^{+/-} \oplus N^-$, $E^{H^-} \xi = E^{H^{+/-}} \xi + E^{N^-} \xi$. But, since $X \perp N^-$, the last term is zero, and consequently $E^{H^-} \xi \in H^{+/-}$. Hence $\bar{E}^{H^-} X \subset H^{+/-}$. This establishes the first part. The second follows by symmetry. □

4. Observability, constructibility, and minimality. Let X be a splitting subspace, and consider the orthogonal decomposition

$$X = \bar{E}^X H^+ \oplus [X \cap (H^+)^\perp] \tag{4.1}$$

given by Lemma 2.1. An element in the subspace $X \cap (H^+)^\perp$ cannot be distinguished from zero by observing the future $\{y(t);\, t \geqq 0\}$ and is therefore called *unobservable*, in analogy with deterministic systems theory [22, p. 52]. The splitting subspace X is said to be *observable* if the unobservable subspace is trivial, i.e. $X \cap (H^+)^\perp = 0$. Likewise,

$$X = \bar{E}^X H^- \oplus [X \cap (H^-)^\perp] \tag{4.2}$$

and we call X *constructible* if the unconstructible subspace $X \cap (H^-)^\perp = 0$.

The above definitions of observability and constructibility, introduced by Ruckebusch in [42], are in complete agreement with the corresponding concepts in deterministic systems theory. To illustrate this point, let us consider the finite-dimensional stochastic system (1.7), which can be solved to yield

$$y(t) = C e^{At} x(0) + \int_0^t C e^{A(t-\sigma)} B \, du(\sigma). \tag{4.3}$$

Now, X is observable if and only if $\bar{E}^X H^+ = X$, i.e.

$$x_k(0) \in \overline{\text{span}}\, \{\hat{y}_i(t);\, t \geqq 0,\, i = 1, 2, \cdots, m\} \tag{4.4}$$

for $k = 1, 2, \cdots, n$, where $\hat{y}_k(t) := E^X y_k(t)$. For $t \geqq 0$, $\hat{y}(t) = C e^{At} x(0)$, since then the components of the second term in (4.3) are orthogonal to X. Therefore $\{\hat{y}(t);\, t \geqq 0\}$ is the output of the linear dynamic system

$$\begin{aligned} \dot{z} &= Az, \qquad z(0) = x(0), \\ \hat{y} &= Cz, \qquad t \geqq 0. \end{aligned} \tag{4.5}$$

The question of observability of X is thus reduced to determining if $x(0)$ can be solved in terms of $\{\hat{y}(t);\, t \geqq 0\}$ which happens if and only if (4.5) is observable in the usual sense of deterministic systems theory [22]. Similarly, X is constructible if and only if $x(0)$ can be solved in terms of $\{\hat{y}(t);\, t \leqq 0\}$. But, from the backward system (1.9), we see that $\{\hat{y}(t);\, t \leqq 0\}$ is the output of

$$\begin{aligned} \dot{\bar{z}} &= -\bar{A}\bar{z}, \qquad \bar{z}(0) = x(0), \\ \hat{y} &= C\bar{z}, \qquad t \leqq 0 \end{aligned} \tag{4.6}$$

and therefore X is constructible if and only if (4.6) is.

In the general setting, observability and constructibility can be characterized as follows.

THEOREM 4.1. [28]. *Let $X \sim (S, \bar{S})$ be a splitting subspace. Then X is observable if and only if*

$$\bar{S} = H^+ \vee S^\perp \tag{4.7}$$

and constructible if and only if

$$S = H^- \vee \bar{S}^\perp. \tag{4.8}$$

Proof. The observability condition $X \cap (H^+)^\perp = 0$ is equivalent to $X^\perp \vee H^+ = H$, which, in view of Corollary 3.2, can be written $(S^\perp \oplus \bar{S}^\perp) \vee H^+ = H$. Since $H^+ \subset \bar{S} \perp \bar{S}^\perp$, this is equivalent to $(H^+ \vee S^\perp) \oplus \bar{S}^\perp = H$, which is the same as (4.7). The proof of the constructibility part is analogous. □

The following result, first presented in [42] in a somewhat different formulation, is an immediate consequence of Corollary 3.3 and Theorem 4.1.

COROLLARY 4.1. (Ruckebusch). *A splitting subspace is minimal if and only if it is both observable and constructible.*

This statement may at first sight seem analogous to the central result in classical deterministic realization theory that a realization is minimal if and only if it is both observable and reachable. However, as we shall see below, it is in fact of quite a different nature, involving both the forward and the backward realization. This was indicated in § 1. In terms of the discussion there, Corollary 4.1 states that X is minimal if and only if conditions (b) and (b)′ both hold.

Defining the *observability map* $\mathcal{O}: X \to H^+$ to be $\mathcal{O}\xi = E^{H^+}\xi$, the decomposition (4.1) of X into an observable and an unobservable subspace is seen to be identical to the well-known relation

$$X = \overline{\text{range } \mathcal{O}^*} \oplus \ker \mathcal{O} \tag{4.9}$$

[48, p. 205], where $\mathcal{O}^*: H^+ \to X$ is the adjoint operator $\mathcal{O}^*\lambda = E^X\lambda$, and ker denotes null space. Consequently X is observable if and only if $\mathcal{O}$ is one-one or, equivalently, $\mathcal{O}^*$ maps onto a dense subset of X. The splitting subspace X is said to be *exactly observable* if $\mathcal{O}^*$ is onto.

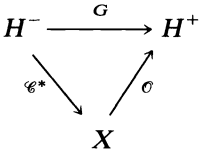
Similarly (4.2) can be written

$$X = \overline{\text{range } \mathscr{C}^*} \oplus \ker \mathscr{C} \tag{4.10}$$

where $\mathscr{C}: X \to H^-$ is the *constructibility map* $\mathscr{C}\xi = E^{H^-}\xi$. The splitting subspace X is constructible if and only if $\mathscr{C}$ is one-one or, equivalently, $\mathscr{C}^*: H^- \to X$ maps densely onto; it is *exactly constructible* if $\mathscr{C}^*$ is onto.

According to Proposition 2.1 (vi), the splitting property $H^- \perp H^+ | X$ is equivalent to $G = \mathcal{O}\mathscr{C}^*$, where $G: H^- \to H^+$ is the map $G\lambda = E^{H^+}\lambda$. This can be described by the commutative diagram

(4.11)



Such a factorization is said to be *canonical* if the first map (here $\mathscr{C}^*$) has a range which is dense in X and the second map (here $\mathcal{O}$) is one-one. In view of Corollary 4.1, we can summarize this in

PROPOSITION 4.1. *Let $G: H^- \to H^+$ be the map $G\lambda = E^{H^+}\lambda$. Then a subspace X is a splitting subspace if and only if the diagram* (4.11) *commutes. This splitting subspace is minimal if and only if the factorization is canonical.*

A splitting subspace X is *exactly canonical* if it is both exactly observable and exactly constructible. These conditions are technical and do not occur in the minimality criteria. However, certain results are much easier to prove in the finite-dimensional case (Theorem 4.3 is a case in point), and the reason for this is that the attribute "exact" is redundant in this case. Thus the technical difficulties are due to the lack of exactness rather than to infinite dimensions. The following lemma, found in [43, p. 28], relates exact canonicity to G having a closed range.

LEMMA 4.1. (Ruckebusch). *If G has a closed range, then all minimal splitting subspaces are exactly canonical. If one splitting subspace is exactly canonical, the G has a closed range.*

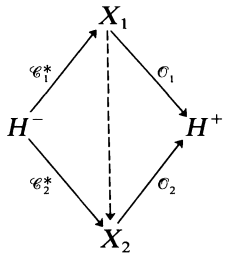
Proof. Recall that if a map has a closed range, then so does its adjoint [48, p. 205]; this will be used several times in the proof. Let X be a minimal splitting subspace. Then $G = \mathcal{O}\mathscr{C}^*$, and $\mathscr{C}^*H^-$ is dense in X (Proposition 4.1). Clearly $GH^- = \mathcal{O}\mathscr{C}^*H^- \subset \mathcal{O}X$. We want to show that, if GH^- is closed, then $GH^- = \mathcal{O}X$ so that $\mathcal{O}$,

and hence $\mathcal{O}^*$, has a closed range, i.e. X is exactly observable. To this end, let $\xi \in X$ be arbitrary. Then there is a sequence $\{\xi_n\}$ in $\mathcal{C}^*H^-$ such that $\xi_n \to \xi$. But $\mathcal{O}\xi_n \in GH^-$, and, since $\mathcal{O}$ is continuous, $\mathcal{O}\xi_n \to \mathcal{O}\xi \in GH^-$. Hence $\mathcal{O}X = GH^-$ as required. In the same way, we use the adjoint factorization $G^* = \mathcal{C}\mathcal{O}^*$, which is also canonical, to prove that X is exactly constructible. Conversely, assume that X is exactly canonical. Then $\mathcal{C}^*H^- = X$, and therefore, since $\mathcal{O}X$ is closed, $GH^- = \mathcal{O}\mathcal{C}^*H^-$ is closed. □

The following theorem ties together the geometric concept of minimality with that based on dimension.

THEOREM 4.2. *All minimal splitting subspaces have the same dimension.*

Proof. Let us first assume that G has a closed range. Let X_1 and X_2 be any two minimal splitting subspaces. Then there are two canonical factorizations $G = \mathcal{O}_1\mathcal{C}_1^* = \mathcal{O}_2\mathcal{C}_2^*$ (Proposition 4.1) which are in fact exactly canonical (Lemma 4.1). Consider the commutative diagram



in which $\mathcal{C}_1^*$ and $\mathcal{C}_2^*$ are onto and $\mathcal{O}_1$ and $\mathcal{O}_2$ are one-one. Then, using the argument of Kalman [22, pp. 256–258], we see that there is a bijective map from X_1 to X_2 (dotted arrow). Consequently X_1 and X_2 are isomorphic vector spaces, and therefore they have the same dimension. It remains to consider the case in which G does not have a closed range. But then, by Lemma 4.1, no minimal splitting subspace is exactly canonical, and consequently all are infinite-dimensional. Therefore, since H is a separable Hilbert space, all X have dimension $\aleph_0$. □

Next we shall give an alternative characterization of the class of minimal Markovian splitting subspaces which involves only the space S [or the space $\bar{S}$], and consequently, as we shall see below, only the forward [or the backward] realization. As a preliminary, first note that Theorem 4.1 has the following corollary.

COROLLARY 4.2. *The subspace X is an observable splitting subspace if and only if there is a subspace $S \supset H^-$ such that*

$$X = \bar{E}^S H^+. \tag{4.12}$$

It is a constructible splitting subspace if and only if there is a subspace $\bar{S} \supset H^+$ such that

$$X = \bar{E}^{\bar{S}} H^-. \tag{4.13}$$

The subspaces S and $\bar{S}$ are those of Theorem 3.1, *i.e.* $X \sim (S, \bar{S})$.

Proof. Suppose that $X \sim (S, \bar{S})$ is an observable splitting subspace. Then $X = E^S\bar{S}$ (Corollary 3.1), which together with the observability condition (4.7) yields (4.12). Conversely, suppose there is an $S \supset H^-$ such that (4.12) holds. Define $\bar{S} := H^+ \vee S^\perp$. Then S and $\bar{S}$ intersect perpendicularly (Theorem 2.2) and $X = E^S\bar{S}$. Hence $X \sim (S, \bar{S})$ is a splitting subspace (Corollary 3.1) which is observable (Theorem 4.2). The rest follows from the symmetric argument. □

There are now two representations for the class of minimal Markovian splitting subspaces, one based on (4.12), the other on (4.13). We shall only state the first, the second being the symmetric one. Phrased in terms of the finite-dimensional analysis

of § 1, Theorem 4.3 states that minimality is equivalent to conditions (b)+(c); this we shall see in § 7.

THEOREM 4.3. [31]. *Assume that N^- and N^+ are full range. Then X is a minimal Markovian splitting subspace if and only if*

$$X = \bar{E}^S H^+ \tag{4.14}$$

for some S satisfying the invariance condition (3.7a) *and*

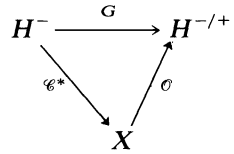
$$H^- \subset S \subset (N^+)^\perp. \tag{4.15}$$

The correspondence $X \leftrightarrow S$ is one-one, S being given by $S = H^- \vee X$.

Consequently, the class of minimal Markovian splitting subspaces forms a lattice, induced by the subspaces S: the greatest lower bound of X_1 and X_2 is the X corresponding to $S_1 \cap S_2$; the least upper bound corresponds to $S_1 \vee S_2$. Hence $X_1 < X_2$ if and only if $S_1 \subset S_2$. This lattice has the minimum element $H^{+/-}$, corresponding to $S = H^-$, and the maximum element $H^{-/+}$, corresponding to $S = (N^+)^\perp$.

To establish Theorem 4.3 it just remains to prove that *an observable splitting subspace $X \sim (S, \bar{S})$ is minimal if and only if $S \subset (N^+)^\perp$*. Then the rest follows from Corollary 4.2 and Theorem 3.2. The only-if part of this statement is immediate. In fact, since $S = H^- \vee X$ (Theorem 3.1), it follows from Theorem 3.4. The proof of the if-part, however, is more difficult. It can be found in [31]; also see Theorem 7.3 below. (Note that the proof in [28] is incorrect.)

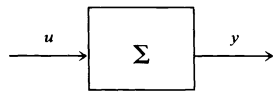
However, in the special case that the map G has a closed range, the proof is easy. Then G maps onto $H^{-/+}$. Moreover, since $X \subset S \perp N^+$, $\mathcal{O}X \subset H^{-/+}$ (Theorem 3.5). Consequently, we may without restriction replace (4.11) by



In this diagram, G is onto. Furthermore, since from the diagram $\mathcal{O}X \supset GH^-$, $\mathcal{O}$ is onto. By observability, $\mathcal{O}$ is one-one and therefore the inverse $\mathcal{O}^{-1}: H^{-/+} \to X$ is well defined and onto. Consequently, $\mathcal{C}^* = \mathcal{O}^{-1}G$ is onto, i.e. X is constructible; hence X is minimal (Corollary 4.1).

5. Reconciliation with systems theory. We wish to pinpoint the similarities and the differences between the state space constructions in deterministic and stochastic realization theory from an abstract systems-theoretical point of view. To this end, let us first briefly review some basic concepts of the standard state space construction in deterministic systems theory [22], [15].

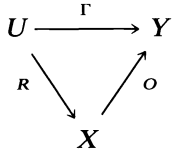
Consider an external description of a continuous-time, constant, linear dynamical system Σ, which we illustrate as a "black box"



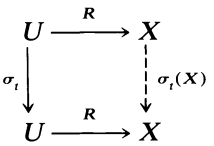
with input u and output y. Let U be a space of input functions u which are identically zero for $t > 0$, and let Y be a space of output functions y which are identically zero for $t < 0$. Let $\Gamma: U \to Y$ be the (linear) restricted input-output map defined by Σ. (Consequently, we apply the inputs up to time zero and observe the outputs from time

zero on.) The input space U is invariant under the operation $(\sigma_t u)(\tau) = u(\tau + t)$ of shifting the function a distance $t \geqq 0$ to the left, i.e. $\sigma_t U \subset U$ for $t \geqq 0$.

Two inputs $u_1, u_2 \in U$ are (*Nerode*) *equivalent* if the corresponding outputs Γu_1 and Γu_2 coincide, i.e. $u_1 - u_2 \in \ker \Gamma$. Then a minimal state space is obtained by forming the quotient space $X = U/\ker \Gamma$. If R is the projection onto the quotient space, there is an injective map O so that the diagram



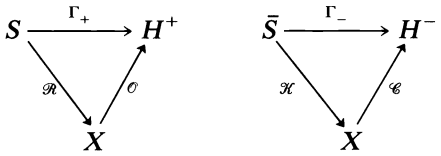
commutes [46, p. 23]. Hence we have a canonical factorization of Γ through the minimal state space X. (A noncanonical factorization will yield a nonminimal realization [22].) The semigroup $\{e^{At};\ t \geqq 0\}$, determining the dynamics of the realization, is then isomorphic to the family of maps making the diagrams



commute.

In the *stochastic* realization problem only the output process is given, and therefore the choice of input space is somewhat arbitrary. While the minimal state space in the deterministic theory is essentially unique, there are many solutions to the stochastic problem, each minimal Markovian splitting subspace $X \sim (S, \bar{S})$ giving rise to a minimal state space. As it turns out, each such state space is best described by two realizations, one evolving forward in time having S as input space and H^+ as output space, and another evolving backward with $\bar{S}$ as input space and H^- as output space. In § 6 we shall see that (under suitable conditions) there are two orthogonal increment processes u and $\bar{u}$ such that $S = H^-(du)$ and $\bar{S} = H^+(d\bar{u})$. These processes, called the *generating processes* of X, will be the input processes of respectively the forward and the backward realization of X.

THEOREM 5.1. *Let X be a subspace of H, and set $S := H^- \vee X$ and $\bar{S} := H^+ \vee X$. Then $X \sim (S, \bar{S})$ is a splitting subspace if and only if the diagrams*



commute, the maps being defined as $\Gamma_+\lambda = E^{H^+}\lambda$, $\Gamma_-\lambda = E^{H^-}\lambda$, $\mathscr{R}\lambda = E^X\lambda$, $\mathscr{K}\lambda = E^X\lambda$, $\mathscr{O}\lambda = E^{H^+}\lambda$, and $\mathscr{C}\lambda = E^{H^-}\lambda$ with domains as indicated. If one diagram commutes, then so does the other. The left factorization is canonical if and only if X is observable, the right one if and only if X is constructible, and both if and only if X is minimal. The maps $\mathscr{R}$ and $\mathscr{K}$ are always onto. Moreover,

(5.1a) $$X \supset S \ominus \ker \Gamma_+,$$

(5.1b) $$X \supset \bar{S} \ominus \ker \Gamma_-$$

with equality in (5.1a) *if and only if the left factorization is canonical and equality in* (5.1b) *if and only if the right factorization is canonical.*

Note that $\mathcal{O}$ and $\mathscr{C}$ are the observability and constructibility maps defined in § 4. Following standard terminology in systems theory, $\mathcal{R}$ is the *reachability map* and $\mathcal{K}$ the *controllability map.*

Proof. By Proposition 2.1 (vi), the left factorization $\Gamma_+ = \mathcal{O}\mathcal{R}$ is equivalent to $H^+ \perp S|X$ and the right one $\Gamma_- = \mathscr{C}\mathcal{K}$ to $H^- \perp \bar{S}|X$. But these conditional orthogonality conditions are both equivalent to $H^- \perp H^+|X$ (Proposition 2.1), the splitting property. Since $X \subset S$ and $X \subset \bar{S}$, $\mathcal{R}$ and $\mathcal{K}$ are obviously onto. Therefore the left factorization is canonical if and only if $\mathcal{O}$ is one-one, i.e. X is observable, and the right one is canonical if and only if $\mathscr{C}$ is one-one, i.e. X is constructible. Then, the minimality statement follows from Corollary 4.1. Since $\ker \Gamma_+ = S \cap (H^+)^\perp$, it follows from Lemma 2.1 that $S \ominus \ker \Gamma_+ = \bar{E}^S H^+$. But, in view of the splitting property (3.3), $\bar{E}^S H^+ = \bar{E}^X H^+$, which is the observable subspace of X; see (4.1). Hence (5.1a) holds, and there is equality if and only if $\bar{E}^X H^+ = X$, i.e. X is observable. The proof of the symmetric statement involving (5.1b) is analogous. □

Observing that $S \ominus \ker \Gamma_+$ and $\bar{S} \ominus \ker \Gamma_-$ are representations of the quotient spaces $S/\ker \Gamma_+$ and $\bar{S}/\ker \Gamma_-$ respectively, the analogy with the deterministic construction is apparent. Note, however, that in order for X to be a minimal splitting subspace, and hence correspond to a minimal state space, *both* diagrams need to be canonical. This is because the input space in the stochastic problem is not fixed but may change with X.

Some of the geometric results of §§ 3 and 4 can be inferred directly from the diagrams. Clearly we always have

$$\ker \mathcal{R} \subset \ker \Gamma_+ \tag{5.2}$$

with equality if and only if $\mathcal{O}$ is one-one. In fact, except for the elements in $\ker \mathcal{R}$ which are sent to the zero point in X and onto the zero point in H^+, there may be a subset of S whose image in X is nontrivial but then mapped onto the zero point of H^+. This happens if and only if $\mathcal{O}$ fails to be one-one. However, $\ker \mathcal{R} = S \ominus X = \bar{S}^\perp$ (Corollary 3.2), and therefore (5.2) can be written $\bar{S}^\perp \subset S \cap (H^+)^\perp$ or, equivalently, $\bar{S} \supset H^+ \vee S^\perp$, i.e. (3.8a). Equality yields the observability condition (4.7). Likewise, the corresponding relation between $\ker \mathcal{K}$ and $\ker \Gamma_-$ yields the constructibility condition (4.8).

Construction of semigroups for the stochastic problem requires that X is Markovian, in which case the input space S is invariant under the shift $\{U_t^*;\ t \geqq 0\}$ and $\bar{S}$ is invariant under $\{U_t;\ t \geqq 0\}$ (Theorem 3.2). These shifts play the role of $\{\sigma_t;\ t \geqq 0\}$ in the deterministic theory.

THEOREM 5.2. [30]. *Let $X \sim (S, \bar{S})$ be a Markovian splitting subspace. For each $t \geqq 0$, let $U_t(X): X \to X$ be the compressed shift $U_t(X)\xi = E^X U_t \xi$ and $U_t(X)^*: X \to X$ its adjoint $U_t(X)^*\xi = E^X U_{-t}\xi$. Then, for $t \geqq 0$, the diagrams*

$$\begin{array}{ccc} S & \xrightarrow{\mathcal{R}} & X \\ {\scriptstyle U_t^*}\downarrow & & \downarrow{\scriptstyle U_t(X)^*} \\ S & \xrightarrow{\mathcal{R}} & X \end{array} \qquad \begin{array}{ccc} \bar{S} & \xrightarrow{\mathcal{K}} & X \\ {\scriptstyle U_t}\downarrow & & \downarrow{\scriptstyle U_t(X)} \\ \bar{S} & \xrightarrow{\mathcal{K}} & X \end{array}$$

commute. Moreover, $\{U_t(X);\ t \geqq 0\}$ and $\{U_t(X)^*;\ t \geqq 0\}$ are strongly continuous contraction semigroups, and for each $\xi \in X$ and $t \geqq 0$,

$$E^S U_t \xi = U_t(X)\xi, \tag{5.3a}$$

$$E^{\bar{S}} U_{-t}\xi = U_t(X)^*\xi. \tag{5.3b}$$

If X is proper, both $U_t(X)$ and $U_t(X)^*$ tend strongly to zero as $t \to \infty$.

Proof. Let $t \geqq 0$ and $\lambda \in \bar{S}$. Then, since $\bar{S} = X \oplus S^{\perp}$ (Corollary 3.2),

$$E^X U_t \lambda = E^X U_t E^X \lambda + E^X U_t E^{S^{\perp}} \lambda. \tag{5.4}$$

Here the last term is zero, for $U_t S^{\perp} \subset S^{\perp} \perp X$ (Theorem 3.2 and Corollary 3.2). Therefore,

$$E^X U_t \lambda = E^X U_t E^X \lambda, \tag{5.5}$$

and consequently $\mathscr{K} U_t = U_t(X)\mathscr{K}$ as required. Also, since $S \perp \bar{S}|X$ and $U_t\lambda \in \bar{S}$, the left member of (5.5) can be exchanged for $E^S U_t \lambda$. Therefore, since $X \subset \bar{S}$, (5.3a) follows. The symmetric argument yields $\mathscr{R} U_t^* = U_t(X)^* \mathscr{R}$ and (5.3b). Since $U_t U_s = U_{t+s}$, it follows from (5.5) that $U_t(X)U_s(X) = U_{t+s}(X)$, i.e. $\{U_t(X);\, t \geqq 0\}$ is a semigroup, which is strongly continuous since $\{U_t\}$ is. Clearly $U_t(X)$ is a contraction, for U_t is unitary. If X is proper, $\bigcap_{t \leqq 0} U_t S = 0$ and hence, in view of (5.3a) and the identity $E^S U_t = U_t E^{U_t^* S}$, we get $\| U_t(X)\xi \| = \| E^{U_{-t} S} \xi \| \to 0$ as $t \to \infty$ proving the last statement of the theorem. The family $\{U_t(X)^*;\, t \geqq 0\}$ is merely the adjoint semigroup with the same properties. □

Following the pattern of this section, in §§ 8 and 9 we shall assign to each proper Markovian splitting subspace X two realizations with the systems-theoretical properties of Theorem 5.1, a forward one with input space S and semigroup $\{U_t(X)^*\}$ and a backward one with input space $\bar{S}$ and semigroup $\{U_t(X)\}$. Therefore we shall call $\{U_t(X)^*;\, t \geqq 0\}$ and $\{U_t(X);\, t \geqq 0\}$ the *forward* and *backward Markovian semigroups* of X respectively.

6. Generating processes. By representing the random variables as Wiener integrals we shall next derive functional models for the geometric results presented above.

To this end, let us first define a p-dimensional *Wiener process on the real line* $\mathbb{R}$ to be a real centered Gaussian vector process $\{u(t);\, t \in \mathbb{R}\}$ which has (almost surely) continuous sample functions and independent (and hence orthogonal) increments such that

$$E\{du(t)\, du(t)'\} = I\, dt. \tag{6.1}$$

Although we shall only be interested in the increments of u, it is convenient to set $u(0) = 0$. Defining $H(du)$ to be the Hilbert space generated by the components of $\{u(t) - u(s);\, t, s \in \mathbb{R}\}$, we have the orthogonal decomposition

$$H(du) = H^-(du) \oplus H^+(du) \tag{6.2}$$

where $H^-(du)$ and $H^+(du)$ are the subspaces corresponding respectively to the increments $\{u(t) - u(s);\, t, s \leqq 0\}$ and $\{u(t) - u(s);\, t, s \geqq 0\}$.

It is well known [38] that to any $\eta \in H(du)$ there is a function f in $\mathscr{L}_p^2(\mathbb{R})$, the space of p-dimensional *real* vector functions square-integrable on $\mathbb{R}$, such that

$$\eta = \sum_{i=1}^{p} \int_{-\infty}^{\infty} f_i(-t)\, du_i(t), \tag{6.3}$$

where the integral is defined in quadratic mean. We shall write (6.3) as

$$\eta = \int_{-\infty}^{\infty} f(-t)\, du(t) \tag{6.4}$$

i.e. we shall think of the function f as a row vector and the process u as a column vector; this convention will be maintained through the rest of the paper. Let $I_u: \mathscr{L}_p^2(\mathbb{R}) \to H(du)$ be the map defined by (6.4), i.e. $\eta = I_u f$. Then $\langle I_u f, I_u g \rangle = \int_{-\infty}^{\infty} f(t) g(t)'\, dt$, the inner product of f and g in $\mathscr{L}_p^2(\mathbb{R})$, i.e. I_u is an isometry. Since it is also onto, I_u is unitary.

It is not hard to see that the vector process

$$\hat{u}(i\omega) := \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-i\omega t}-1}{it}\, du(t) \tag{6.5}$$

defined on the imaginary axis $\mathbb{I}$, has much the same properties as u with[5]

$$E\{d\hat{u}(i\omega)\, d\hat{u}(i\omega)^*\} = \frac{1}{2\pi} I\, d\omega, \tag{6.6}$$

and therefore we can think of it as a vector *Wiener process on the imaginary axis.* Also, it is well known [38, p. 147] that, for each $f \in \mathscr{L}_p^2(\mathbb{R})$,

$$\int_{-\infty}^{\infty} f(-t)\, du(t) = \int_{-\infty}^{\infty} \hat{f}(i\omega)\, d\hat{u}(i\omega) \tag{6.7}$$

where $\omega \to \hat{f}(i\omega)$ is the Fourier transform

$$\hat{f}(i\omega) = \int_{-\infty}^{\infty} e^{-i\omega t} f(t)\, dt \tag{6.8}$$

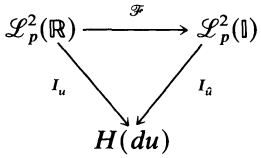
extended to all of $\mathscr{L}_p^2(\mathbb{R})$ in the usual manner [10], [38]. The space $\mathscr{L}_p^2(\mathbb{I})$ of all such $\hat{f}$ is a Hilbert space with inner product

$$(\hat{f}, \hat{g}) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(i\omega)\hat{g}(-i\omega)'\, d\omega,$$

and the map $\mathscr{F}: \mathscr{L}_p^2(\mathbb{R}) \to \mathscr{L}_p^2(\mathbb{I})$ defined by $\hat{f} = \mathscr{F}f$ is unitary. (We define $\hat{f}$ in the style of the Laplace transform in order to conform with usual nomenclature in systems theory.) In view of (6.7), the map $I_{\hat{u}}: \mathscr{L}_p^2(\mathbb{I}) \to H(du)$ defined by

$$I_{\hat{u}}\hat{f} = \int_{-\infty}^{\infty} \hat{f}(i\omega)\, d\hat{u}(i\omega) \tag{6.9}$$

is also unitary, and the diagram



commutes. Taking $f(-t)$ to be the indicator function of the interval $[0, t]$, (6.7) yields the spectral representation

$$u(t) = \int_{-\infty}^{\infty} \frac{e^{i\omega t}-1}{i\omega}\, d\hat{u}(i\omega). \tag{6.10}$$

If we let f vary over all functions in $\mathscr{L}_p^2(\mathbb{R})$ which vanish on the negative [positive] real axis, (6.4) generates $H^-(du)$ $[H^+(du)]$. This motivates the introduction of the *Hardy spaces* $\mathscr{H}_p^2$ and $\bar{\mathscr{H}}_p^2$. Let $\mathscr{H}_p^2$ $[\bar{\mathscr{H}}_p^2]$ be the subspace of those $\hat{f} \in \mathscr{L}_p^2(\mathbb{I})$ for which $f := \mathscr{F}^{-1}\hat{f}$ vanishes on the negative [positive] real line. Then, $H^-(du) = I_{\hat{u}}\mathscr{H}_p^2$ and $H^+(du) = I_{\hat{u}}\bar{\mathscr{H}}_p^2$, i.e.

$$\mathscr{L}_p^2(\mathbb{I}) = \mathscr{H}_p^2 \oplus \bar{\mathscr{H}}_p^2 \tag{6.11}$$

[5] Here * denotes conjugate transpose.

is the isomorphic image of (6.2) under $I_{\hat{u}}^{-1}$. Clearly $\bar{\mathscr{H}}_p^2 = \{i\omega \to g(-i\omega) | g \in \mathscr{H}_p^2\}$, and therefore $\bar{\mathscr{H}}_p^2$ is sometimes called the *conjugate* Hardy space. The Hardy spaces $\mathscr{H}_p^2$ and $\bar{\mathscr{H}}_p^2$ can also be defined as *bona fide* function spaces of functions analytic in the open right and left complex half planes respectively, the limits of which as the imaginary axis is approached perpendicularly are the elements of $\mathscr{H}_p^2$ and $\bar{\mathscr{H}}_p^2$ as defined above [14], [18], [45]. Therefore the functions of $\mathscr{H}_p^2$ will sometimes be called *analytic* and those of $\bar{\mathscr{H}}_p^2$ *coanalytic*. In the same way we define $\mathscr{H}_{p\times p}^\infty$ to be the space of all $p \times p$-matrix-valued functions bounded and analytic in the open right half plane, or, alternatively, as the corresponding subspace of $\mathscr{L}_{p\times p}^\infty(\mathbb{I})$. Here we shall think of these functions as defined on the imaginary axis, but it is useful to keep the other interpretation in mind.

Our program now is to assign to each proper Markovian splitting subspace $X \sim (S, \bar{S})$ a pair $(u, \bar{u})$ of Wiener processes on the real line such that $H^-(du) = S$, $H^+(d\bar{u}) = \bar{S}$, and $H(du) = H(d\bar{u}) = H$. Through the isomorphisms $I_{\hat{u}}$ and $I_{\hat{\bar{u}}}$ we shall then transform the geometry of §§ 2–5 to the Hardy space $\mathscr{H}_p^2$ in which the appropriate mappings take a particularly simple form.

Recall that the given m-dimensional process $\{y(t); t \in \mathbb{R}\}$ is stationary, Gaussian, mean-square continuous, and centered. From now on, we shall also assume that y is *purely nondeterministic* in the (strong) sense that both $(H^-)^\perp$ and $(H^+)^\perp$ are full range. Then, y has a spectral representation

$$y(t) = \int_{-\infty}^{\infty} e^{i\omega t}\, d\hat{y}(i\omega) \tag{6.12}$$

where $\{\hat{y}(s); s \in \mathbb{I}\}$ is an independent-increment process such that

$$E\{d\hat{y}(i\omega)\, d\hat{y}(i\omega)^*\} = \frac{1}{2\pi}\Phi(i\omega)\, d\omega. \tag{6.13}$$

Here the $m \times m$-matrix function Φ is the *spectral density* of y, and $\hat{y}$ is given by

$$\hat{y}(i\omega) = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-i\omega t} - 1}{it}\, y(t)\, dt, \tag{6.14}$$

where the limit is in quadratic mean [9].

Since y is purely nondeterministic, $\Phi(i\omega)$ has a constant rank $p \leqq m$ (for almost all ω), and it admits a factorization

$$W(s)\, W(-s)' = \Phi(s) \tag{6.15}$$

where W is an $m \times p$-matrix function whose rows belong to $\mathscr{L}_p^2(\mathbb{I})$ [38, p. 114]. There are many such W, and we call them *full-rank spectral factors*. More specifically, the condition that $(H^-)^\perp[(H^+)^\perp]$ is full range implies that there are W with rows in $\mathscr{H}_p^2(\bar{\mathscr{H}}_p^2)$.

To each full-rank spectral factor W we associate a unique p-dimensional Wiener process (on the real line), namely (6.10) with

$$d\hat{u} = W^{-L}\, d\hat{y} \tag{6.16}$$

where W^{-L} is any left inverse of W, i.e. a $p \times m$-matrix function such that $W^{-L}W = I$. Although, in general, W has more than one left inverse, u is uniquely determined by (6.16). In fact, let W^{-L} and $W^{-L} + \Delta$ be two left inverses. Then $\Delta W = 0$, and consequently, because of (6.15), $\Delta\Phi\Delta^*\, d\omega = 0$, and therefore the uniqueness is established. For example, we may take $W^{-L} = (W'W)^{-1}W'$. Despite the fact that $WW^{-L} \neq I$ in general, $WW^{-L}\, d\hat{y} = d\hat{y}$, i.e.

$$d\hat{y} = W\, d\hat{u}. \tag{6.17}$$

To see this, form $E\{(I - WW^{-L})\, d\hat{y}\, d\hat{y}^*(I - WW^{-L})^*\}$, which, in view of (6.13) and (6.15), equals zero.

The class $\mathcal{U}$ of Wiener processes u defined in this way is characterized as follows.

PROPOSITION 6.1. [28]. *Let u be a vector Wiener process defined on the real line. Then $u \in \mathcal{U}$ if and only if $H(du) = H$. In this case, for each $t \in \mathbb{R}$,*

$$U_t I_{\hat{u}} = I_{\hat{u}} \chi_t \tag{6.18}$$

where $\chi_t : \mathcal{L}_p^2(\mathbb{I}) \to \mathcal{L}_p^2(\mathbb{I})$ is multiplication by $e^{i\omega t}$.

Proof. First suppose that $u \in \mathcal{U}$. Then, by (6.17),

$$y(t) = \int e^{i\omega t} W\, d\hat{u} \tag{6.19}$$

and consequently, in view of (6.7), $y_k(t) \in H(du)$ for $k = 1, 2, \cdots, m$. Hence $H \subset H(du)$. On the other hand, it follows from (6.10), (6.16) and (6.14) that $H(du) \subset H$, and therefore $H(du) = H$. Conversely, assume that $H(du) = H$. Then, by (6.7), there exists a matrix function W with rows in $\mathcal{L}_p^2(\mathbb{I})$ such that $y(0) = \int W\, d\hat{u}$. From this it is seen that W is a spectral factor and that $d\hat{y} = W\, d\hat{u}$, but it remains to show that W is full rank. However, since $H(du) \subset H$, for each $t \in \mathbb{R}$, there is a matrix function G_t such that $u(t) = \int G_t\, d\hat{y}$; i.e. $u(t) = \int G_t W\, d\hat{u}$. Hence, by (6.10), $G_t W = (e^{i\omega t} - 1)/i\omega I$ for all $t \in \mathbb{R}$. Consequently, W must have full rank. □

COROLLARY 6.1. *Let $u \in \mathcal{U}$. Then, for $k = 1, 2, \cdots, p$,*

$$U_t[u_k(\tau) - u_k(\sigma)] = u_k(\tau + t) - u_k(\sigma + t) \tag{6.20}$$

and consequently $U_t H^-(du)$ is the subspace generated by the components of $\{u(\tau) - u(\sigma) | \tau, \sigma \leqq t\}$.

Proof. In view of (6.10),

$$u(\tau) - u(\sigma) = \int_{-\infty}^{\infty} \frac{e^{i\omega\tau} - e^{i\omega\sigma}}{i\omega}\, d\hat{u} \tag{6.21}$$

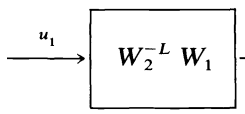
and therefore (6.20) follows from (6.18). □

How are the processes in $\mathcal{U}$ related to each other? It is immediately clear that, if u_1 and u_2 correspond to the spectral factors W_1 and W_2 respectively, then

$$d\hat{u}_2 = W_2^{-L} W_1\, d\hat{u}_1. \tag{6.22}$$

The $p \times p$-matrix function $W_2^{-L} W_1$ is uniquely defined (independent of the choice of left inverse), because, just as above, $d\hat{u}_2 = P_1\, d\hat{u}_1$ and $d\hat{u}_2 = P_2\, d\hat{u}_1$ imply that $(P_1 - P_2)(P_1 - P_2)^* = 0$, i.e. $P_1 = P_2$. Also, it follows from (6.15) that the values of $W_2^{-L} W_1$ on $\mathbb{I}$ are unitary matrices. Therefore we can think of u_2 being obtained by passing u_1 through a filter with the transfer function $W_2^{-L} W_1$:

(6.23a)



In engineering language, such an object is called an *all-pass filter.* Moreover, for any $f \in \mathcal{L}_p^2(\mathbb{I})$, $I_{\hat{u}_2} f = I_{\hat{u}_1} f W_2^{-L} W_1$, i.e.

$$I_{\hat{u}_1}^{-1} I_{\hat{u}_2} = M_{W_2^{-L} W_1} \tag{6.23b}$$

where, here as in the sequel, $M_Q : \mathcal{L}_2(\mathbb{I}) \to \mathcal{L}_2(\mathbb{I})$ denotes multiplication from the right by Q, i.e. $M_Q f = fQ$.

Clearly, if W is a full-rank spectral factor, then so is WT for any constant unitary $p \times p$ matrix T. However, the corresponding processes in $\mathcal{U}$ are related to each other by a trivial coordinate transformation, and therefore we shall regard them as equivalent. The transformation (6.23) is interesting only if $W_2^{-L}W_1$ is nonconstant.

A matrix function $Q \in \mathcal{H}_{p\times p}^{\infty}$ with the property that $Q(i\omega)$ are unitary matrices for almost all ω is said to be *inner* [14], [18], [45]. In particular, $W_2^{-L}W_1$ is inner if it belongs to $\mathcal{H}_{p\times p}^{\infty}$. The following lemma, which is a corollary of a famous theorem by Beurling, generalized to vector functions by Lax [14], [18], [45], states that the all-pass filter (6.23a) is *causal* if and only if $W_2^{-L}W_1$ is inner.

LEMMA 6.1. [28]. *Let u_1 and u_2 be two processes in $\mathcal{U}$, and let W_1 and W_2 be the corresponding spectral factors. Then $W_2^{-L}W_1$ is inner if and only if*

$$H^-(du_2) \subset H^-(du_1). \tag{6.24}$$

Proof. Set $\mathcal{Z} := I_{\hat{u}_1}^{-1} H^-(du_2)$. Since $U_t H^-(du_2) \subset H^-(du_2)$ for $t \leqq 0$ (Corollary 6.1), $\chi_t \mathcal{Z} \subset \mathcal{Z}$ for $t \leqq 0$ (Proposition 6.1). A subspace $\mathcal{Z}$ with this property is called *invariant*. Moreover, since $H^-(du_2)$ is full range, then so is $\mathcal{Z}$, in the sense that the closed linear hull of $\{\chi_t \mathcal{Z};\ t \in \mathbb{R}\}$ is all of $\mathcal{L}_p^2(\mathbb{I})$. Since $I_{\hat{u}_1}^{-1} H^-(du_1) = \mathcal{H}_p^2$, (6.24) is equivalent to $\mathcal{Z} \subset \mathcal{H}_p^2$. Now, by the Beurling–Lax theorem, the invariant full range subspaces of $\mathcal{H}_p^2$ are precisely the subspaces of the form $\mathcal{H}_p^2 Q$ where Q is inner. But, in view of (6.23b) $\mathcal{Z} = \mathcal{H}_p^2 W_2^{-L} W_1$. Therefore, if $W_2^{-L}W_1$ is inner, (6.24) holds. Conversely, if (6.24) holds, $W_2^{-L}W_1$ must be inner. In fact, if $\mathcal{H}_p^2 Q_1 = \mathcal{H}_p^2 Q_2$ where both Q_1 and Q_2 take unitary values in $\mathbb{I}$, then $Q_1 = TQ_2$ where T is a constant unitary matrix [18]. □

Referring to the alternative definitions of $\mathcal{H}_p^2$ and $\bar{\mathcal{H}}_p^2$, a full-rank spectral factor with rows in $\mathcal{H}_p^2$ will be called *analytic*, and one with rows in $\bar{\mathcal{H}}_p^2$ *coanalytic*. Let $\mathcal{U}^-$ and $\mathcal{U}^+$ respectively be the corresponding subclasses of $\mathcal{U}$.

LEMMA 6.2. [28]. *Let $u \in \mathcal{U}$. Then, $u \in \mathcal{U}^-$ if and only if $H^-(du) \supset H^-$, and $u \in \mathcal{U}^+$ if and only if $H^+(du) \supset H^+$.*

Proof. By definition, $u \in \mathcal{U}^-$ is equivalent to $W_k \in \mathcal{H}_p^2$ for $k = 1, 2, \cdots, m$, where W_k is the kth row of the spectral factor W corresponding to u. Under the isomorphism $I_{\hat{u}}$ this is equivalent to

$$y_k(0) \in H^-(du) \quad \text{for } k = 1, 2, \cdots, m. \tag{6.25}$$

For this to hold it is clearly sufficient that $H^- \subset H^-(du)$. Conversely, suppose that (6.25) holds. Then, in view of Corollary 6.1, $y_k(t) \subset H^-(du)$ for $t \leqq 0$ and $k = 1, 2, \cdots, m$. This implies that $H^- \subset H^-(du)$. The proof of the symmetric statement is analogous. □

We are now in a position to tie together the results of this section with the geometric theory presented in the beginning of the paper. The link is provided by the following theorem.

THEOREM 6.1. [28]. (i) *Let S be a subspace such that $S \supset H^-$ and $S^\perp$ is full range. Then*

$$U_t S \subset S \quad \textit{for } t \leqq 0 \tag{6.26}$$

if and only if there are an analytic full-rank spectral factor W and a corresponding $u \in \mathcal{U}^-$ such that

$$S = H^-(du). \tag{6.27}$$

The spectral factor W and the process u are unique modulo multiplication from the right (respectively the left) by the same constant unitary matrix.

(ii) *Let $\bar{S}$ be a subspace such that $\bar{S} \supset H^+$ and $\bar{S}^\perp$ is full range. Then*

$$U_t\bar{S} \subset \bar{S} \quad \text{for } t \geqq 0 \tag{6.28}$$

if and only if there are a coanalytic full-rank spectral factor $\bar{W}$ and a corresponding $\bar{u} \in \mathcal{U}^+$ such that

$$\bar{S} = H^+(d\bar{u}). \tag{6.29}$$

Here $\bar{W}$ and $\bar{u}$ enjoy the same uniqueness properties as in (i).

Proof. (i) Let $v \in \mathcal{U}$ be arbitrary, and let V be the corresponding spectral factor. (Since there are full-rank spectral factors [38, p. 114], $\mathcal{U}$ is nonempty.) Set $\mathcal{Z} := I_{\hat{v}}^{-1}S$. In view of Proposition 6.1, (6.26) is equivalent to

$$\chi_t\mathcal{Z} \subset \mathcal{Z} \quad \text{for } t \leqq 0. \tag{6.30}$$

Moreover, since both S and $S^\perp$ are full range, then so are $\mathcal{Z}$ and $\mathcal{Z}^\perp$ (in the sense of the proof of Lemma 6.1). Therefore, there is a $p \times p$ matrix function $i\omega \to Q(i\omega)$ taking values which are unitary matrices such that $\mathcal{Z} = \mathcal{H}_p^2 Q$ [18]. Define $W := VQ^{-1}$. Clearly W is a full-rank spectral factor; let $u \in \mathcal{U}$ be the corresponding element in $\mathcal{U}$. The function Q is unique up to multiplication by a constant regular matrix [18] and hence the same is true for W and u. Then, by (6.23b), $I_{\hat{v}}^{-1}I_{\hat{u}} = M_Q$, i.e. $I_{\hat{v}}M_Q = I_{\hat{u}}$. Therefore, since $S = I_{\hat{v}}\mathcal{Z}$, we have $S = I_{\hat{v}}M_Q\mathcal{H}_p^2 = I_{\hat{u}}\mathcal{H}_p^2 = H^-(du)$. Since $S \supset H^-$, Lemma 6.2 implies that $u \in \mathcal{U}^-$. This concludes the if-part of (i); the only-if part follows immediately from Corollary 6.1. The proof of (ii) is analogous. □

Consequently, each proper Markovian splitting subspace $X \sim (S, \bar{S})$ is completely determined by a pair $(u, \bar{u})$ of Wiener processes, one in $\mathcal{U}^-$ and the other in $\mathcal{U}^+$; in fact

$$X = H^-(du) \cap H^+(d\bar{u}). \tag{6.31}$$

The processes are called respectively the *forward and backward generating processes* of X.

7. Hardy space representation of Markovian splitting subspaces. The goal of any description of dynamic phenomena is to obtain differential (or difference) equation representations of the relevant state variables, such as (1.7) and (1.9). To achieve this goal, in this section we go through an intermediate step in which the basic objects representing the dynamics are pairs of transfer functions $(W, \bar{W})$, one causal and the other anticausal. We shall arrive at a concrete coordinate-free state-space description in terms of analytic functions which can be computed from W and $\bar{W}$. The appropriate mathematical setting for representing causal and anticausal transfer functions is the theory of Hardy spaces. Notice that, while in the finite-dimensional setting differential equations can be obtained through straightforward algebraic calculations involving the appropriate analytic functions (§ 8), the general situation requires considerably more care (§ 9). The advantage of working with transfer function descriptions, i.e. the Hardy space setting, is that very detailed structural information about the state-space representations is obtained without having to introduce unnecessary finite-dimensionality conditions from the beginning.

Our next task is therefore to transfer the splitting subspace geometry to the Hardy space setting. To this end we need the following lemma.

LEMMA 7.1. [29]. *Let $u_1, u_2 \in \mathcal{U}$ be such that $H^-(du_1) \vee H^-(du_2) = H$, and let W_1 and W_2 be the corresponding spectral factors. Then $H^-(du_1)$ and $H^+(du_2)$ intersect perpendicularly if and only if $W_2^{-L}W_1$ is an inner function.*

Proof. By Theorem 2.2 (ii), $H^-(du_1)$ and $H^+(du_2)$ intersect perpendicularly if and only if $H^-(du_2) \subset H^-(du_1)$. But this is equivalent to $W_2^{-L}W_1$ being *inner* (Lemma 6.1). □

As we have seen, each proper Markovian splitting subspace $X \sim (S, \bar{S})$ is characterized by a pair of generating processes $(u, \bar{u})$; we have $S = H^-(du)$ and $\bar{S} = H^+(d\bar{u})$. Let $(W, \bar{W})$ be the corresponding pair of spectral factors. The condition $S \supset H^-$ is equivalent to W being analytic, $\bar{S} \supset H^+$ to $\bar{W}$ being coanalytic (Lemma 6.2), and the perpendicular intersection of S and $\bar{S}$ to

$$K := \bar{W}^{-L}W \tag{7.1}$$

being inner (Lemma 7.1). The function K is called the *structural function* of X and will play a major role in what follows. It corresponds to the *scattering matrix* in Lax-Phillips scattering theory [23].

Now, by Corollary 3.2, we have $X = S \ominus \bar{S}^{\perp}$, i.e. $X = H^-(du) \ominus H^-(d\bar{u})$, and therefore, in view of (6.23b), $I_{\hat{u}}X = \mathscr{H}_p^2 \ominus (\mathscr{H}_p^2 K)$. (Remember that $I_{\hat{u}}$ is unitary, and therefore orthogonality is preserved.) Define $\mathscr{H}(J) := \mathscr{H}_p^2 \ominus (\mathscr{H}_p^2 J)$ for any inner function J. Then

$$X = \int_{-\infty}^{\infty} \mathscr{H}(K)\, d\hat{u}. \tag{7.2}$$

Together with $d\hat{u} = W^{-L}\, d\hat{y}$ this yields a representation in terms of y. Consequently, we have established the following Hardy space version of Theorem 3.1.

THEOREM 7.1. [28]. *A subspace X is a proper Markovian splitting subspace if and only if*

$$X = \int_{-\infty}^{\infty} \mathscr{H}(\bar{W}^{-L}W)\, W^{-L}\, d\hat{y} \tag{7.3}$$

for some pair $(W, \bar{W})$ of full-rank spectral factors such that W is analytic, $\bar{W}$ is coanalytic, and $K := \bar{W}^{-L}W$ is inner. The correspondence $X \leftrightarrow (W, \bar{W})$ is one-one (modulo multiplication from the right by constant unitary matrices).

Instead applying $I_{\hat{u}}$ to $X = \bar{S} \ominus S^{\perp}$ (Corollary 3.2), we have the symmetric representation

$$X = \int_{-\infty}^{\infty} \bar{\mathscr{H}}(K^*)\, d\hat{\bar{u}} \tag{7.4}$$

where $\bar{\mathscr{H}}(J) := \bar{\mathscr{H}}_p^2 \ominus \bar{\mathscr{H}}_p^2 J$ for each conjugate inner function J. (A function J is *conjugate inner* if its inverse J^* is inner.) Consequently, since $d\hat{\bar{u}} = \bar{W}^{-L}d\hat{y}$, we can replace (7.3) by

$$X = \int_{-\infty}^{\infty} \bar{\mathscr{H}}(W^{-L}\bar{W})\, \bar{W}^{-L}\, d\hat{y} \tag{7.5}$$

in Theorem 7.1.

Which pairs of spectral factors $(W, \bar{W})$ correspond to *minimal* splitting subspaces? To answer this question we need to take a closer look at the classes of analytic and coanalytic full-rank spectral factors.

By assumption, H^- satisfies the conditions of Theorem 6.1 (i), and hence there is a $u_- \in \mathscr{U}^-$ such that

$$H^-(du_-) = H^-. \tag{7.6}$$

This is the (*forward*) *innovation process* of y. Let W_- denote the corresponding analytic

spectral factor. Since $d\hat{y} = W_- \, d\hat{u}_-$, (6.12) yields $I_{\hat{u}_-}^{-1} a y(t) = \chi_t a W_-$ for any row vector a in $\mathbb{R}^m$, and therefore applying $I_{\hat{u}_-}^{-1}$ to (7.6) we obtain

$$\overline{\text{span}}\,\{\chi_t a W_-;\, t \leqq 0,\, a \in \mathbb{R}^m\} = \mathscr{H}_p^2. \tag{7.7}$$

Such a function W_- is called *outer* [14], [18]. There is only one spectral factor with this property. Consequently, we shall call W_- the *outer* (or *minimum phase*) *spectral factor.*

All other analytic full-rank spectral factors have the property $H^-(du) \supset H^-$, where u is the corresponding process in $\mathscr{U}^-$ (Lemma 6.2), i.e. $H^-(du) \supset H^-(du_-)$. Therefore, $Q := W_-^L W$ is inner (Lemma 6.1) so that we have the inner-outer factorization

$$W = W_- Q. \tag{7.8}$$

To see this, use the following lemma.

LEMMA 7.2. *Let W_1 and W_2 be full-rank spectral factors. Then*

$$W_1 W_1^{-L} W_2 = W_2. \tag{7.9}$$

Proof. Form $(W_1 W_1^{-L} W_2 - W_2)(W_1 W_1^{-L} W_2 - W_2)^*$. Since $W_2 W_2^* = W_1 W_1^*$, we see that this is zero. □

Likewise, Theorem 6.1 (ii) implies that there is a $\bar{u}_+ \in \mathscr{U}^+$ such that

$$H^+(d\bar{u}_+) = H^+. \tag{7.10}$$

This is the *backward innovation process* of y. The corresponding spectral factor $\bar{W}_+$ satisfies

$$\overline{\text{span}}\,\{\chi_t a \bar{W}_+;\, t \geqq 0,\, a \in \mathbb{R}^m\} = \bar{\mathscr{H}}_p^2 \tag{7.11}$$

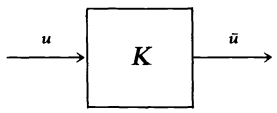
and is therefore called the *conjugate outer spectral factor.* In the same way as above, we show that any coanalytic full-rank spectral factor $\bar{W}$ can be written

$$\bar{W} = \bar{W}_+ \bar{Q} \tag{7.12}$$

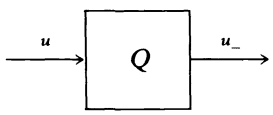
where $\bar{Q} := \bar{W}_+^{-L} \bar{W}$ is conjugate inner. The factorizations (7.8) and (7.12) are unique (modulo trivial coordinate transformations).

Consequently each proper Markovian splitting subspace is characterized by a triplet $(K, Q, \bar{Q}^*)$ consisting of the structural function K and the *forward and backward spectral inner factors* Q and $\bar{Q}^*$. These define three causal all-pass filters with the following inputs and outputs.

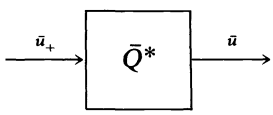
(7.13a)



(7.13b)



(7.13c)



We shall call $(K, Q, \bar{Q}^*)$ the *inner triplet* of X.

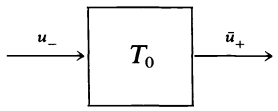
Our question on minimality can now be answered in terms of certain coprimeness conditions on these inner functions. Before turning to this, let us define a few concepts. If P_1 and P_2 are inner functions, then so is $P_3 := P_1P_2$; P_1 is a *left inner divisor* of P_3 ($P_1|_L P_3$ for short) and P_2 is a *right inner divisor* of P_3 ($P_2|_R P_3$): (These notations will also be used for conjugate inner functions.) Two inner functions are *left* (*right*) *coprime* if they have no nontrivial (i.e. nonconstant) common left (right) inner divisor. If P_1 and P_2 are right (left) coprime, then there is no cancellation in the factorization $P = P_1P_2^*$ ($P = P_1^*P_2$); we say that it is *coprime.* If there is such a factorization of P, it is unique (modulo multiplications from the right (respectively the left) by a constant unitary matrix) [14].

THEOREM 7.2. [29]. *The proper Markovian splitting subspace X with inner triplet $(K, Q, \bar{Q}^*)$ is observable if and only if K and $\bar{Q}^*$ are left coprime and constructible if and only if K and Q are right coprime.*

Proof. Let $(u, \bar{u})$ be the generating processes of X. Then the constructibility condition $S = H^- \vee \bar{S}^\perp$ can be written $H^-(du) = H^-(du_-) \vee H^-(d\bar{u})$. Applying $I_{\hat{u}}^{-1}$ to this, and using (6.23b), we obtain $\mathscr{H}_p^2 = (\mathscr{H}_p^2 Q) \vee (\mathscr{H}_p^2 K)$ which holds if and only if Q and K are right coprime [18]. In the same way we see that the observability condition $\bar{S} = H^+ \vee S^\perp$ is equivalent to the conjugate inner functions $\bar{Q}$ and K^* being right coprime, which is the same as K and $\bar{Q}^*$ being left coprime. □

The interplay between the past and the future of y can be described by the all-pass filter

(7.14)



transforming the forward innovation process u_- into the backward innovation process $\bar{u}_+$; it has the transfer function $T_0 := \bar{W}_+^{-L} W_-$. This is *not* a causal all-pass filter, unless H^- and H^+ intersect perpendicularly. For each proper Markovian splitting subspace X with inner triplet $(K, Q, \bar{Q}^*)$, the function T_0 has the factorization

$$T_0 = \bar{Q}KQ^*. \tag{7.15}$$

In view of Lemma 7.2, this follows by simple calculation, but it can also be seen by putting the boxes in (7.13) in series, after having reversed (7.13b) and (7.13c). By Theorem 7.2 and Corollary 4.1, X is minimal if and only if there is no cancellation in (7.15), i.e. the factorizations $T := \bar{Q}K$ and $\bar{T} := KQ^*$ are coprime.

What has been established so far in this section holds under the assumption that X is proper. Therefore, we may ask under what conditions at least all minimal splitting subspaces are proper.

THEOREM 7.3. [28]. *Set $T_0 := \bar{W}_+^{-L} W_-$, and let N^- and N^+ be given by* (3.12). *Then the following statements are equivalent.*

(i) *All minimal splitting subspaces are proper.*

(ii) *Both N^- and N^+ are full range.*

(iii) *There are inner functions J_1, J_2, J_3, and J_4 such that*

$$T_0 = J_1J_2^* = J_3^*J_4. \tag{7.16}$$

Proof. (i)⇒(iii): The predictor space $H^{+/-}$ is a minimal splitting subspace such that $Q = I$. Hence the second of the factorizations (7.16) follows from (7.15). In the same way the first of relations (7.16) follows from the fact that $H^{-/+}$ is a minimal splitting subspace with $\bar{Q} = I$. (iii)⇒(ii): The first of relations (7.16) yields $W_-J_2 =$

$\bar{W}_+J_1$. Since J_1 is inner, the spectral factor $W := W_-J_2$ is analytic. Let $u \in \mathcal{U}^-$ be the corresponding generating process. Since $\bar{W}_+^L W = J_1$ is inner, $H^-(d\bar{u}_+) \subset H^-(du)$, i.e. $H^+ = H^+(d\bar{u}_+) \supset H^+(du)$. Moreover, since $u \in \mathcal{U}^-$, $H^-(du) \supset H^-$ (Lemma 6.2), or, equivalently, $(H^-)^\perp \supset H^+(du)$. Hence, $N^+ := H^+ \cap (H^-)^\perp \supset H^+(du)$. Therefore, since $H^+(du)$ is full range, then so is N^+. The second relation (7.16) yields $W_-J_4^* = \bar{W}_+J_3^*$. Here $\bar{W} := \bar{W}_+J_3^*$ is a coanalytic spectral factor, and therefore the corresponding $\bar{u} \in \mathcal{U}^+$ satisfies $H^+(d\bar{u}) \supset H^+$, or, equivalently $(H^+)^\perp \supset H^-(d\bar{u})$. Also, $\bar{W}^{-L}W_- = J_4$ is inner, and hence, by Lemma 6.1, $H^-(d\bar{u}) \subset H^-$. Consequently, $N^- := H^- \cap (H^+)^\perp \supset H^-(d\bar{u})$, showing that N^- is full range. (ii)$\Rightarrow$(i): Let $X \sim (S, \bar{S})$ be a minimal splitting subspace. Then, from the proof of Theorem 3.4, we see that $S \subset (N^+)^\perp$ and $\bar{S} \subset (N^-)^\perp$, i.e. $S^\perp \supset N^+$ and $\bar{S}^\perp \supset N^-$. Therefore, if N^+ and N^- are full range, then the same must be true for $S^\perp$ and $\bar{S}^\perp$. Hence X is proper. □

A unitary function T_0 has the property (7.16) if and only if it is *strictly noncyclic*, i.e. the orthogonal complement in $\mathcal{H}_p^2$ of the range of the Hankel operator $H_{T_0}: \bar{\mathcal{H}}_p^2 \to \mathcal{H}_p^2$ defined by $H_{T_0}f = P^{\mathcal{H}_p^2}T_0 f$ (where $P^{\mathcal{X}}$ denotes the orthogonal projection onto $\mathcal{X}$) is full range [14, p. 254]. Therefore, with a slight misuse of notation, we shall say that the process y is *strictly noncyclic* if the conditions of Theorem 7.3 hold. For example, a scalar process y with spectral density $\Phi(i\omega) = (1+\omega^2)^{-3/2}$ will not satisfy these conditions; in this case $H^{+/-} = H^-$ and $H^{-/+} = H^+$ [10, p. 99]. However, it can be shown that all processes with rational spectral density are strictly noncyclic.

COROLLARY 7.1. *Suppose that y is strictly noncyclic. Then the predictor spaces $H^{+/-}$ and $H^{-/+}$, defined by* (3.13), *are proper. Let $(K_-, Q_-, \bar{Q}_-^*)$ and $(K_+, Q_+, \bar{Q}_+^*)$ respectively be their inner triplets. Then $Q_- = I$ and $\bar{Q}_+ = I$; the other inner functions are the unique solutions of the coprime factorizations*

$$T_0 = \bar{Q}_-K_- = K_+Q_+^*. \tag{7.17}$$

Proof. The factorization (7.17) was derived in the first part of the proof of Theorem 7.3. Since $H^{+/-}$ and $H^{-/+}$ are minimal, the coprimeness follows from Theorem 7.2 and Corollary 4.1. □

Now, in § 3, we saw that $H^{+/-} \sim (H^-, (N^-)^\perp)$, and hence its generating processes are $(u_-, \bar{u}_-)$, where u_- is the innovation process of y and $\bar{u}_- \in \mathcal{U}^+$ is determined, through Theorem 6.1, by

$$H^+(d\bar{u}_-) = (N^-)^\perp. \tag{7.18}$$

The analytic spectral factor is the outer spectral factor W_-, and the coanalytic one is $\bar{W}_- := \bar{W}_+\bar{Q}_-$. In the same way, $H^{-/+}$ has generating processes $(u_+, \bar{u}_+)$, where $u_+ \in \mathcal{U}^-$ is defined by

$$H^-(du_+) = (N^+)^\perp \tag{7.19}$$

and $\bar{u}_+$ is the backward innovation of y, and its spectral factors are $W_+ := W_-Q_+$ and $\bar{W}_+$, the conjugate outer spectral factor.

Next, we shall take a closer look at the minimal Markovian splitting subspaces of a strictly noncyclic process y.

THEOREM 7.4. [31]. *Suppose that y is strictly noncyclic. Let $X \sim (S, \bar{S})$ be a Markovian splitting subspace. Then the following conditions are equivalent.*

(i) *X is minimal.*

(ii) *X is observable and $S \subset (N^+)^\perp$.*

(iii) *X is constructible and $\bar{S} \subset (N^-)^\perp$.*

The proof of this theorem[6], which can be found in [31] and will not be repeated here, is based on the observation that the structural functions of any two X satisfying (ii) or (iii) have the same invariant factors. The *invariant factors* of a $p \times p$ inner function K are scalar inner functions $k_1, k_2, \cdots, k_p$ defined in the following way. Set $\gamma_0 = 1$, and, for $i = 1, 2, \cdots, p$, define γ_i to be the greatest common inner divisor of all $i \times i$ minors of K. Then set $k_i := \gamma_i / \gamma_{i-1}$ for $i = 1, 2, \cdots, p$. Clearly these functions are inner, for γ_{i-1} divides γ_i. (Two inner functions with the same invariant factors are called *quasi-equivalent* [14].) Consequently we have the following important corollary, the significance of which will become evident in § 10.

THEOREM 7.5. [31]. *Suppose that y is strictly noncyclic. Let K_1 and K_2 be the structural functions of two minimal Markovian splitting subspaces. Then K_1 and K_2 have the same invariant factors.*

To illustrate this result, let us consider the following example [31]. Let y be a two-dimensional process with the rational spectral density

$$\Phi(s) = \frac{1}{(s^2-1)(s^2-4)} \begin{bmatrix} 17-2s^2 & -(s+1)(s-2) \\ -(s-1)(s+2) & 4-s^2 \end{bmatrix}.$$

Then it can be seen that the structural function of $H^{+/-}$ is

$$K_-(s) = \frac{s-1}{(s+1)(s+2)} \begin{bmatrix} s-1.2 & 1.6 \\ 1.6 & s+1.2 \end{bmatrix}$$

and that the one of $H^{-/+}$ is

$$K_+(s) = \frac{s-1}{(s+1)(s+2)} \begin{bmatrix} s-70/37 & 24/37 \\ 24/37 & s+70/37 \end{bmatrix}.$$

These functions look quite different, but they have the same invariant factors, namely

$$k_1(s) = \frac{s-1}{s+1} \quad \text{and} \quad k_2(s) = \frac{(s-1)(s-2)}{(s+1)(s+2)},$$

and are therefore quasi-equivalent.

In the scalar case ($m = 1$), quasi-equivalence reduces to equality.

COROLLARY 7.2. *Suppose that y is scalar and strictly noncyclic. Then all minimal Markovian splitting subspaces have the same structural function.*

Conditions (ii) and (iii) of Theorem 7.4 suggest the following definitions for minimality of spectral factors [41]. An analytic full-rank spectral factor W is *minimal* if the corresponding $u \in \mathcal{U}^-$ satisfies the condition $H^-(du) \subset (N^+)^\perp$; a coanalytic full-rank spectral factor $\bar{W}$ is *minimal* if its $\bar{u} \in \mathcal{U}^+$ satisfies $H^+(d\bar{u}) \subset (N^-)^\perp$. These definitions are justified by the following result.

COROLLARY 7.3. *Let y be strictly noncyclic. Then there is a one-one correspondence between the class of minimal Markovian splitting subspaces X and the class of minimal analytic (coanalytic) spectral factors $W[\bar{W}]$ (modulo multiplication from the right by constant unitary matrices). The correspondence $X \leftrightarrow W$ $[X \leftrightarrow \bar{W}]$ is that of Theorem* 7.1.

Proof. In view of Theorem 7.4 and the observability condition $\bar{S} = H^+ \vee S^\perp$ (Theorem 4.1), there is a one-one correspondence between minimal $X \sim (S, \bar{S})$ and $u \in \mathcal{U}^-$ such that $S = H^-(du) \subset (N^+)^\perp$, i.e. to minimal analytic spectral factors. Here the correspondence $u \leftrightarrow S$ is by Theorem 6.1 and is hence modulo the transformations described there. The proof of the symmetric statement is analogous. □

[6] Theorem 7.4 was first stated in [28], but there is a nontrivial gap in the proof. The same incomplete argument was used in [41], [43].

Note, however, that a Markovian splitting subspace X need not be minimal even if both its analytic and coanalytic spectral factors are minimal; the only thing we can say in this case is that $X \subset H^{\square}$, the frame space.

In § 8, we show that, if $W[\bar{W}]$ is rational, it is minimal if and only if its degree is minimal. This is the concept of minimality mentioned in § 1.

The scalar version of the following result is due to Ruckebusch [41].

PROPOSITION 7.2. [41], [28]. *Suppose that y is strictly noncyclic. Then* (i) $W := W_- Q$ *is a minimal analytic spectral factor if and only if* $Q|_L Q_+$; *and* (ii) $\bar{W} := \bar{W}_+ \bar{Q}$ *is a minimal coanalytic spectral factor if and only if* $\bar{Q}|_L \bar{Q}_-$.

Proof. Let $u \in \mathcal{U}^-$ be the Wiener process of W. Then, by definition, W is minimal if and only if $H^-(du) \subset H^-(du_+)$, which is equivalent to $P := W^{-L} W_+$ being inner (Lemma 6.1). Now, in view of Lemma 7.2, $Q_+ := W_-^{-L} W_+ = W_-^{-L} W W^{-L} W_+ = QP$. Therefore, W is minimal if and only if $Q|_L Q_+$. This establishes (i); (ii) is proved in the same way. □

Now, by Corollary 7.3 and Proposition 7.2, there is a one-one correspondence (modulo trivial transformations) between minimal Markovian splitting subspaces X and left inner divisors Q of Q_+. This provides a parametrization $\{X_Q; Q|_L Q_+\}$ of the class of minimal Markovian splitting subspaces which introduces a natural *partial ordering* of this class, under which $X_{Q_1} < X_{Q_2}$ if and only if $Q_1|_L Q_2$. Here there are a minimal element $X_I = H^{+/-}$ and a maximal element $X_{Q_+} = H^{-/+}$. Obviously this is the lattice structure described in the end of § 4. (A similar parameterization can of course be obtained in terms of the conjugate inner functions $\bar{Q}$ such that $\bar{Q}|_L \bar{Q}_-$.)

Given a left inner divisor Q of Q_+, how do we determine X_Q? The inner triplet $(K, Q, \bar{Q}^*)$ can be determined from the factorization (7.15) as described in the following lemma.

LEMMA 7.3. *Suppose y is strictly noncyclic. Let Q be a left inner divisor of* Q_+, *and define* $T := T_0 Q$. *Then, T has a unique* (*modulo constant unitary factors*) *coprime factorization*

$$T = \bar{Q} K \tag{7.20}$$

where K is inner, $\bar{Q}$ *is conjugate inner and K and* $\bar{Q}^*$ *are left coprime. Moreover,* $(K, Q, \bar{Q}^*)$ *is the inner triplet of* X_Q.

Proof. Let $(K, Q, \bar{Q})$ be the inner triplet of X_Q. Then (7.20) follows from (7.15). Since X_Q is observable, K and $\bar{Q}^*$ are left coprime (Theorem 7.2). As pointed out above, the coprime factorization is unique, in the sense described in the lemma [14]. Since we do not distinguish between equivalent inner triplets (differing only by constant unitary factors), the lemma follows. □

(For the relationship between the factorization (7.20) and the corresponding Hankel operators, the reader is referred to [30].) Consequently, in view of Theorem 7.1, we have the following representation theorem for the class of minimal Markovian splitting subspaces.

THEOREM 7.6. *Suppose that y is strictly noncyclic. Then a subspace X of H is a minimal Markovian splitting subspace if and only if*

$$X = \int_{-\infty}^{\infty} \mathcal{H}(K) Q^* \, d\hat{u}_- \tag{7.21}$$

for some $Q|_L Q_+$, *where K is the inner factor in the coprime factorization* (7.20) *and* u_- *is the innovation process of y.*

An alternative formulation of this theorem goes as follows. (Here $P^{\mathcal{Z}}$ denotes orthogonal projection onto the subspace $\mathcal{Z}$ and $\bar{P}^{\mathcal{Z}}\mathcal{H}$ the closure of $P^{\mathcal{Z}}\mathcal{H}$.)

THEOREM 7.7. [28]. *Suppose that y is strictly noncyclic. Then a subspace X of H is a minimal Markovian splitting subspace if and only if*

$$X = \int_{-\infty}^{\infty} [\bar{P}^{\mathscr{H}_p^2 Q^*}(\bar{\mathscr{H}}_p^2 T_0)]\, d\hat{u}_- \tag{7.22}$$

for some left inner divisor Q of Q_+.

Proof. We need to show that X_Q is given by (7.22). Let $u \in \mathscr{U}^-$ be the forward generating process of X_Q, and let W be the corresponding analytic spectral factor. Then $W^{-L}W_- = Q^*$. Now, in view of Corollary 4.2 and (7.10),

$$X_Q = \bar{E}^{H^-(du)} H^+(d\bar{u}_+). \tag{7.23}$$

By (6.23b), $H^-(du)$ and $H^+(d\bar{u}_+)$ correspond, under the isomorphism $I_{\hat{u}}$, to $\mathscr{H}_p^2 Q^*$ and $\bar{\mathscr{H}}_p^2 T_0$ respectively, and therefore (7.22) follows from (7.23). □

Of course, there are also backward versions of Theorems 7.6 and 7.7 in which $\bar{Q}$ plays the role of Q.

8. Stochastic realizations: the finite-dimensional case.

PROPOSITION 8.1. *Let X be a proper Markovian splitting subspace. Then X is finite-dimensional if and only if its structural function K is rational.*

Proof. Let $(u, \bar{u})$ be the generating processes of X. Then, by Corollary 3.1, $X = E^{H^-(du)} H^+(d\bar{u})$, the isomorphic image of which under $I_{\hat{u}}^{-1}$ is $P^{\mathscr{H}_p^2}(\bar{\mathscr{H}}_p^2 K)$. Consequently, X is isomorphic to the range of the Hankel operator $H_K: \bar{\mathscr{H}}_p^2 \to \mathscr{H}_p^2$ defined by $H_K f = P^{\mathscr{H}_p^2} fK$, which, by [14, Thm. 3.8, p. 256], is finite-dimensional if and only if K is rational. □

Now, let X be a finite-dimensional, but not necessarily minimal, Markovian splitting subspace with structural function K and generating processes $(u, \bar{u})$. Then K is rational, and there is a coprime factorization

$$K(s) = \bar{D}(s) D(s)^{-1} \tag{8.1}$$

where D and $\bar{D}$ are real invertible $p \times p$ polynomial matrices which are right coprime, i.e. any common right divisor is unimodular[7] [14], [47]. The matrix polynomial D and $\bar{D}$ are unique (modulo a common unimodular factor). To maintain the symmetry between the past and the future in our presentation we also note that

$$K^*(s) = D(s) \bar{D}(s)^{-1}. \tag{8.2}$$

The following result shows that $\mathscr{H}(K)$, the isomorphic image of X under $I_{\hat{u}}$, consists of rational functions which are *strictly proper*, i.e., in each component, the numerator polynomial is of lower degree than the denominator polynomial.

THEOREM 8.1. [29]. *Let the inner function K have the polynomial-matrix-fraction representation* (8.1). *Then*

$$\mathscr{H}(K) = \{gD^{-1} | g \in \mathbb{R}^p[s];\ gD^{-1} \text{ strictly proper}\} \tag{8.3}$$

where $\mathbb{R}^p[s]$ is the vector space of p-dimensional row vectors of polynomials.

For the proof we need the following lemma.

LEMMA 8.1. *If K is rational, the space $\mathscr{H}(K)$ consists of strictly proper rational functions.*

Proof. Set $k := \det K$. Then $\mathscr{H}_p^2 k \subset \mathscr{H}_p^2 K$ [14, p. 187], and consequently $\mathscr{H}(K) \subset \mathscr{H}(kI)$. Therefore, it is no restriction to study the scalar case $p = 1$. In fact, if K is

[7] A matrix polynomial is *unimodular* if it has a polynomial inverse.

rational, then so is k. So, if we can prove that the scalar $\mathcal{H}(k)$ consists of rational functions, the same holds true for $\mathcal{H}(kI)$ and thus for $\mathcal{H}(K)$. A scalar rational inner function k is a finite Blaschke product [14], [18], i.e. a *finite* product of coprime functions $k_i(s) := (s-s_i)^{\nu_i}(s+\bar{s}_i)^{-\nu_i}$, where, for each i, s_i is a complex number, $\bar{s}_i$ its complex conjugate, and ν_i an integer. Then $\mathcal{H}^2 k = \cap_i \mathcal{H}^2 k_i$, and hence $\mathcal{H}(k) = \vee_i \mathcal{H}(k_i)$, so it is enough to show that any $\mathcal{H}(k_i)$ consists of rational functions. To this end, we quote from [10, p. 34] that

$$e_j(s) = \frac{1}{s+\bar{s}_i}\left[\frac{s-s_i}{s+\bar{s}_i}\right]^j, \qquad j = 0, 1, 2, \cdots \tag{8.4}$$

is an orthogonal basis in $\mathcal{H}^2$. However, $e_j k_i = e_{j+\nu_i}$, and hence $\mathcal{H}^2 k_i$ is spanned by $\{e_{\nu_i}, e_{\nu_{i+1}}, \cdots\}$. Therefore, $\mathcal{H}(k_i)$ is the span of $\{e_0, e_1, \cdots, e_{\nu_{i-1}}\}$, which is a space of strictly proper rational functions. Consequently, the same is true for $\mathcal{H}(K)$, as required. □

Proof of Theorem 8.1. It is not hard to show that

$$\mathcal{H}(K) = \{f \in \mathcal{H}_p^2 | fK^* \in \bar{\mathcal{H}}_p^2\} \tag{8.5}$$

[18, p. 75]. In view of (8.2), this may be written

$$\mathcal{H}(K) = \{gD^{-1} \in \mathcal{H}_p^2 | g\bar{D}^{-1} \in \bar{\mathcal{H}}_p^2\}. \tag{8.6}$$

Since $gD^{-1} \in \mathcal{H}(K)$ is rational (Lemma 8.1), then so is g. Any rational g such that $gD^{-1} \in \mathcal{H}_p^2$ and $g\bar{D}^{-1} \in \bar{\mathcal{H}}_p^2$ must be analytic in the whole complex plane, and hence $g \in \mathbb{R}^p[s]$. By Lemma 8.1, gD^{-1} is strictly proper. □

COROLLARY 8.1. [29]. *Let X be a finite-dimensional Markovian splitting subspace with structural function* (8.1). *Then the corresponding spectral factors W and $\bar{W}$ are strictly proper rational. In fact,*

$$W(s) = N(s)D(s)^{-1}, \tag{8.7a}$$

$$\bar{W}(s) = N(s)\bar{D}(s)^{-1} \tag{8.7b}$$

for some $m \times p$-matrix polynomial N.

Proof. By the definition (7.1), $W = \bar{W}K$ (Lemma 7.2), i.e. $\bar{W} = WK^*$, and therefore, in view of (8.5) and the fact that W is analytic and $\bar{W}$ is coanalytic, the rows of W belong to $\mathcal{H}(K)$. Hence, by Theorem 8.1, W is strictly proper rational and has a representation (8.7a). However, (7.1) and (8.1) yield $\bar{W}\bar{D} = WD$, which is precisely N. Hence (8.7b) follows. Since $\bar{W}$ is square-integrable, it must be strictly proper. □

It is important to note that the factorization (8.7) need *not* be coprime. The significance of this will be made clear below.

We proceed to construct a basis in X. To this end, we shall choose the arbitrary unimodular factor in (8.1) so that (i) if n_i is the degree of the ith column of $\left[\begin{smallmatrix} D \\ \bar{D} \end{smallmatrix}\right]$, then $n_1 + n_2 + \cdots + n_p = n$, where n is the common degree of $\det D$ and $\det \bar{D}$; and (ii) D and $\bar{D}$ are *column proper*, i.e. the highest-degree coefficient matrices D_h and $\bar{D}_h$ are full rank; here D_h ($\bar{D}_h$) is the constant matrix whose ith column consists of the coefficients of s^{n_i} in the ith column of D ($\bar{D}$). It is always possible to choose D and $\bar{D}$ in this way, and there are procedures to achieve it [13], [20], [47]. With this representation, we have

$$D(s) = D_h\{\text{diag}\{s^{n_1}, s^{n_2}, \cdots, s^{n_p}\} + D_0\Pi(s)\}, \tag{8.8a}$$

$$\bar{D}(s) = \bar{D}_h\{\text{diag}\{s^{n_1}, s^{n_2}, \cdots, s^{n_p}\} + \bar{D}_0\Pi(s)\} \tag{8.8b}$$

where $\text{diag}\{s^{n_1}, s^{n_2}, \cdots, s^{n_p}\}$ is the $p \times p$ matrix with $s^{n_1}, s^{n_2}, \cdots, s^{n_p}$ on the diagonal

and zeros elsewhere, the $n \times p$-matrix polynomial $\Pi(s)$ is the transpose of

$$\Pi(s)' := \begin{bmatrix} s^{n_1-1}, \cdots, s, 1 & & \\ & s^{n_2-1}, \cdots, s, 1 & \\ & & \ddots \\ & & s^{n_p-1}, \cdots, s, 1 \end{bmatrix} \tag{8.9}$$

(where empty spaces are zeros), and D_0 and $\bar{D}_0$ are constant $p \times n$ matrices.

LEMMA 8.2. [29]. *The n rows of the $n \times p$ matrix $\Pi(s)D(s)^{-1}$ of rational functions form a basis in $\mathscr{H}(K)$.*

Proof. The rows of ΠD^{-1} are clearly linearly independent. It remains to show that they span $\mathscr{H}(K)$. In view of Theorem 8.1, this amounts to showing that gD^{-1} is strictly proper for precisely those $g \in \mathbb{R}^p[s]$ which can be written $a\Pi(s)$ for some row vector $a \in \mathbb{R}^n$, i.e. for those $g \in \mathbb{R}^p[s]$ with $\deg g_i < n_i$ for $i = 1, 2, \cdots, p$. By Cramer's rule, $[D(s)^{-1}]_{ij} = (-1)^{i+j}\Delta_{ji}(s)/\Delta(s)$, where $\Delta := \det D$ and Δ_{ji} is the determinant of the matrix obtained by deleting row j and column i in D. Hence, Δ_{ji} is a sum of products of one element from each of all columns of D except the ith, and consequently $\deg \Delta_{ji} \leqq n - n_i$. Since D_h is full rank, for each i, there is a j such that equality holds. In fact, in each column i of D there is a row j such that deleting row j and column i in D_h leaves a nonsingular matrix. Hence in forming Δ_{ji} there is at least one product that contains only factors of highest degree. Since $\deg \Delta = n$, gD^{-1} is therefore strictly proper if and only if $\deg g_i < n_i$ for $i = 1, 2, \cdots, p$ as required. □

THEOREM 8.2. [29]. *Let X be a finite-dimensional Markovian splitting subspace with structural function* (8.1) *and spectral factors* (8.7). *Then, for each $t \in \mathbb{R}$, the components of the random vector*

$$x(t) = \int_{-\infty}^{\infty} e^{i\omega t}\Pi(i\omega)N(i\omega)^{-L}\, d\hat{y} \tag{8.10}$$

form a basis in U_tX. Moreover,

$$y(t) = Cx(t), \tag{8.11}$$

where the matrix C is uniquely determined by identifying coefficients of like power of s in

$$N(s) = C\Pi(s). \tag{8.12}$$

The process x also has the representations

$$x(t) = \int_{-\infty}^{\infty} e^{i\omega t}\Pi(i\omega)D(i\omega)^{-1}\, d\hat{u}, \tag{8.13a}$$

$$x(t) = \int_{-\infty}^{\infty} e^{i\omega t}\Pi(i\omega)\bar{D}(i\omega)^{-1}\, d\hat{\bar{u}} \tag{8.13b}$$

where $(u, \bar{u})$ are the generating processes of X.

Proof. In view of Lemma 8.2, it is immediately clear from (7.2) that $\{x_1(0), x_2(0), \cdots, x_n(0)\}$, as defined by (8.13a), is a basis in X. Then, it follows from (6.18) that $\{x_1(t), x_2(t), \cdots, x_n(t)\}$ is a basis in U_tX for each $t \in \mathbb{R}$. From (6.16) and (8.7a) we have that $D^{-1}\, d\hat{u} = N^{-L}\, d\hat{y}$, and hence (8.10) and (8.13a) are equivalent. The equivalence of (8.13a) and (8.13b) follows from $d\hat{\bar{u}} = K\, d\hat{u}$ and (8.1). In the proof of Corollary 8.1, we saw that the rows of W belong to $\mathscr{H}(K)$. Therefore, by Lemma 8.2, there is an $m \times n$ matrix C such that $W(s) = C\Pi(s)D(s)^{-1}$; hence, in view of (8.7a), (8.12) holds. It is easy to see that C is uniquely determined by this relation. Inserting this expression for W in (6.12)+(6.17) and observing (8.13a) we have (8.11). □

In particular, it follows from Theorem 8.2 that

$$U_t X = \{ax(t) | a \in \mathbb{R}^n\} \tag{8.14}$$

which is precisely (1.4). Hence x is a *state process* of X.

It remains to find two Markovian representations for the state process x, a forward one generated by u and a backward one driven by $\bar{u}$. To this end, we define the $n \times n$ matrices A and $\bar{A}$ and the $n \times p$ matrices B and $\bar{B}$ as

$$\begin{aligned} A &:= J - \Pi(0) D_0, & \bar{A} &:= -J + \Pi(0) \bar{D}_0, \\ B &:= \Pi(0) D_h^{-1}, & \bar{B} &:= \Pi(0) \bar{D}_h^{-1}. \end{aligned} \tag{8.15}$$

Here J is the block diagonal matrix

$$J = \operatorname{diag} \{J_{n_1}, J_{n_2}, \cdots, J_{n_p}\} \tag{8.16}$$

where J_k is the $k \times k$ shift matrix with ones on the superdiagonal and zeros elsewhere. The pair $[J, \Pi(0)]$ is known as the *Brunovsky canonical form*, and $\{n_1, n_2, \cdots, n_p\}$ are known as its *indices*.

THEOREM 8.3. *Let $\{x(t); t \in \mathbb{R}\}$ be the state process (8.10) of the Markovian splitting subspace*

$$X = \{ax(0) | a \in \mathbb{R}^n\} \tag{8.17}$$

and let A, B, $\bar{A}$, and $\bar{B}$ be defined by (8.15). Then A and $\bar{A}$ have the same eigenvalues, all located in the open left half plane, and $[B, AB, A^2B, \cdots]$ and $[\bar{B}, \bar{A}\bar{B}, \bar{A}^2\bar{B}, \cdots]$ have full rank. Moreover, x has the two representations

$$x(t) = \int_{-\infty}^{\infty} e^{A(t-\sigma)} B \, du(\sigma), \tag{8.18a}$$

$$x(t) = -\int_{-\infty}^{\infty} e^{\bar{A}(\sigma - t)} \bar{B} \, d\bar{u}(\sigma) \tag{8.18b}$$

where $(u, \bar{u})$ are the generating processes of X and the integrals are defined in quadratic mean.

Proof. A simple calculation yields $(sI - J)\Pi(s) = \Pi(0) \operatorname{diag} \{s^{n_1}, s^{n_2}, \cdots, s^{n_p}\}$, and consequently $(sI - A)\Pi(s) = BD(s)$, i.e.

$$\Pi(s) D(s)^{-1} = (sI - A)^{-1} B. \tag{8.19}$$

It is well known and easy to show that $[B, AB, A^2B, \cdots]$ has full rank, and therefore (8.19) has degree n [20], [47]. Consequently $\det D(s)$ and $\det (sI - A)$ have the same zeros, i.e. the eigenvalues of A and the zeros of $\det D(s)$ coincide. In the same way, we see that

$$\Pi(s) \bar{D}(s)^{-1} = (sI + \bar{A})^{-1} \bar{B}, \tag{8.20}$$

and therefore, since $[\bar{B}, \bar{A}\bar{B}, \bar{A}^2\bar{B}, \cdots]$ is full rank [20], [47], the eigenvalues of $\bar{A}$ are the zeros of $\det \bar{D}(-s)$. In view of (8.1), $\det K = \det \bar{D} / \det D$, which is a finite Blaschke product [14], [18]. Such a function has all its poles in the open left half plane, and the zeros of its numerator polynomial are related to those of its denominator polynomial by a simple change of sign. Consequently, the zeros of $\det D(s)$ and $\det \bar{D}(-s)$ coincide, i.e. A and $\bar{A}$ have the same eigenvalues, and they are located in the open left half plane. The rows of (8.19) belong to $\mathscr{H}_p^2$ (Lemma 8.2), and the inverse Fourier transform of $(i\omega I - A)^{-1} B$ is $e^{At} B$ for $t \geqq 0$ and zero otherwise. Consequently, in view of (6.7)

and (6.18), (8.18a) follows from (8.13a). In the same way, (8.18b) follows from (8.13b). In fact, $(i\omega I + \bar{A})^{-1}$ has inverse Fourier transform $e^{-\bar{A}t}$ for $t \leqq 0$ and zero otherwise, its rows belonging to $\bar{\mathscr{H}}_p^2$. □

Therefore, given a finite-dimensional Markovian splitting subspace X with generating processes $(u, \bar{u})$, there are a forward stochastic realization

$$\Sigma \begin{cases} dx = Ax\,dt + B\,du, \\ y = Cx \end{cases} \tag{8.21a}$$

and a backward one

$$\bar{\Sigma} \begin{cases} dx = \bar{A}x\,dt + \bar{B}\,d\bar{u}, \\ y = Cx \end{cases} \tag{8.21b}$$

such that (8.17) holds; this follows from (8.11) and (8.18). We shall call them the *standard* (*forward and backward*) *realizations* of X. The fact that Σ is forward and $\bar{\Sigma}$ is backward is seen from (8.18), but it can also be illustrated by (3.6) rewritten as

$$H = H^-(d\bar{u}) \oplus X \oplus H^+(du), \tag{8.22}$$

i.e. the components of the state $x(0)$ are orthogonal to the future increments of u and to the past increments of $\bar{u}$.

From Theorem 8.3 it also follows that Σ is always reachable and $\bar{\Sigma}$ is always controllable, with these terms defined as in § 1. The circumstances under which Σ is observable and $\bar{\Sigma}$ is constructible is described by the following theorem.

THEOREM 8.4. *Let X be a finite-dimensional Markovian splitting subspace with spectral factors $(W, \bar{W})$ and standard realizations* (8.21). *Then, W is the transfer function of Σ, and the following conditions are equivalent.*

(i) *X is observable.*

(ii) *Σ is observable.*

(iii) *The factorization $W = ND^{-1}$ of Corollary* 8.1 *is coprime, i.e. N and D are right coprime.*

Symmetrically, $\bar{W}$ is the transfer function of $\bar{\Sigma}$ and the following conditions are equivalent.

(iv) *X is constructible.*

(v) *Σ is constructible.*

(vi) *The factorization $\bar{W} = N\bar{D}^{-1}$ is coprime.*

Proof. We shall only consider the first part. The second follows by symmetry. In view of (8.12), (8.19) and (8.7a), we have

$$W(s) = C(sI - A)^{-1}B \tag{8.23}$$

and consequently W is the transfer function of Σ. Then, the equivalence of (ii) and (iii) follows from [14, p. 41] or [20, p. 439], so it only remains to show that (i) and (ii) are equivalent. With $S = H^-(du)$, it follows from (8.17a) and (8.11) that

$$E^S a y(t) = aC\,e^{At}x(0) \tag{8.24}$$

for any row vector $a \in \mathbb{R}^m$, and consequently

$$\bar{E}^S H^+ = \overline{\operatorname{span}}\,\{aCe^{At};\ t \geqq 0,\ a \in \mathbb{R}^m\}x(0). \tag{8.25}$$

By Corollary 4.2, the left member of (8.25) equals X if and only if X is observable. On the other hand, Σ is observable if and only if the range of $\{e^{A't}C';\ t \geqq 0\}$ is dense in $\mathbb{R}^n$ [22]. Therefore, it follows from (8.17) that (i) and (ii) are equivalent. □

We shall say that a finite-dimensional stochastic realization is *minimal* if there is no other realization with a state process x of smaller dimension. Together with Theorem 8.2, the following result implies that Σ and $\bar{\Sigma}$ are minimal if and only if X is minimal.

THEOREM 8.5. *A finite-dimensional splitting subspace is minimal if and only if its dimension is minimal.*

Proof. Let X be a splitting subspace. First, assume that there is a splitting subspace X_1 of smaller dimension than X. By Corollary 3.5, X_1 contains a minimal splitting subspace X_2. Since $\dim X_2 \leqq \dim X_1 < \dim X$, Theorem 4.2 implies that X is non-minimal. Second, suppose that X is not minimal. Then it contains a minimal splitting subspace as a proper subspace (Corollary 3.5), and thus X cannot have minimal dimension. □

By Theorem 8.4 and Corollary 4.1, it is not enough for the stochastic realization Σ to be both reachable and observable to be minimal as is the case in deterministic realization theory; for this to happen the backward realization $\bar{\Sigma}$ must be constructible also, or, alternatively, the analytic spectral factor W must be minimal in the sense described in § 7. In the finite-dimensional case under discussion, minimality of spectral factors can be related to their degrees, as the following result shows.

COROLLARY 8.2. *Let X be a finite-dimensional Markovian splitting subspace with spectral factors $(W, \bar{W})$. Then*

$$\dim X \geqq \deg W \tag{8.26}$$

with equality if and only if X is observable, and

$$\dim X \geqq \deg \bar{W} \tag{8.27}$$

with equality if and only if X is constructible. Moreover, $W[\bar{W}]$ is minimal if and only if its degree is as small as possible.

Proof. By Theorem 8.2, $\dim X$ equals n, the degree of $\det D$. But, since $W = ND^{-1}$, $\deg W \leqq n$, with equality if and only if N and D are right coprime, which, in view of Theorem 8.4, holds if and only if X is observable. Now, suppose that X is observable. Then, $\deg W = \dim X$. Since W is minimal if and only if X is minimal (Corollary 7.4) and X is minimal if and only if $\dim X$ is minimal (Theorem 8.5), W is minimal if and only if $\deg W$ is minimal. The proofs of the statements concerning $\bar{W}$ are analogous. □

In view of Theorem 7.4, we have also established the following result.

COROLLARY 8.3. *Let X be a finite-dimensional Markovian splitting subspace with spectral factors $(W, \bar{W})$ and standard realizations* (8.21). *Then the following conditions are equivalent.*

(i) *X is minimal.*
(ii) *Σ is minimal.*
(iii) *Σ is observable and W is minimal.*
(iv) *$\bar{\Sigma}$ is minimal.*
(v) *$\bar{\Sigma}$ is constructible and $\bar{W}$ is minimal.*
(vi) *Σ is observable and $\bar{\Sigma}$ is constructible.*

As an application of Theorem 8.4, let us give an alternative characterization of the class of minimal Markovian splitting subspaces in the case that y is a *scalar* process. Then, N, D, and $\bar{D}$ are scalar and $\bar{D}(s) = D(-s)$, for K is a finite scalar Blaschke product. Minimality of X requires that both Condition (iii) and Condition (vi) in Theorem 8.4 are satisfied, i.e. $N(s)$ and $\psi(s) := D(s)D(-s)$ are coprime. This is clearly equivalent to coprimeness of $\varphi(s) := N(s)N(-s)$ and $\psi(s)$. Therefore, we can characterize the class of minimal Markovian splitting subspaces in the following way.

Write the rational density Φ of y as $\Phi = \varphi/\psi$ where φ and ψ are coprime polynomials. For each polynomial solution N of

$$N(s)N(-s) = \varphi(s) \tag{8.28}$$

form

$$X = \int_{-\infty}^{\infty} \left\{ \frac{g(s)}{N(s)} \middle| \deg g < \frac{1}{2} \deg \psi \right\} d\hat{y} \tag{8.29}$$

where $\deg g < n$ means that g is a polynomial of degree less than n. It follows from (8.10), (8.17), and what has been said above that this procedure produces precisely the minimal splitting subspaces of y.

9. Stochastic realizations: the general case. In § 8, given a Markovian splitting subspace X of finite dimension n, we constructed a state process $\{x(t);\, t \in \mathbb{R}\}$ taking values in $\mathbb{R}^n$ and forward and backward differential equation representations for it. The main point of this construction is a convenient choice of basis in X. In this basis the matrix representations e^{At} and $e^{\bar{A}t}$ of the Markov semigroups $\{U_t(X)^*\}$ and $\{U_t(X)\}$ can be found almost by inspection from the matrix fraction representation (8.1) of the structural function K. This immediately leads to the forward and backward realizations (8.21) of the process y in the familiar state space form. So, in the finite-dimensional case, the passage from any solution of the abstract realization problem, i.e. a Markovian splitting subspace $X \sim (S, \bar{S})$ and a corresponding Markov semigroup $\{U_t(X);\, t \in \mathbb{R}\}$, is merely a question of coordinatization.

On the other hand, the theory developed up to § 8 is absolutely independent of any restrictions of the dimension of X. The natural question to ask at this point is thus the following. Given a Markovian splitting subspace of possibly infinite dimensions, when is it possible to obtain differential equations representations for $\{y(t);\, t \in \mathbb{R}\}$ of the type (1.7) and (1.9)?

This is basically a *representation problem* in which one seeks a global description in terms of local or infinitesimal data. As such it has no meaningful solution in general. Obtaining differential equation representations for a process with nonrational spectrum necessarily involves restrictions of a technical nature (essentially smoothness conditions) on the underlying spectral factors. The elucidation of these conditions is one of the goals of this section. Note that there are several possible mathematical frameworks for infinite-dimensional Markov processes as solutions of stochastic differential equations (e.g. [17] and [49]), all of which coincide when specialized to the finite-dimensional case. Here we shall work in a setting which looks most natural to us, but other approaches are possible.

The problem dealt with in this section might seem relevant only from a purely theoretical point of view. However, we remark that many engineering problems involve random processes with nonrational spectra, e.g. turbulence, wave spectra, gyroscopic noise, etc. In practical problems, these spectra must be approximated, and finite-dimensional approximate realizations must be constructed. Understanding the exact structure of the infinite-dimensional state space models for these processes is probably the best way to gain insight into the approximation process and to design efficient finite-dimensional filters.

An important feature of the construction in § 8 is that $x(0)$ is a basis in X so that the *state space* $\mathbb{R}^n$, i.e. the space in which the state process $\{x(t);\, t \in \mathbb{R}\}$ takes values, and the splitting subspace X have the same dimension (and are therefore isomorphic). Choosing the state space in this way insures that the forward realization Σ is reachable

and the backward realization $\bar{\Sigma}$ is controllable. Of course we could have achieved the same thing by taking as the state space any vector space $\mathscr{X}$ isomorphic to X such as, for example, an n-dimensional vector space of polynomials in the style of Fuhrmann [14], thereby obtaining a coordinate-free representation.

In this section we shall assume that X is a possibly infinite-dimensional (not necessarily minimal) proper Markovian splitting subspace with spectral factors $(W, \bar{W})$ and generating processes $(u, \bar{u})$. As before, it is reasonable to take as the state space a Hilbert space isomorphic to X. In this paper, we shall choose $\mathscr{X} := I_u^* X$ as the state space of the forward realization and $\bar{\mathscr{X}} := I_{\bar{u}}^* X$ as the state space in the backward one. then, by (6.7), (7.2) and (7.4), $\mathscr{X} = \mathscr{F}^* \mathscr{H}(K)$ and $\bar{\mathscr{X}} = \mathscr{F}^* \mathscr{H}(K^*)$ where K is the structural function of X.

As explained in § 5, the forward realization should, in an abstract sense, be a stochastic dynamical system with input u and semigroup $\{U_t(X)^*; t \geqq 0\}$. With our present choice of state space we should therefore take

$$e^{At} := I_u^* U_t(X)^* I_u. \tag{9.1}$$

Of course, as should be, $\{e^{At}; t \geqq 0\}$ is a strongly continuous contraction semigroup on $\mathscr{X}$ (Theorem 5.2), and the infinitesimal generator A is in general an unbounded operator with domain $\mathscr{D}(A)$ dense in $\mathscr{X}$.

In the same way, the backward realization should have input $\bar{u}$ and a semigroup isomorphic to $\{U_t(X); t \geqq 0\}$. We take

$$e^{\bar{A}t} := I_{\bar{u}}^* U_t(X) I_{\bar{u}} \tag{9.2}$$

defining a strongly continuous contraction semigroup on the state space $\bar{\mathscr{X}}$ of the backward realization.

It remains to determine maps $B: \mathbb{R}^p \to \mathscr{X}$ and $C: \mathscr{X} \to \mathbb{R}^m$ for the forward realization and $\bar{B}: \mathbb{R}^p \to \bar{\mathscr{X}}$ and $\bar{C}: \bar{\mathscr{X}} \to \mathbb{R}^m$ for the backward realization having the appropriate properties. We would like these maps to be bounded.

We begin with the forward realization. Let $\xi \in X$ be arbitrary, and let $f \in \mathscr{X}$ be the corresponding point in the state space, i.e. $\xi = I_u f$. Then

$$U_t \xi = \int_{-\infty}^{\infty} f(-\sigma)\, du(\sigma + t) = \int_{-\infty}^{\infty} f(t - \sigma)\, du(\sigma). \tag{9.3}$$

But, $\mathscr{F} f \in \mathscr{H}(K) \subset \mathscr{H}_p^2$, and therefore f vanishes on negative real line so that

$$U_t \xi = \int_{-\infty}^{t} f(t - \sigma)\, du(\sigma). \tag{9.4}$$

Consequently, since $S = H^-(du)$, (5.3a) yields

$$U_t(X)\xi = \int_{-\infty}^{0} f(t - \sigma)\, du(\sigma). \tag{9.5}$$

It follows from (9.1) that $U_t(X)\xi = I_u e^{A^* t} f$, and hence

$$(e^{A^* t} f)(\tau) = \begin{cases} f(t + \tau) & \text{for } \tau \geqq 0, \\ 0 & \text{for } \tau < 0. \end{cases} \tag{9.6}$$

Therefore, whenever defined, $A^* f$ is the derivative of f in the $\mathscr{L}^2$ sense [3].

Now, following a standard construction [3], define $\mathscr{X}$ to be the domain $\mathscr{D}(A^*)$ of the unbounded operator A^* equipped with the graph topology

$$\langle f, g \rangle_{\mathscr{X}} = \langle f, g \rangle + \langle A^* f, A^* f \rangle \tag{9.7}$$

where now $\langle f, g\rangle := \int_0^\infty f(t)g(t)'\,dt$ is the inner product in $\mathscr{X}$. Since an infinitesimal generator such as A^* is a closed operator with a dense domain [48], $\mathscr{Z}$ is a Hilbert space which is densely embedded in $\mathscr{X}$. The topology of $\mathscr{Z}$ is stronger than that of $\mathscr{X}$, and therefore all continuous linear functionals on $\mathscr{X}$ are continuous on $\mathscr{Z}$ as well. Consequently, we can think of the dual space $\mathscr{X}^*$ as embedded in the dual space $\mathscr{Z}^*$. Then, identifying $\mathscr{X}^*$ with $\mathscr{X}$, we have

$$\mathscr{Z} \subset \mathscr{X} \subset \mathscr{Z}^* \tag{9.8}$$

where $\mathscr{Z}$ is dense in $\mathscr{X}$ which in turn is dense in $\mathscr{Z}^*$. We shall write (f, f^*) to denote the value of the functional $f^* \in \mathscr{Z}^*$ evaluated at $f \in \mathscr{Z}$ (or, by reflexivity, the value at f^* of f regarded as a functional on $\mathscr{Z}^*$). Clearly, the bilinear form (f, f^*) coincides with the inner product $\langle f, f^*\rangle$ whenever $f^* \in \mathscr{X}$. Since A^*f is the derivative of f, $\mathscr{Z}$ is a subspace of the Sobolev space $H^1(\mathbb{R}^+)$, and $\mathscr{Z}^*$ is a space of distributions [3].

Next, define $D: \mathscr{Z} \to \mathscr{X}$ to be the differentiation operator on $\mathscr{Z}$. Then $Df = A^*f$ for all $f \in \mathscr{Z}$, but, since $\|Df\| \leqq \|f\|_{\mathscr{Z}}$, D is a continuous map. Its adjoint $D^*: \mathscr{X} \to \mathscr{Z}^*$ is the extension of A to $\mathscr{X}$, because $(f, D^*g) = \langle A^*f, g\rangle$. We collect some well-known properties of D in the following lemma.

LEMMA 9.1. *The map $(I - D): \mathscr{Z} \to \mathscr{X}$ is bijective, and it has a bounded inverse $(I - D)^{-1}: \mathscr{X} \to \mathscr{Z}$. Moreover,*

$$\|f\|_{\mathscr{Z}}^2 \leqq \|(I - D)f\|^2 \leqq 2\|f\|_{\mathscr{Z}}^2. \tag{9.9}$$

Proof. Since $\{U_t(X); t \geqq 0\}$ is a strongly continuous contraction semigroup (Theorem 5.2), then so is $\{e^{A^*t}; t \geqq 0\}$. Consequently, D is dissipative, i.e. $\langle Df, f\rangle \leqq 0$ for all $f \in \mathscr{Z}$, and $(I - D)$ maps $\mathscr{Z}$ onto $\mathscr{X}$ [48, p. 250]. The dissipative property implies that

$$\|(I - D)f\|^2 \geqq \|f\|^2 + \|Df\|^2 \tag{9.10}$$

and therefore $(I - D)$ is also injective. Hence, $(I - D)^{-1}: \mathscr{X} \to \mathscr{Z}$ is defined on all of $\mathscr{X}$, and, due to (9.10), $\|(I - D)^{-1}g\|_{\mathscr{Z}} \leqq \|g\|$, i.e. $(I - D)^{-1}$ is a bounded map. The first of inequalities (9.9) is precisely (9.10), whereas the second follows from the inequality $(a - b)^2 \leqq 2(a^2 + b^2)$. □

We shall construct a shift realization much along the lines of infinite-dimensional deterministic realization theory [5], [6], [14], [15], [19]. Note, however, that, in comparison with this work, our set-up has been transposed. This is necessary in order to obtain the appropriate relation between observability (constructibility) of X and its standard forward (backward) realization, as we shall see below.

Let $f \in \mathscr{Z}$. Since $\mathscr{Z}$ is a *bona fide* function space, we can evaluate f at each point, and consequently, in view of (9.6),

$$f(t) = (e^{A^*t}f)(0). \tag{9.11}$$

Since $\mathscr{Z}$ is a subspace of the Sobolev space $H^1(\mathbb{R}^+)$; the evaluation operator is bounded [3], [16]. However, we want it defined on $\mathscr{X}$, and for this we need the operator $(I - D)$ of Lemma 9.1. Since A^* commutes with e^{A^*t}, then so does $(I - D)$. Therefore, (9.11) yields

$$f(t) = [(I - D)^{-1}\, e^{A^*t}(I - D)f](0). \tag{9.12}$$

Now, recalling that $(I - D)^{-1}$ maps $\mathscr{X}$ onto $\mathscr{Z}$ (Lemma 9.1),

$$B^*g = [(I - D)^{-1}g](0) \tag{9.13}$$

defines a bounded map $B^*: \mathscr{X} \to \mathbb{R}^p$. Let $B: \mathbb{R}^p \to \mathscr{X}$ be its adjoint. Then, (9.12) may be

written

$$f(t) = B^* e^{A^* t}(I - D)f, \tag{9.14}$$

and therefore, if e_k is the kth unit axis vector in $\mathbb{R}^p$,

$$f_k(t) = \langle B^* e^{A^* t}(I - D)f, e_k \rangle_{\mathbb{R}^p} = \langle (I - D)f, e^{At} B e_k \rangle, \qquad k = 1, 2, \cdots, p. \tag{9.15}$$

Together, (9.4) and (9.15) yield, for each $\xi \in I_u \mathscr{Z}$, the representation

$$U_t \xi = \sum_{k=1}^{p} \int_{-\infty}^{t} \langle g, e^{A(t-\sigma)} B e_k \rangle du_k(\sigma) \tag{9.16}$$

where $g = (I - D) I_u^* \xi$. It can be shown that if the structural function K is analytic in some strip $-\alpha < \operatorname{Re}(s) \leqq 0$ of the complex plane, the integral

$$x(t) = \int_{-\infty}^{t} e^{A(t-\sigma)} B \, du(\sigma) \tag{9.17}$$

is well defined [32], and hence it defines an $\mathscr{X}$-valued *state process* $\{x(t);\ t \in \mathbb{R}\}$, i.e. a Hilbert-space-valued process with nuclear covariance operator [16]. If so, (9.16) can be written

$$U_t \xi = \langle g, x(t) \rangle. \tag{9.18}$$

If the integral (9.17) is not well defined, we can interpret the state process $\{x(t);\ t \in \mathbb{R}\}$ as a *generalized stochastic process* in the sense of [17], in which case (9.18) is merely shorthand for (9.16), rather than a *bona fide* inner product.

Note that, when g varies over $\mathscr{X}$, f ranges over $\mathscr{Z}$ (Lemma 9.1), and hence ξ ranges over $I_u \mathscr{Z}$ which is dense in X. Consequently,

$$X = \operatorname{cl} \{\langle g, x(0) \rangle | g \in \mathscr{X}\} \tag{9.19}$$

where cl stands for closure (in the topology of H). This should be compared with (8.17) in the finite-dimensional case, of which it is a generalization: recall that the state space $\mathbb{R}^n$ corresponds to $\mathscr{X}$ here.

It is important to note that we must take closure in (9.19). This means that processes with components of type $\{U_t \xi;\ t \in \mathbb{R}\}$ can be represented as outputs of a stochastic dynamical system with state process $\{x(t);\ t \in \mathbb{R}\}$ if and only if $\xi \in I_u \mathscr{Z}$, which is only a dense subset of X. Therefore, in particular, we must have

$$y_k(0) \in I_u \mathscr{Z} \quad \text{for } k = 1, 2, \cdots, m \tag{9.20}$$

in order to have y as an output. This condition can be characterized in the following ways.

PROPOSITION 9.1. *Let X be a proper Markovian splitting subspace with analytic spectral factor W. Let w denote the inverse Fourier transform of W and Γ the infinitesimal generator of $\{U_t(X);\ t \geqq 0\}$. Then, the following conditions are equivalent to* (9.20).

(i) $y_k(0) \in \mathscr{D}(\Gamma)$ *for* $k = 1, 2, \cdots, m$.

(ii) *The rows* $w_1, w_2, \cdots, w_m$ *of* w *belong to* $\mathscr{Z}$.

(iii) *The rows of* $i\omega W(i\omega) - N$ *belong to* $\mathscr{H}_p^2$ *for some constant* $m \times p$ *matrix* N.

Proof. First note that, by construction, $\mathscr{D}(\Gamma) = I_u \mathscr{Z}$, and therefore (9.20) and (i) are the same. Since $I_u = I_{\hat{u}} \mathscr{F}$, it follows from (6.12)+(6.17) that

$$w_k = I_u^* y_k(0) \quad \text{for } k = 1, 2, \cdots, m. \tag{9.21}$$

Hence the equivalence of (i) and (ii) is immediate; that of (ii) and (iii) follows from [23, Lemma 3.1]. □

If the conditions of Proposition 9.1 are satisfied, the inner products

$$(Cg)_k = \langle (I - D)w_k, g\rangle, \qquad k = 1, 2, \cdots, m \tag{9.22}$$

are well defined, and, they define a bounded operator $C: \mathscr{X} \to \mathbb{R}^p$ such that

$$y(t) = \int_{-\infty}^{t} C\, e^{A(t-\sigma)} B\, du(\sigma), \tag{9.23}$$

as can be seen from (9.16) and (9.21). If the integral (9.17) is well defined, this may be written

$$y(t) = Cx(t); \tag{9.24}$$

otherwise we may interpret (9.24) in the generalized sense mentioned above, i.e. simply as (9.23). We shall call (9.23) the *standard forward realization* of X.

How natural are the conditions of Proposition 9.1? For any (forward) stochastic realization

$$\begin{aligned} dx &= Ax\, dt + B\, du, \\ y &= Cx \end{aligned} \tag{9.25}$$

with x a strong solution, we must have

$$E^{H^-(du)} a[y(h) - y(0)] = \int_0^h E^{H^-(du)} aCAx(t)\, dt \tag{9.26}$$

for any row vector $a \in \mathbb{R}^m$ and $h \geqq 0$. Using (5.3a), it is easy to see that this implies $\|[U_t(X) - I]y(0)\| \leqq kh$ and hence, as in [33], Condition (i) of Proposition 9.1, providing a justification for this condition. However, it should be noted that, even if (9.17) is well defined, it is not automatically true that x is a strong solution of the stochastic differential equation in (9.25) [8].

Next, we shall investigate the systems-theoretical properties of the realization (9.23). Let us begin with reachability. Recall that (9.23) is *reachable* if $\bigcap_{t \geqq 0} \ker B^* e^{A^* t} = 0$ [14]. But, in view of (9.14), $B^* e^{A^* t} g = 0$ for all $t \geqq 0$ if and only if $f := (I - D)^{-1} g$ is identically zero, i.e. if and only if $g = 0$. Hence, (9.23) is reachable.

The realization (9.23) is said to be *observable* if $\bigcap_{t \geqq 0} \ker C e^{At} = 0$ [14]. To determine if this holds, form

$$(C e^{At} g)_k = \langle (I - D)w_k, e^{At} g\rangle = \langle (I - D)\, e^{A^* t} w_k, g\rangle \tag{9.27}$$

where we have used the fact that D and $e^{A^* t}$ commute. Define the vector space

$$\mathscr{M} := \operatorname{span}\{e^{A^* t} w_k;\ t \geqq 0,\ k = 1, 2, \cdots, m\}. \tag{9.28}$$

From (9.27) it follows then that (9.23) is observable if and only if $(I - D)\mathscr{M}$ is dense in $\mathscr{X}$. However, in view of (9.9), this is equal to $\mathscr{M}$ being dense in $\mathscr{Z}$ (in $\mathscr{Z}$ topology).

On the other hand, $X \sim (S, \bar{S})$ is an observable splitting subspace if and only if the vector space

$$M := \operatorname{span}\{E^S y_k(t);\ t \geqq 0,\ k = 1, 2, \cdots, m\} \tag{9.29}$$

is dense in X (Corollary 4.3). In view of Theorem 5.2, $E^S y_k(t) = U_t(X) y_k(0)$, and therefore, by (9.1) and (9.21), $M = I_u \mathscr{M}$.

Now, suppose that (9.23) is observable. Then $\mathscr{M}$ is dense in $\mathscr{Z}$ and hence in $\mathscr{X}$ (weaker topology). Consequently, M is dense in X, i.e. X is observable. Next, suppose that X is observable. Then M is dense in X, and hence $\mathscr{M}$ is dense in $\mathscr{X}$. Therefore,

we have the situation

$$\mathscr{M} \subset \mathscr{Z} \subset \mathscr{X} \tag{9.30}$$

where the vector space $\mathscr{M}$ is dense in the Hilbert space $\mathscr{X}$. Since the topology of $\mathscr{Z}$ is stronger than that of $\mathscr{X}$, (9.30) does not automatically imply that $\mathscr{M}$ is dense in $\mathscr{Z}$ as required; $\mathscr{Z}$ is said to be *normal* if this favorable situation occurs [3, p. 101]. However, it can be shown that the dissipative property of $\{e^{A^*t};\ t \geqq 0\}$ implies that $\mathscr{Z}$ is normal [32]. Consequently, the realization (9.23) is observable if and only if X is observable.

We collect these observations in the following theorem.

THEOREM 9.1. [32]. *Let X be a proper Markovian splitting subspace with forward generating process u, and let $\mathscr{X} := I_u X$. Then*

$$X = \mathrm{cl}\left\{ \sum_{k=1}^{p} \int_{-\infty}^{t} \langle g, e^{A(t-\sigma)} B e_k \rangle \, du_k(\sigma) \,|\, g \in \mathscr{X} \right\} \tag{9.31}$$

where $\{e^{At};\ t \geqq 0\}$ is the strongly continuous contraction semigroup on $\mathscr{X}$ defined by (9.1) *and $B: \mathbb{R}^p \to \mathscr{X}$ is the adjoint of* (9.13). *If the structural function of X is analytic in some strip $-\alpha < \mathrm{Re}\,(s) \leqq 0$ of the complex plane, the integral*

$$x(t) = \int_{-\infty}^{t} e^{A(t-\sigma)} B \, du(\sigma) \tag{9.32}$$

is well defined and defines an $\mathscr{X}$-valued random process $\{x(t);\ t \in \mathbb{R}\}$ in terms of which (9.31) *can be written*

$$X = \mathrm{cl}\,\{\langle g, x(0) \rangle | g \in \mathscr{X}\}. \tag{9.33}$$

If the conditions of Proposition 9.1 *are satisfied, there is a map $C: \mathscr{X} \to \mathbb{R}^m$, defined by* (9.22), *such that*

$$y(t) = \int_{-\infty}^{t} C e^{A(t-\sigma)} B \, du(\sigma) \tag{9.34}$$

which, in the case that (9.32) *is well defined, yields*

$$y(t) = Cx(t). \tag{9.35}$$

This is a reachable forward realization which is observable if and only if X is observable.

The construction of the corresponding backward realization is analogous, exchanging $\mathbb{R}^+$ for $\mathbb{R}^-$ everywhere. Let $\bar{\mathscr{Z}}$ be $\mathscr{D}(\bar{A}^*)$ equipped with graph topology, and let $\bar{D}: \bar{\mathscr{Z}} \to \bar{\mathscr{X}}$ be the (bounded) differentiation operator on $\bar{\mathscr{Z}}$. Let $\langle \cdot, \cdot \rangle$ denote the inner product in $\bar{\mathscr{X}}$. Then, we can proceed as above to obtain, for each $\xi \in I_{\bar{u}} \bar{\mathscr{Z}}$, the representation

$$U_t \xi = \sum_{k=1}^{p} \int_{t}^{\infty} \langle g, e^{\bar{A}(\sigma - t)} \bar{B} e_k \rangle \, d\bar{u}_k(\sigma) \tag{9.36}$$

where $g = (I - \bar{D})^{-1} I_{\bar{u}}^* \xi$, and $\bar{B}: \mathbb{R}^p \to \bar{\mathscr{X}}$ is the adjoint of

$$\bar{B}^* g = [(I - \bar{D})^{-1} g](0). \tag{9.37}$$

Now, if one of the three equivalent conditions

(9.38)
(i) $y_k(0) \in \mathscr{D}(\Gamma^*)$, $k = 1, 2, \cdots, m$,
(ii) $\bar{w}_k := I_{\bar{u}}^* y_k(0) \in \bar{\mathscr{Z}}$, $k = 1, 2, \cdots, m$,
(iii) the rows of $i\omega \bar{W}(i\omega) - \bar{N}$ belong to $\bar{\mathscr{H}}_p^2$ for some constant $m \times p$ matrix $\bar{N}$

hold, we may define a bounded linear operator $\bar{C}: \bar{\mathscr{X}} \to \mathbb{R}^m$ by the relations

$$(\bar{C}g)_k = \langle (I - \bar{D})\bar{w}_k, g\rangle, \qquad k = 1, 2, \cdots, m \tag{9.39}$$

and then we have the *standard backward realization*

$$y(t) = \int_t^\infty \bar{C}\, e^{\bar{A}(\sigma - t)} \bar{B}\, d\bar{u}(\sigma). \tag{9.40}$$

Following the convention set up in §§ 1 and 8, we shall say that (9.40) is *controllable* if $\bigcap_{t \geqq 0} \ker \bar{B}^* e^{\bar{A}^* t} = 0$ and *constructible* if $\bigcap_{t \geqq 0} \ker \bar{C} e^{\bar{A} t} = 0$. It is then easy to check that Theorem 9.1 has the following "backward" version.

THEOREM 9.2. [32]. *Let X be a proper Markovian splitting subspace with backward generating process $\bar{u}$. Set $\bar{\mathscr{X}} := I_{\bar{u}} X$. Then*

$$X = \operatorname{cl}\left\{ \sum_{k=1}^{p} \int_t^\infty \langle g, e^{\bar{A}(\sigma - t)} \bar{B} e_k \rangle\, du_k(\sigma) \,|\, g \in \bar{\mathscr{X}} \right\} \tag{9.41}$$

where $\{e^{\bar{A}t}; t \geqq 0\}$ is the strongly continuous contraction semigroup (9.2) on $\bar{\mathscr{X}}$, and $\bar{B}: \mathbb{R}^p \to \bar{\mathscr{X}}$ is the adjoint of (9.37). If the structural function of X is analytic in some strip $-\alpha < \operatorname{Re}(s) \leqq 0$ of the complex plane, there is an $\bar{\mathscr{X}}$-valued random process $\{\bar{x}(t); t \in \mathbb{R}\}$ defined by

$$\bar{x}(t) = \int_t^\infty e^{\bar{A}(\sigma - t)} \bar{B}\, d\bar{u}(\sigma) \tag{9.42}$$

so that (9.41) may be written

$$X = \operatorname{cl}\{\langle g, \bar{x}(0)\rangle | g \in \bar{\mathscr{X}}\}. \tag{9.43}$$

Moreover, if the conditions (9.38) hold, there is a map $\bar{C}: \bar{\mathscr{X}} \to \mathbb{R}^p$, defined by (9.39), such that (9.40) holds, and hence, if (9.42) is well defined,

$$y(t) = \bar{C}\bar{x}(t). \tag{9.44}$$

This is a controllable backward realization which is constructible if and only if X is constructible.

Consequently, for X to have both a forward and backward realization we must have

$$y_k(0) \in \mathscr{D}(\Gamma) \cap \mathscr{D}(\Gamma^*), \qquad k = 1, 2, \cdots, m. \tag{9.45}$$

Questions of this sort are studied in [33].

10. State space isomorphism. There is an important difference between stochastic and deterministic realization theory which manifests itself already in the finite-dimensional case. In the deterministic theory, there is an essentially unique minimal realization (modulo trivial coordinate transformations). This is not the case in the stochastic theory. Two different minimal Markovian splitting subspaces give rise to realizations with probabilistically different state processes. Therefore, it is important to investigate the relationship between realizations of different minimal X.

In this section we shall study the class of standard forward realizations (9.23) of minimal Markovian splitting subspaces; the corresponding results for backward realizations are analogous and will not be mentioned. Our main goal is to clarify the connections between triplets (A, B, C) of such forward realizations. In the finite-dimensional case, this link is provided by the Yakubovic-Kalman-Popov or Positive Real Lemma, to which we shall return below.

For the rest of the paper we shall *assume that y is strictly noncyclic.* Then the class of minimal Markovian splitting subspaces can be parameterized by the left inner divisors Q of Q_+ (Theorem 7.6), and this parametrization, denoted $\{X_Q; Q|_L Q_+\}$, induces a lattice structure on the class under which $X_{Q_2} < X_{Q_1}$ if and only if $Q_2|_L Q_1$; see § 7.

Let $(K_1, Q_1, \bar{Q}_1^*)$ and $(K_2, Q_2, \bar{Q}_2^*)$ be the inner triplets of two minimal Markovian splitting subspaces, X_{Q_1} and X_{Q_2}. Then, it follows from (7.15) that

$$\bar{Q}_2^* \bar{Q}_1 K_1 = K_2 Q_2^* Q_1. \tag{10.1}$$

LEMMA 10.1. *The following statements are equivalent.*

(i) $X_{Q_2} < X_{Q_1}$.

(ii) $V_1 := Q_2^* Q_1$ *is inner.*

(iii) $V_2 := \bar{Q}_2^* \bar{Q}_1$ *is inner.*

If these conditions are satisfied, then

$$K_1 V_1^* = V_2^* K_2 \tag{10.2}$$

with K_1 and V_1 right coprime and K_2 and V_2 left coprime.

Proof. Let $X_{Q_1} \sim (S_1, \bar{S}_1)$ and $X_{Q_2} \sim (S_2, \bar{S}_2)$. Then, by Lemma 6.1, (ii) is equivalent to

$$S_2 \subset S_1 \tag{10.3}$$

and (iii) is equivalent to $\bar{S}_2^\perp \subset \bar{S}_1^\perp$, i.e.

$$\bar{S}_2 \supset \bar{S}_1. \tag{10.4}$$

But, since X_{Q_1} and X_{Q_2} are minimal, (10.3) and (10.4) are equivalent (Corollary 3.3), establishing the equivalence of (ii) and (iii). Now, $Q_1 = Q_2 V_1$. Hence (i) and (ii) are equivalent, and, since K_1 and Q_1 are right coprime (Theorem 7.2 and Corollary 4.1), then so are K_1 and V_1. Likewise, since $\bar{Q}_2^* = V_2 \bar{Q}_1^*$, the left coprimeness of K_2 and V_2 follows from that of K_2 and $\bar{Q}_2^*$ (Theorem 7.2 and Corollary 4.1). Relation (10.2) is the same as (10.1). □

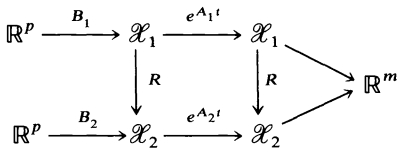
The following theorem describes the intertwining of the triplets (A_1, B_1, C_1) and (A_2, B_2, C_2) corresponding to two minimal Markovian splitting subspaces, X_1 and X_2, which are ordered.

THEOREM 10.1. *Let X_1 and X_2 be two minimal Markovian splitting subspaces such that $X_2 < X_1$, and let Σ_1 and Σ_2 be the corresponding standard forward realizations with state spaces $\mathscr{X}_1$ and $\mathscr{X}_2$. Then the map $R: \mathscr{X}_1 \to \mathscr{X}_2$ defined by*

$$Rf = P^{\mathscr{X}_2} \mathscr{F} M_{Q_1^* Q_2} \mathscr{F}^* f \tag{10.5}$$

is injective with dense range, and the following diagram commutes,

(10.6)



where indices refer to Σ_1 and Σ_2.

Proof. Let K_1 and K_2 be the structural functions of X_1 and X_2, and let $\Sigma_t(K_i): \mathscr{H}(K_i) \to \mathscr{H}(K_i)$ be the restricted shifts

$$\Sigma_t(K_i) f = P^{\mathscr{H}(K_i)} \chi_t f \tag{10.7}$$

for $t \geqq 0$ and $i = 1, 2$. Since there are inner functions V_1 and V_2 such that $V_2 K_1 = K_2 V_1$ (Lemma 10.1), there is a map $\hat{R}^*: \mathscr{H}(K_2) \to \mathscr{H}(K_1)$ such that

$$\hat{R}^* \Sigma_t(K_2) = \Sigma_t(K_1) \hat{R}^* \tag{10.8}$$

[14, Thm. 14.8, p. 203]. This map is given by

$$\hat{R}^* f = P^{\mathscr{H}(K_1)} M_{Q_2^* Q_1} f \tag{10.9}$$

and, in view of the coprimeness conditions of Lemma 10.1, $\hat{R}^*$ is injective with dense range [14, Thm. 14.11, p. 206]. Therefore the same is true for the adjoint $\hat{R}: \mathscr{H}(K_1) \to \mathscr{H}(K_2)$ and for $R := \mathscr{F}^* \hat{R} \mathscr{F}: \mathscr{X}_1 \to \mathscr{X}_2$, which is the map of the theorem. It remains to show that the diagram commutes. To this end, first note that $e^{A_i^* t} = \mathscr{F}^* \Sigma_t(K_i) \mathscr{F}$ for $i = 1, 2$, and therefore (10.8) is equivalent to

$$R e^{A_1 t} = e^{A_2 t} R. \tag{10.10}$$

Then the same intertwining must hold for the resolvents, i.e. in particular

$$R(I - A_1)^{-1} = (I - A_2)^{-1} R \tag{10.11}$$

(Lemma 9.1), and therefore

$$(I - A_1^*) R^* = R^* (I - A_2^*). \tag{10.12}$$

Now, if W_1 and W_2 are the analytic spectral factors of X_1 and X_2, then $W_1 = W_2 Q_2^* Q_1$. But, in view of (8.5), $a W_1 \in \mathscr{H}(K_1)$ for any row vector $a \in \mathbb{R}^m$, and hence $a W_1 = \hat{R}^* a W_2$. Consequently

$$a w_1 = R^* a w_2 \tag{10.13}$$

where $w_1 := \mathscr{F}^* W_1$ and $w_2 := \mathscr{F}^* W_2$. Now, from the definition (9.22) it is easy to see that

$$C_i^* a = (I - A_i^*) a w_i \tag{10.14}$$

for $i = 1, 2$. (Recall that $A^* f = Df$.) Consequently, in view of (10.12) and (10.13), $C_1^* = R^* C_2^*$, i.e.

$$C_1 = C_2 R. \tag{10.15}$$

This together with (10.10) proves that the diagram commutes. □

The parts of diagram (10.6) involving B_1 and B_2 add nothing to the theorem but have been added to remind the reader that the two horizontal chains of arrows realize different functions, namely w_1 and w_2. This situation differs of course from that in the deterministic "state space isomorphism" theorems [22, p. 258].

A map which is injective with dense range such as R in Theorem 10.1 will be called *quasi-invertible*. In the finite-dimensional case, this is the same as invertible, and therefore, in this case, the condition $X_2 < X_1$ of Theorem 10.1 is unnecessary, for we have also a diagram with the arrows reversed. In particular, the semigroups $\{e^{A_1 t};\ t \geqq 0\}$ and $\{e^{A_2 t};\ t \geqq 0\}$ are then similar.

In the infinite-dimensional situation, a natural generalization of similarity is quasisimilarity. We say that the semigroups $\{e^{A_1 t};\ t \geqq 0\}$ and $\{e^{A_2 t};\ t \geqq 0\}$ are *quasisimilar* if there are quasi-invertible maps $R_1: \mathscr{X}_1 \to \mathscr{X}_2$ and $R_2: \mathscr{X}_2 \to \mathscr{X}_1$ such that

$$\begin{aligned} R_1 e^{A_1 t} &= e^{A_2 t} R_1, \\ R_2 e^{A_2 t} &= e^{A_1 t} R_2. \end{aligned} \tag{10.16}$$

Only the first of relations (10.16) is given by Theorem 10.1, and then only if $X_2 < X_1$. If we also had the other, the ordering assumption would be unnecessary also in the

infinite-dimensional case, for quasisimilarity is an equivalence relation [14, p. 74]. That this favorable situation actually happens follows from the next theorem, the proof of which can be found in [31].

THEOREM 10.2. [31]. *Let Σ_1 and Σ_2 be the forward standard realizations corresponding to two arbitrary minimal Markovian splitting subspaces. Then the corresponding semigroups $\{e^{A_1 t};\ t \geqq 0\}$ and $\{e^{A_2 t};\ t \geqq 0\}$ are quasisimilar, i.e. they satisfy* (10.16).

This implies that, as far as the rectangular part of the diagram (10.6) is concerned, the ordering condition $X_2 < X_1$ can be dispensed with. Whether this is true for the diagram as a whole is as yet an open question.

By [14, Thm. 15.18, p. 220], the semigroups are quasisimilar if and only if the corresponding structural functions K_1 and K_2 are quasi-equivalent, i.e. have the same invariant factors, and therefore Theorem 10.2 is equivalent to Theorem 7.5. This allows us to draw the conclusion that the infinitesimal generators A corresponding to minimal Markovian splitting subspaces have the same eigenvalues. To see this, just note that these eigenvalues are the poles of the common determinant of the structural functions [23, Thm. 3.2, p. 70], [14, Thm. 13.8, p. 195].

Theorem 10.2 can also be stated in terms of Jordan models. For a discussion of this concept, see, for example, [14, p. 214].

COROLLARY 10.1. [31]. *All semigroups $\{e^{At};\ t \geqq 0\}$ corresponding to minimal Markovian splitting subspaces have the same Jordan model, i.e. they are all quasisimilar to the direct sum*

$$\Sigma_t(k_1) \oplus \Sigma_t(k_2) \oplus \cdots \oplus \Sigma_t(k_p) \tag{10.17}$$

where $k_1, k_2, \cdots, k_p$ are the common invariant factors of the structural functions, and the restricted shifts $\Sigma_t(k_i)$, $i = 1, 2, \cdots, p$ and $t \geqq 0$, are defined as in (10.7) *but for a scalar Hardy space.*

As an application of Theorem 10.1, we shall next derive an infinite-dimensional version of the Positive Real Lemma equations. For this we shall need the following two lemmas.

LEMMA 10.2. *Let A and B be defined by* (9.1) *and* (9.13). *Then*

$$AP + PA^* + BB^* = 0 \tag{10.18}$$

where $P: \mathscr{X} \to \mathscr{X}$ is the positive self-adjoint operator

$$P = (I - A)^{-1}(I - A^*)^{-1}. \tag{10.19}$$

Proof. Let $f_i \in \mathscr{Z}$, $i = 1, 2$. Then, recalling that $A^* f = Df$ for $f \in \mathscr{Z}$, where D is the differentiation operator, integration by parts yields

$$\langle A^* f_1, f_2 \rangle + \langle f_1, A^* f_2 \rangle = \int_0^\infty (\dot{f}_1 f_2' + f_1 \dot{f}_2')\, dt = -f_1(0) f_2(0)'. \tag{10.20}$$

Also, by the definition (9.13),

$$\begin{aligned} \langle (I - A^*) f_1, BB^*(I - A^*) f_2 \rangle &= \langle B^*(I - A^*) f_1, B^*(I - A^*) f_2 \rangle_{\mathbb{R}^p} \\ &= f_1(0) f_2(0)'. \end{aligned} \tag{10.21}$$

Now, let $g_i \in \mathscr{X}$, $i = 1, 2$, be arbitrary. Then, by Lemma 9.1, $f_i := (I - A^*)^{-1} g_i \in \mathscr{Z}$ for $i = 1, 2$. Inserting this into (10.20) and (10.21) and adding the relations, we obtain

$$\langle g_1, APg_2 \rangle + \langle g_1, PA^* g_2 \rangle + \langle g_1, BB^* g_2 \rangle = 0 \tag{10.22}$$

where we have used the fact that A^* and $(I - A^*)^{-1}$ commute. This yields (10.18). □

The operator P is actually the *state covariance operator* in the sense that

$$E\{\langle g_1, x(0)\rangle\langle g_2, x(0)\rangle\} = \langle g_1, Pg_2\rangle. \tag{10.23}$$

To see this, note that, by (9.4) and (9.18),

$$\langle g, x(0)\rangle = \int_{-\infty}^{0} [(I-A^*)^{-1}g](-\sigma)\,du(\sigma) \tag{10.24}$$

where, in general, the left member should be understood in the sense of (9.16). In passing, we recall that the state process $\{x(t);\, t\in\mathbb{R}\}$ is a *bona fide* $\mathscr{X}$-valued random process if and only if the operator P is nuclear [16].

LEMMA 10.3. *Let* $\Lambda:\mathbb{R}^+\to\mathbb{R}^{m\times m}$ *be the covariance*

$$\Lambda(t) = E\{y(t)y(0)'\}, \qquad t\geqq 0 \tag{10.25}$$

and let A *and* C *be defined by* (9.1) *and* (9.22). *Then*

$$\Lambda(t) = C\,e^{At}PC^* \tag{10.26}$$

where P *is the state covariance operator* (10.19).

Proof. Since $C^*a = (I-A^*)aw$ for any row vector $a\in\mathbb{R}^m$, and $(I-A)^{-1}$ commutes with e^{At}, we have

$$C\,e^{At}PC^*a = C(I-A)^{-1}\,e^{At}aw, \tag{10.27}$$

and therefore

$$[C\,e^{At}PC^*]_{kj} = \langle (I-A^*)w_k, (I-A)^{-1}\,e^{At}w_j\rangle_{\mathscr{X}} = \langle w_k, e^{At}w_j\rangle_{\mathscr{X}}. \tag{10.28}$$

But $\mathscr{F}^*\,e^{At}\mathscr{F} = \Sigma_t(K)^*$ and $W_k = \mathscr{F}w_k$. Hence

$$[C\,e^{At}PC^*]_{kj} = \langle W_k, P^{\mathscr{H}(K)}e^{-i\omega t}W_j\rangle_{\mathscr{H}(K)} = \langle W_k, e^{-i\omega t}W_j\rangle_{\mathscr{L}^2_{p}(\mathbb{I})} \tag{10.29}$$

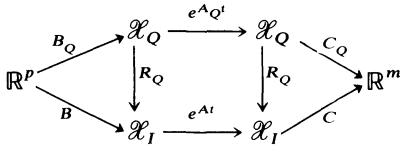
because, by (8.5), $W_k\in\mathscr{H}(K)$. Consequently, (10.26) follows from the Bochner representation

$$\Lambda(t) = \int_{-\infty}^{\infty} e^{i\omega t}\Phi(i\omega)\,d\omega. \tag{10.30}$$

□

To compare the standard forward realizations of different minimal Markovian splitting subspaces, we must reduce them to the same state space. In view of the ordering condition of Theorem 10.1, the most suitable choice of common state space is $\mathscr{X}_I$, the state space of the minimal element $X_I := H^{+/-}$ of the lattice. Given the standard forward realization Σ_Q of an arbitrary minimal Markovian splitting subspace X_Q, the reduction will be according to the diagram

(10.31)



defining a new (A, B, C) for X_Q which has state space $\mathscr{X}_I$. Here R_Q is the map (10.5) with $Q_1 := Q$ and $Q_2 := I$. Then, when X_Q varies over the lattice of minimal Markovian splitting subspaces, $A := A_I$ and $C := C_I$ are fixed, whereas B varies.

THEOREM 10.2. *Let X_Q be an arbitrary minimal Markovian splitting subspace, and let (A, B, C) be defined by* (10.31). *Then*

$$\begin{aligned} &AP + PA^* + BB^* = 0, \\ &PC^* = G \end{aligned} \tag{10.32}$$

where the positive self-adjoint operator $P: \mathscr{X}_I \to \mathscr{X}_I$, defined by

$$P := (I - A)^{-1} R_Q R_Q^* (I - A^*)^{-1} \tag{10.33}$$

is the state covariance operator in the fixed state space representation, and $G: \mathbb{R}^m \to \mathscr{X}$ is given by

$$G := (I - A)^{-1}(I - A^*)^{-1} C^*. \tag{10.34}$$

Proof. By Lemma 10.2, $A_Q P_Q + P_Q A_Q^* + B_Q B_Q^* = 0$, where $P_Q := (I - A_Q)^{-1} \times (I - A_Q^*)^{-1}$. Transforming this via (10.31) and (10.11) yields the first of relations (10.32). To derive the second relation (10.32), reduce the representation (10.26) to the fixed state space $\mathscr{X}_I$. Comparing the expressions for $\Lambda(t)$ thus obtained corresponding to Σ_Q and Σ_I respectively, we have

$$C e^{At}[PC^* - P_I C^*] = 0 \quad \text{for all } t \geqq 0. \tag{10.35}$$

Since Σ_I is observable (Theorem 9.1), this implies that $PC^* = P_I C^*$, which is precisely G. □

We have thus shown that all standard forward realizations $\{\Sigma_Q; Q|_L Q_+\}$ reduced to the common fixed state space $\mathscr{X}_I$ satisfy equations akin to those of the Positive Real Lemma [2], [11], [12]. Note, however, that in our case the representation is coordinate-free.

REFERENCES

[1] H. AKAIKE, *Markovian representation of stochastic processes by canonical variables*, SIAM J. Control, 13 (1975), pp. 165–173.

[2] B. D. O. ANDERSON, *The inverse problem of stationary covariance generation*, J. Statist. Phys., 1 (1969), pp. 133–147.

[3] J.-P. AUBIN, *Applied Functional Analysis*, Wiley-Interscience, New York, 1979.

[4] F. BADAWI, A. LINDQUIST AND M. PAVON, *A stochastic realization approach to the smoothing problem*, IEEE Trans. Automat. Control, 24 (1979), pp. 878–888.

[5] J. S. BARAS AND R. W. BROCKETT, *H^2-functions and infinite-dimensional realization theory*, SIAM J. Control, 13 (1975), pp. 221–241.

[6] J. S. BARAS AND P. DEWILDE, *Invariant subspaces methods in linear multivariable distributed systems and lumped distributed network synthesis*, Proc. IEEE, 64 (1976), pp. 160–178.

[7] A. BENSOUSSAN, *Filtrage optimal des systèmes linéaires*, Dunod, Paris, 1971.

[8] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems*, Springer-Verlag, Berlin, 1978.

[9] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1953.

[10] H. DYM AND H. P. MCKEAN, *Gaussian Processes, Function Theory, and the Inverse Spectral Problem*, Academic Press, New York, 1976.

[11] P. FAURRE, *Réalisations markoviennes de processus stationnaires*, Research Report No. 13, March 1973, IRIA Laboria, Le Chesney, France.

[12] P. FAURRE, M. CLERGET AND F. GERMAIN, *Opérateurs rationnels positifs*, Dunod, Paris, 1979.

[13] G. D. FORNEY, JR., *Minimal bases of rational vector spaces, with applications to multivariable linear systems*, SIAM J. Control, 13 (1975), pp. 453–520.

[14] P. A. FUHRMANN, *Linear Operators and Systems in Hilbert Space*, McGraw-Hill, New York, 1981.

[15] ———, *On realization of linear systems and applications to some questions in stability*, Math. Systems Theory, 8 (1974), pp. 132–141.

[16] I. M. GELFAND AND N. YA. VILENKIN, *Generalized Functions*, Vol. 4, Academic Press, New York, 1964.
[17] I. GIKHMAN AND A. V. SKOROKHOD, *Introduction to the Theory of Random Processes*, Saunders, Philadelphia, 1965.
[18] H. HELSON, *Lectures on Invariant Subspaces*, Academic Press, New York, 1976.
[19] J. W. HELTON, *Systems with infinite-dimensional state space: The Hilbert space approach*, Proc. IEEE, 64 (1976), pp. 145-160.
[20] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[21] R. E. KALMAN, *Linear stochastic filtering-reappraisal and outlook*, Proc. Symp. System Theory, Polytechnic Institute of Brooklyn, 1965, pp. 197-205.
[22] R. E. KALMAN, P. L. FALB AND M. A. ARBIB, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1979.
[23] P. D. LAX AND R. S. PHILLIPS, *Scattering Theory*, Academic Press, New York, 1967.
[24] A. LINDQUIST, M. PAVON AND G. PICCI, *Recent trends in stochastic realization theory*, in Prediction Theory and Harmonic Analysis: The Pesi Masani Volume, V. Mandrekar and H. Salehi, eds., North-Holland, Amsterdam, 1983.
[25] A. LINDQUIST AND G. PICCI, *On the stochastic realization problem*, SIAM J. Control Optim., 17 (1979), pp. 365-389.
[26] ———, *A state-space theory for stationary stochastic processes*, Proc. 21st Midwest Symposium on Circuits and Systems, Ames, IO, August, 1978.
[27] ———, *A Hardy space approach to the stochastic realization problem*, Proc. 1978 Conference on Decision and Control, San Diego, pp. 933-939.
[28] ———, *Realization theory for multivariate stationary gaussian processes* I: *State space construction*, Proc. 4th International Symposium on the Mathematical Theory of Networks and Systems, July 1979, Delft, Holland, pp. 140-148.
[29] ———, *Realization theory for multivariate stationary gaussian processes* II: *State space theory revisited and dynamical representations of finite dimensional state spaces*, in Proc. 2nd International Conference on Information Sciences and Systems, Patras, Greece, July 1979, Reidel, Dordrecht, pp. 108-129.
[30] ———, *State space models for gaussian stochastic processes*, in Stochastic Systems: The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, 1981.
[31] ———, *On a condition for minimality of Markovian splitting subspaces*, Systems Control Lett., 1 (1982), pp. 264-269.
[32] ———, *Infinite dimensional stochastic realizations of continuous-time stationary vector processes*, Topics in Operators and Systems, H. Dym and I. Gohberg, eds., Birkhäuser Verlag, Berlin, 1984.
[33] ———, *Forward and backward semimartingale representations for stationary increment processes*, Stochastics, to appear.
[34] H. P. MCKEAN, *Brownian motion with several dimensional time*, Theory Prob. Appl., VIII (1963), pp. 353-354.
[35] J. NEVEU, *Processus aléatoires gaussiens*, Les Presses de l'Université de Montréal, Montreal, 1968.
[36] M. PAVON, *Stochastic realization and invariant directions of the matrix Riccati equation*, SIAM J. Control Optim., 18 (1980), pp. 155-180.
[37] G. PICCI, *Stochastic realization of Gaussian processes*, Proc. IEEE, 64 (1976), pp. 112-122.
[38] YU. A. ROZANOV, *Stationary Random Processes*, Holden-Day, New York, 1967.
[39] G. RUCKEBUSCH, *Répresentations markoviennes de processus gaussiens stationnaires*, Ph.D. thesis, Univ. Paris VI, 1975.
[40] ———, *Répresentations markoviennes de processus gaussiens stationnaires*, C.R. Acad. Sc. Paris, Ser. A, 282 (1976), pp. 649-651.
[41] ———, *Factorisations minimales de densités spectrales et répresentations markoviennes*, Proc. 1re Colloque AFCET-SMF, Palaiseau, France, 1978.
[42] ———, *A state space approach to the stochastic realization problem*, Proc. 1978 International Symposium on Circuits and Systems, New York.
[43] ———, *Théorie géométrique de la répresentation markovienne*, Thèse de doctorat d'etat, Univ. Paris VI, 1980.
[44] ———, *On the structure of minimal Markovian representations*, Nonlinear Stochastic Problems, R. Bucy and J. M. F. Moura, eds., Reidel, Dordrecht, 1983.
[45] B. SZ.-NAGY AND C. FOIAS, *Harmonic Analysis of Operators on Hilbert Space*, North-Holland, Amsterdam, 1970.
[46] S. MACLANE AND G. BIRKHOFF, *Algebra*, Macmillan, New York, 1967.
[47] W. A. WOLOVICH, *Linear Multivariate Systems*, Springer-Verlag, Berlin, 1974.

[48] K. YOSIDA, *Functional Analysis*, Springer-Verlag, Berlin, 1965.
[49] F. BADAWI, A. LINDQUIST AND M. PAVON, *A stochastic realization approach to the smoothing problem*, IEEE Trans. Automat. Control, 24 (1979), pp. 878–888.
[50] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, Berlin, 1976.
[51] M. PAVON, *New results on the interpolation problem for continuous-time stationary increment processes*, SIAM J Control Optim., 22 (1984), pp. 133–142.
[52] J. C. WILLEMS, *System theoretic models for the analysis of physical systems*, Ricerche di Automatica, 10 (1979), pp. 71–104.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 6, November 1985

ADDITIVE CONTROL OF STOCHASTIC LINEAR SYSTEMS WITH FINITE HORIZON*

PAO-LIU CHOW†, JOSÉ-LUIS MENALDI† AND MAURICE ROBIN‡

Abstract. We consider a dynamic system whose state is governed by a linear stochastic differential equation with time-dependent coefficients. The control acts additively on the state of the system. Our objective is to minimize an integral cost which depends upon the evolution of the state and the total variation of the control process. It is proved that the optimal cost is the unique solution of an appropriate free boundary problem in a space-time domain. By using some decomposition arguments, the problems of a two-sided control, i.e. optimal corrections, and the case with constraints on the resources, i.e. finite fuel, can be reduced to a simpler case of only one-sided control, i.e. a monotone follower. These results are applied to solving some examples by the so-called method of similarity solutions.

Key words. dynamic programming, stochastic processes, free boundary problems, degenerate second order parabolic equations

Introduction. In this paper, we wish to control a linear stochastic differential equation in the sense of Itô by using additive strategies, i.e. the evolution of the state is subjected to

$$y(s) = x + \nu(s-t) + \int_t^s (a(\lambda)y(\lambda) + b(\lambda))\, d\lambda + \int_t^s \sigma(\lambda)\, dw(\lambda - t) \quad \text{for every } s \geqq t, \tag{1}$$

where $a(\cdot)$, $b(\cdot)$, $\sigma(\cdot)$ are given deterministic functions, $(w(s), s \geqq 0)$ is a standard Wiener process, x is the initial state at the time t and $(\nu(s), s \geqq 0)$ stands for the control which is a progressively measurable process with locally bounded variation.

The expected cost takes the form

$$J_{xt}(\nu) = E\left\{\int_t^T f(y(s), s)\, ds + c(t)\nu(0) + \int_t^T c(s)\, d|\nu|(s-t)\right\}, \tag{2}$$

with $f(\cdot,\cdot)$, $c(\cdot)$ given, $|\nu|$ denoting the variation of the process ν and T being the finite horizon. Hence, the optimal cost function is

$$u(x, t) = \inf\{J_{xt}(\nu): \nu\} \quad \text{for every } x, t. \tag{3}$$

The entire paper is devoted to the one-dimensional case, i.e. x belonging to $\mathbb{R}$; however, most of the results can be extended to multidimensional situations.

A formal application of the dynamic programming principle yields the complementary problem

$$\begin{aligned} &\max\{Au - f, |Du| - c\} = 0 \quad \text{in } \mathbb{R} \times [0, T[, \\ &u(\cdot, T) = 0 \quad \text{in } \mathbb{R}, \end{aligned} \tag{4}$$

* Received by the editors June 28, 1983, and in revised form April 25, 1984. This research was supported in part by the U.S. Army, Research Office under contract DAAG29-83-K-0014.

† Department of Mathematics, Wayne State University, Detroit, Michigan 48202.

‡ INRIA, Domaine de Voluceau, B.P. 105, Rocquencourt, 78153 Le Chesnay, Cedex, France.

for the optimal cost (3), where the operators

$$
\begin{aligned}
Au &= -\frac{\partial u}{\partial t} - \frac{1}{2}\sigma^2(t)\frac{\partial^2 u}{\partial x^2} - (a(t)x + b(t))\frac{\partial u}{\partial x}, \\
Du &= -\frac{\partial u}{\partial x},
\end{aligned} \tag{5}
$$

and $|\cdot|$ denotes the absolute value of a real number.

It is clear that (4) can be regarded either as a variational inequality or as a free boundary problem. In contrast with the classical aspect of the problem (4) we mention that among our assumptions it is allowed to have degeneracy, i.e., $\sigma(t) = 0$, and that we are interested in the characteristics of an optimal policy of the control as well as a possible computation of that optimal strategy. Moreover, we seek a suitable decomposition of (4) into problems, typically of the form

$$
\begin{aligned}
&\max\{Au - f, Du - c\} = 0 \quad \text{in } \mathbb{R} \times [0, T[, \\
&u(\cdot, T) = 0 \quad \text{in } \mathbb{R}.
\end{aligned} \tag{6}
$$

Also, we wish to be able to treat the case with constraints on the resources, i.e. in the minimization (3) we add a condition:

(7) the total variation of ν on $[0, T]$ is bounded by a constant K,

where K stands for the total resources available.

On the other hand, we will see that the problem (6), commonly referred to as the "monotone follower," can be obtained as a limit-case of a quasi-variational inequality.

As the main result of this paper, we should mention the characterization of the optimal cost function as the unique solution of the problem (4) or (6) in a certain sense; the proof of the existence of an optimal control; the construction of an optimal control of Markovian type; the reduction to problems of the form (6); and lastly, some properties of regularity for the optimal cost, e.g. locally Lipschitzian derivative of u, even without assuming uniform ellipticity of the operator A in (5).

This problem is motivated by our interest in studying the optimal control of a dissipative dynamical system under uncertainty. In the simplest model, one considers the automative cruise control of an aircraft under an uncertain wind condition. The equation (1) is the equation of motion, where $y(s)$ is the speed; $a(s) < 0$ the coefficient of air resistence; $b(s)$ the thrust force; the white-noise term the dynamic force due to the shifting wind condition, and the formal derivative $\dot{\nu}$ represents the control in the form of a corrective thrust force. We wish to find an optimal control policy ν over the flight time T so that, given a finite amount of fuel for correction, the flight speed will deviate as little as possible to a desirable cruising speed at a minimum fuel cost. This fact is expressed by the equations (2) and (3). The system (1)–(3) has another interesting interpretation in the context of optimal harvesting of randomly fluctuating resource Ludwig [36]. In this case, the equation (1) stands for a controlled linear growth model for the size y of a population, say, in a fishery, where $a > 0$ is the birth rate; the terms b and $(\sigma \dot{w})$ are, respectively, the mean and fluctuating rates of migration, and $\dot{\nu}$ denotes the harvesting rate. For instance, in a finite horizon, we would like to determine the harvesting rate in order to maintain the population size as close as possible to an equilibrium size at a minimum cost.

Let us remark that, when the rate function $a \equiv 0$, similar kinds of problems have been considered by several authors, in particular Bather and Chernoff [4], [5], Benes, Shepp and Witsenhausen [6], Borodowski et al. [12], Bratus [13], Chernousko [17],

[18], Gorbunov [22], Harrison and Taksar [24], Harrison and Taylor [25], Jacka [26], Shreve et al. [55], Karatzas [27], [28], and [41], [42]. The connection with optimal stopping is deeply investigated in Karatzas and Shreve [29], [30].

The methods to be used throughout this article are suggested by the techniques presented in the books of Bensoussan and Lions [8], [9], Fleming and Rishel [20], Friedman [21], Kinderlehrer and Stampacchia [31], and Krylov [32].

We organize the contents of the paper as follows:

1. Statement of the problem and assumptions. Let $(\Omega, \mathcal{T}, P)$ be a probability space, $(w(t), t \geqq 0)$ be a standard Wiener process in $\mathbb{R}$ and $(\mathcal{T}^t, t \geqq 0)$ be a filtration satisfying the usual conditions with respect to $(w(t), t \geqq 0)$, i.e., $(\mathcal{T}^t, t \geqq 0)$ is an increasing right continuous family of completed σ-subalgebras of $\mathcal{T}$ and $(w(t), t \geqq 0)$ is a martingale with respect to $(\mathcal{T}^t, t \geqq 0)$.

Denote by $\mathcal{V}$ the set of controls $\nu(\cdot)$ which are progressively measurable random processes from $[0, +\infty)$ into $\bar{\mathbb{R}}$ (extended real numbers), right continuous having left limits (cad-lag), nonnegative and increasing, i.e.,

$$\nu(0) \geqq 0, \qquad \nu(s) - \nu(t) \geqq 0 \quad \text{for every } s \geqq t \geqq 0. \tag{1.1}$$

The state of the dynamic system is described by the following stochastic equation

$$\begin{aligned} &dy(s) = d\nu(s-t) + (a(s)y(s) + b(s))\, ds + \sigma(s)\, dw(s-t), \qquad s \geqq t, \\ &y(t) = x + \nu(0), \end{aligned} \tag{1.2}$$

where $a(s)$, $b(s)$ and $\sigma^2(s)$ stand for the drift and the convariance terms, and x is the initial state at the time t. Note that $y(s) = y_{xt}(s)$ is a cad-lag random process adapted to $(\mathcal{T}^{s-t}, s \geqq t)$.

To each control ν in $\mathcal{V}$, we associate a cost given by the payoff functional

$$\begin{aligned} J_{xt}(\nu) = E\Bigg\{ &\int_t^T f(y(s), s) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda \right) ds + c(t)\nu(0) \\ &+ \int_t^T c(s) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda \right) d\nu(s-t) \Bigg\}, \end{aligned} \tag{1.3}$$

where f, α, c and T are respectively, the running cost, the discount factor, the instantaneous cost per unit of fuel and the finite horizon.

Our purpose is to characterize the optimal cost

$$\hat{u}(x, t) = \inf \{ J_{xt}(\nu) : \nu \text{ in } \mathcal{V} \} \tag{1.4}$$

and to construct an optimal control $\hat{v}$, i.e.

$$\hat{v} \quad \text{in } \mathcal{V} \text{ such that } \hat{u}(x, t) = J_{xt}(\hat{v}) \tag{1.5}$$

for each initial state (x, t). This problem corresponds to several simple models, e.g., control of a spaceship with unlimited fuel (cf. Bather and Chernoff [4]), optimal control with no turning back (cf. Barron and Jensen [1]), monotone follower problem (cf. in Benes et al. [6, problem # 2], Karatzas [27], [28]), optimal correction problem (cf. Chernousko [17], [18], Borodovskii et al. [12], Bratus [13], Gorbunov [22], optimal control of a dam (cf. Bather [3], Faddy [19]), control of Brownian motion (cf. Rath [50], [51], Chernoff and Petkau [16]) and inventory theory (cf. Bather [2], Menaldi and Rofman [45]).

A similar study will be made for the optimal cost

$$\hat{v}(x, z, t) = \inf\{J_{xt}(\nu): \nu \text{ in } \mathcal{V}, \nu(T) \leqq z\}, \tag{1.6}$$

where the positive constant z stands for the total amount of fuel available. This is associated with the previous cases under constraint of resources, e.g., the control of a spaceship with finite fuel available (cf. Bather and Chernoff [5], problem # 3 in Benes et al. [6]).

Let us summarize the technical assumptions as follows:

(1.7) T is a positive constant,

(1.8) $a(t)$, $b(t)$, $\sigma(t)$, $\alpha(t)$, $c(t)$ are Lipschitz functions from $[0, T]$ into $\mathbb{R}$ and either $c(t) \geqq c_0 > 0$ for every t or $c(t) = 0$ for every t,

(1.9) $f(x, t)$ is a nonnegative continuous function from $\mathbb{R} \times [0, T]$ into $\mathbb{R}$ such that there exist constants $m \geqq 1$, $0 \leqq c \leqq C$ satisfying

$$\begin{aligned} &c|x^+|^m - C \leqq f(x, t) \leqq C(1+|x|^m), \\ &|f(x, t) - f(x', t)| \leqq C(1+|x|^{m-1}+|x'|^{m-1})|x-x'|, \\ &|f(x, t) - f(x, t')| \leqq C(1+|x|^m)|t-t'|, \\ &0 \leqq \frac{\partial^2 f}{\partial x^2}(x, t) \leqq C(1+|x|^q), \qquad q = (m-2)^+, \end{aligned}$$

for every x, x', t, t',

where $(\cdot)^+$ denotes the postive part of a real number, i.e., $x^+ = x$ if $x \geqq 0$ and $x^+ = 0$ if $x \leqq 0$. Note that $\sigma(t)$ could vanish, even everywhere, and then the problem could be degenerate and even deterministic. On the other hand, since the horizon T is finite, without loss of generality, the function $\alpha(t)$ may be assumed to satisfy for every t

$$\alpha(t) \geqq \alpha_0, \qquad \alpha_0 \text{ is positive large enough.} \tag{1.10}$$

Let us introduce two penalized problems associated with (1.4) as follows: $\varepsilon > 0$,

(1.11) $\mathcal{V}_\varepsilon$ is the set of all controls $\nu(\cdot)$ in $\mathcal{V}$ such that $\nu(t)$ is Lipschitz continuous and

$$0 \leqq \frac{d\nu}{dt}(t) \leqq \frac{1}{\varepsilon} \quad \text{for almost every } t \geqq 0,$$

$$\hat{u}^\varepsilon(x, t) = \inf\{J_{xt}(\nu): \nu \text{ in } \mathcal{V}_\varepsilon\} \tag{1.12}$$

and

(1.13) $\mathcal{V}_*$ is the set of impulse controls, i.e., $\nu(\cdot)$ in $\mathcal{V}$ such that there exist stopping times $(\theta_j, j=1,2,\cdots)$ $0 \leqq \theta_j \leqq \theta_{j+1}$, for every $j=1,2,\cdots$, and adapted random variables $(\xi_j, j=1,2,\cdots)$ satisfying

$$\nu(t) = \sum_{j=1}^{\infty} \xi_j I(\theta_j \leqq t) \quad \text{for every } t \geqq 0,$$

where $I(\theta_j \leqq t)$ is the characteristic function of the set $(\theta_j \leqq t)$,

$$J^{\varepsilon}_{xt}(\nu) = J_{xt}(\nu) + \varepsilon E\left\{\sum_{j=1}^{\infty} \exp\left(-\int_t^{t+\theta_j} \alpha(s)\, ds\right)\right\}, \tag{1.14}$$

$$\hat{u}_{\varepsilon}(x,t) = \inf\{J^{\varepsilon}_{xt}(\nu):\ \nu \text{ in } \mathcal{V}_*\}. \tag{1.15}$$

Notice that (1.12) and (1.15) correspond respectively to a classical stochastic control problem (cf. Fleming and Rishel [20]) and an impulse control problem (cf. Bensoussan and Lions [9]). The term "penalized" is used to indicate that the formal Dynamic Programming equations associated with the problems (1.12) and (1.15) are indeed two possible penalizations of the equation (2.4) below.

On the other hand, for $z \geqq 0$ and ν in $\mathcal{V}$, define a cost

$$\begin{aligned} F(x,z,t,v) = E\Bigg\{ & \int_t^{\tau} f(y(s),s) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) ds + c(t)\nu(0) I(t<\tau) \\ & + \int_t^{\tau} c(s) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) d\nu(s-t) \\ & + u^0(y(\tau),\tau) \exp\left(-\int_t^{\tau} \alpha(s)\, ds\right)\Bigg\}, \end{aligned} \tag{1.16}$$

where $\tau = \tau_{zt}$ is the first exit time from $[\nu \leqq z]$ of the process $\nu(s)$, i.e.

$$\tau = \inf\{s \in [t,T]:\ \nu(s-t) > z\}, \tag{1.17}$$

$y^0(s)$ is given by (1.2) with $\nu = 0$, and

$$u^0(x,t) = E\left\{\int_t^T f(y^0(s),s) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) ds\right\} \tag{1.18}$$

represents the cost of free evolution. It is clear that the optimal cost (1.6) corresponding to finite fuel conditions, satisfies

$$\hat{v}(x,z,t) = \inf\{F(x,z,t;\nu):\ \nu \text{ in } \mathcal{V}\}. \tag{1.19}$$

The relation will be used to reduce the problem with constrained resources to the case without constraint.

To conclude this section, let us observe that it is possible to obtain the same optimal cost (1.4) by minimizing the functional cost (1.3), denoted now by $J_{xt}(\mathcal{A})$, over all system controls $\mathcal{A}$, where $\mathcal{A}$ is a set including the probability space $(\Omega, \mathcal{T}, P)$, the filtration, the Wiener process and the control $(\mathcal{T}^t, w(t), \nu(t), t \geqq 0)$. The same idea corresponds to identifying the state process $y_{xt}(s)$ with its probability law P_{xt} on the sample space D of the cad-lag functions. The probability law P_{xt} is characterized by

the conditions

$$P_{xt}(X_t = x) = 1,$$

$$\varphi(X_s, s) - \int_t^s \left[\frac{\partial \varphi}{\partial s} + \sigma^2(\lambda)\frac{\partial^2 \varphi}{\partial x^2} + (a(\lambda)X_\lambda + b(\lambda))\frac{\partial \varphi}{\partial x}\right](X_\lambda, \lambda)\, d\lambda \tag{1.20}$$

$$- \int_t^s \varphi(X_\lambda, \lambda)\, d\nu(\lambda) = M_s$$

is a martingale in $t \leqq s \leqq T$, for every smooth function φ in $\mathbb{R} \times [0, T]$.

More details about this formulation for stochastic control problems can be found in Nisio [49], Bensoussan and Lions [8], Lions and Menaldi [35].

2. The dynamic programming approach. Consider the differential operators

$$Au = -\frac{\partial u}{\partial t} - \frac{1}{2}\sigma^2(t)\frac{\partial^2 u}{\partial x^2} - (a(t)x + b(t))\frac{\partial u}{\partial x} + \alpha(t)u, \tag{2.1}$$

and

$$Bu = -\frac{\partial u}{\partial x} - c(t). \tag{2.2}$$

A heuristic application of the dynamic programming to the penalized problem (1.11), (1.12) yields the following Hamilton-Jacobi-Bellman equation

$$\begin{aligned} &Au + \frac{1}{\varepsilon}(Bu)^+ = f \quad \text{in } \mathbb{R} \times [0, T[, \\ &u(\cdot, T) = 0 \quad \text{in } \mathbb{R} \end{aligned} \tag{2.3}$$

to be satisfied by the optimal cost $\hat{u}^\varepsilon$ defined in (1.12). Then, as ε tends to zero, (2.3) becomes

$$\begin{aligned} &(Au - f) \vee Bu = 0 \quad \text{in } \mathbb{R} \times [0, T[, \\ &u(\cdot, T) = 0 \quad \text{in } \mathbb{R}, \end{aligned} \tag{2.4}$$

where $x \vee y$ denotes the maximum of the two real numbers x and y. Equation (2.4) will be used to characterize the optimal cost $\hat{u}$ given by (1.4).

On the other hand, the quasi-variational inequality associated with the impulse control problem (1.14), (1.15) is

$$\begin{aligned} &(Au - f) \vee (u - Mu - \varepsilon) = 0 \quad \text{in } \mathbb{R} \times [0, T[, \\ &u(\cdot, T) = 0 \quad \text{in } \mathbb{R}, \end{aligned} \tag{2.5}$$

where

$$Mu(x, t) = \inf\{\xi c(t) + u(x + \xi, t) : \xi \geqq 0\}, \tag{2.6}$$

which is satisfied by the optimal cost $\hat{u}_\varepsilon$ defined in (1.15). Moreover, $\hat{u}_\varepsilon$ is indeed the maximum solution of (2.5). Thus, as ε tends to zero, (2.5) becomes

$$\begin{aligned} &(Au - f) \vee (u - Mu) = 0 \quad \text{in } \mathbb{R} \times [0, T[, \\ &u(\cdot, T) = 0 \quad \text{in } \mathbb{R}. \end{aligned} \tag{2.7}$$

Hence, the optimal cost $\hat{u}$ given by (1.4) will be the maximum solution of the equation

(2.7). For details on the two penalized problems one is referred to the books of Fleming and Rishel [20], Bensoussan and Lions [9] and to the works [38], [41] and [52].

2.1. Some estimates. First of all, we will deduce some a priori estimates for the optimal costs (1.4), (1.12) and (1.15).

THEOREM 2.1. *Under the assumptions* (1.7), $\cdots$, (1.10) *the optimal cost* $\hat{u}$ *defined by* (1.4) *is a nonnegative continuous function such that for some constants* $0 < c \leqq C$, *the same* $m \geqq 1$ *of hypothesis* (1.9), *and every* (x, t), (x', t') *in* $\mathbb{R} \times [0, T]$ *we have*

$$
\begin{aligned}
& c|x^+|^m - C \leqq \hat{u}(x, t) \leqq C(1+|x|^m), \\
& |\hat{u}(x, t) - \hat{u}(x', t)| \leqq C(1+|x|^{m-1}+|x'|^{m-1})|x-x'|, \\
& 0 \leqq \frac{\partial^2 \hat{u}}{\partial x^2}(x, t) \leqq C(1+|x|^q), \qquad q = (m-2)^+,
\end{aligned} \tag{2.8}
$$

so, $\hat{u}$ *is convex in the first variable. Moreover, if* $\hat{u}$ *satisfies*

$$
\begin{aligned}
\hat{u}(x, t) \leqq E\Big\{ \int_t^{t'} f(y^0(s), s) \exp\Big(-\int_t^s \alpha(\lambda) d\lambda \Big) ds \\
+ \hat{u}(y^0(t'), t') \exp\Big(-\int_t^{t'} \alpha(s)\, ds \Big) \Big\},
\end{aligned} \tag{2.9}
$$

for every $t' \geqq t \geqq 0$, x *in* $\mathbb{R}$ *and* $y^0(s)$ *given by* (1.2) *with* $\nu = 0$, *i.e. the dynamic programming in the weak sense, then we have*

$$
|\hat{u}(x, t) - \hat{u}(x, t')| \leqq C(1+|x|^m)|t-t'|, \tag{2.10}
$$

for every x *in* $\mathbb{R}$, t, t' *in* $[0, T]$ *and some constant* C.

Proof. Since f has m-polynomial growth as x tends to positive infinity, $\nu(s) \geqq 0$ and for the vanishing control $\nu = 0$

$$
J_{xt}(0) \leqq C(1+|x|^m),
$$

one can restrain the set of admissible controls to those satisfying

$$
\int_t^T |y_{xt}(s)|^m \, ds \leqq C(1+|x|^m) \tag{2.11}
$$

for the same $m \geqq 1$ of (1.9) and a suitable constant C independent of x, t and ν. Similarly, every admissible control ν may satisfy without loss of generality, the inequality

$$
\int_t^T |(y_{xt}(s))^+|^m \, ds \geqq c|x^+|^m - C \tag{2.12}
$$

for some constants $0 < c \leqq C$ independent of x, t and ν. It is clear that from (2.11) and (2.12) one deduces the first condition of (2.8).

Now, using the fact that for some constant C and for every t, x, x' and ν the following estimate holds

$$
\int_t^T |y_{xt}(s) - y_{x't}(s)|^m \, ds \leqq C|x-x'|^m, \tag{2.13}
$$

and starting with

$$|\hat{u}(x,t)-\hat{u}(x',t)| \leqq \sup\{|J_{xt}(\nu)-J_{x't}(\nu)|:\ \nu \text{ in } \mathcal{V} \text{ satisfying (2.9)}\},$$

$$|J_{xt}(\nu)-J_{x't}(\nu)| \leqq CE\left\{\int_t^T (1+|y_{xt}(s)|^{m-1}+|y_{x't}(s)|^{m-1})|y_{xt}(s)-y_{x't}(s)|\,ds\right\},$$

where C is a constant independent of x, t and ν, we obtain the second estimate of (2.8) after applying Hölder's inequality.

In order to get the estimate (2.10), we observe that

$$J_{xt}(\nu) = E\left\{\int_0^{T-t} f(y(t+s),t+s)\exp\left(-\int_0^s \alpha(t+\lambda)\,d\lambda\right)ds + c(t)\nu(0) + \int_0^{T-t} c(t+s)\exp\left(-\int_0^s \alpha(t+\lambda)\,d\lambda\right)d\nu(s)\right\} \tag{2.14}$$

and if $c(s)$ is strictly positive, the set of admissible controls can be restricted to those continuous at $T-t$ and satisfying for every x, t

$$E\{\nu(T-t)\} \leqq C(1+|x|^m), \tag{2.15}$$

for a suitable constant C independent of x, t and ν. If $y(s)$ and $y'(s)$ denote the evolutions associated respectively to x, t, ν and x, t', ν, we have

$$E\{|y(t+s)-y'(t'+s)|^m\} \leqq C|t-t'|^m, \quad \text{for every } s \text{ in } [0,T-t] \tag{2.16}$$

and some constant C independent of x, t, t' and ν. Hence, starting with

$$\hat{u}(x,t)-\hat{u}(x,t') \leqq \sup\{J_{xt}(\nu)-J_{xt'}(\nu):\ \nu \text{ in } \mathcal{V} \text{ satisfying (2.11) and (2.15)}\}, \qquad t' \leqq t$$

and in view of (2.14), (1.9), (1.8), for some constant C,

$$J_{xt}(\nu)-J_{xt'}(\nu) \leqq CE\left\{\left[\int_t^T (1+|y(s)|^m)\,ds+\nu(T-t)\right]|t-t'| + \int_0^{T-t}(1+|y(t+s)|^{m-1}+|y'(t'+s)|^{m-1})|y(t+s)-y'(t'+s)|\,ds\right\},$$

we deduce, for a constant C independent of x, t and t',

$$\hat{u}(x,t')-\hat{u}(x,t) \leqq C(1+|x|^m)|t-t'|, \qquad t' \leqq t, \tag{2.17}$$

after using Hölder's inequality and (2.16). To obtain a similar inequality for $t' > t \geqq 0$ we shall use (2.9) as follows. From Itô's formula applied to a sequence of smooth functions convergent to

$$x \to \hat{u}(x,t')$$

we get, for some constant $C > 0$,

$$E\{\hat{u}(y^0(t'),t')\} \leqq CE\left\{\int_t^{t'}(1+|y^0(s)|^m)\,ds\right\}+\hat{u}(x,t'),$$

in view of (2.8). Since there is a constant C_0 such that

$$E\{|y^0(s)|^m\} \leqq C_0(1+|x|^m), \qquad s \text{ in } [t,t'],$$

the dynamic programming property (2.9) yields

$$\hat{u}(x,t)-\hat{u}(x,t') \leqq C(1+|x|^m)|t-t'|, \qquad t' > t.$$

It is clear that this last estimate and (2.18) imply (2.10).

To estimate the second derivative of $\hat{u}(x, t)$ in x, let us proceed as in Krylov [27]. From

$$\hat{u}(x+\Delta x, t)-2\hat{u}(x, t)+\hat{u}(x-\Delta x, t)$$
$$\leqq \sup\{(J_{x+\Delta x}(\nu)-2J_x(\nu)+J_{x-\Delta x}(\nu)): \nu \text{ in } \mathcal{V} \text{ satisfying (2.9)}\},$$

where the subscript t has been omitted in the functional $J_{xt}(\nu)$, and the equalities

$$f(z+r\Delta x, s)-2f(z, s)+f(z-r\,\Delta x, s)=|\Delta x|^2\int_0^1 d\lambda\int_{-\lambda}^{\lambda}\frac{\partial^2 f}{\partial x^2}(z+\mu r\Delta x, s)|r|^2\,d\mu,$$

$$y_{x+\Delta x}(s)=y_x(s)+\Delta x\exp\left(\int_t^s a(\lambda)\,d\lambda\right),$$

for every x, z, Δx, r and s, we deduce

$$\hat{u}(x+\Delta x, t)-2\hat{u}(x, t)+\hat{u}(x-\Delta x, t)\leqq C(1+|x|^q)|\Delta x|^2,$$

where $q=(m-2)^+$, C is a suitable constant independent of (x, t) in $\mathbb{R}\times[0, T]$ and Δx in $[-1, 1]$. Hence an upper bound for the second derivative in x of $\hat{u}(x, t)$ is obtained.

To complete this proof, we need to show that the optimal cost $\hat{u}(x, t)$ is a convex function in the first variable x. Since the functional $J_{xt}(\nu)$ is simultaneously convex in (x, ν) and the set of controls $\mathcal{V}$ is a convex set, we have

$$\hat{u}(\theta x+(1-\theta)x', t)\leqq \theta J_{xt}(\nu)+(1-\theta)J_{x't}(\nu') \tag{2.18}$$

for every t, x, x', ν, ν' and $0\leqq\theta\leqq 1$. Thus, the inequality (2.18) implies the convexity of function $\hat{u}$. □

COROLLARY 2.1. *Under the same assumptions of Theorem* 2.1 *the optimal cost* $\hat{u}^\varepsilon(x, t)$ *corresponding to the penalized problem* (1.11), (1.12), *is a nonnegative continuous function satisfying conditions* (2.8) *and* (2.10) *uniformly in* $\varepsilon>0$. *Furthermore, the optimal cost* $\hat{u}_\varepsilon(x, t)$, *corresponding to the penalized problem* (1.14), (1.15), *is a nonnegative continuous function satisfying conditions* (2.8), *and* (2.10) *except the bound of the second derivative, uniformly in* $\varepsilon>0$.[1] □

Remark 2.1. The optimal cost $\hat{u}_\varepsilon(x, t)$ given by (1.15) is not convex in general. However a discretization in the time variable t allows us to adopt a technique of Scarf [49] in order to show that $\hat{u}_\varepsilon(x, t)$ is ε-convex in x, i.e. for every (x, t) in $\mathbb{R}\times[0, T]$

$$\varepsilon+\hat{u}_\varepsilon(x+z, t)-\hat{u}_\varepsilon(x, t)\geqq z\frac{\partial\hat{u}_\varepsilon}{\partial x}(x, t)\quad\text{for every } z\geqq 0 \tag{2.19}$$

and any $\varepsilon>0$. On the other hand, a lower bound for the second derivative in x of $\hat{u}_\varepsilon(x, t)$ may be deduced by using the dynamic programming equation (2.3) and the nondegeneracy of $\sigma(t)$.

Define the subset of admissible controls

(2.20) $\mathcal{V}_0$ is the set of all controls $\nu(\cdot)$ in $\mathcal{V}$ such that $\nu(t)$ is uniformly Lipschitz continuous on $[0, +\infty]$, i.e. $0\leqq d\nu(t)/dt\leqq C$, for almost every t and some constant C.

THEOREM 2.2. *Let the assumptions* (1.7), $\cdots$, (1.10) *hold. The infimum of the functional* $J_{xt}(\nu)$, *given by* (1.3), *over the sets* (a) *all controls* ν *in* $\mathcal{V}$, (b) *all Lipschitz controls* ν *in* $\mathcal{V}_0$, (c) *all impulse controls* ν *in* $\mathcal{V}_*$, *is always the same. Moreover, the*

[1] Note that $\hat{u}^\varepsilon$ and $\hat{u}_\varepsilon$ satisfy the condition (2.9). See Theorems 2.3 and 2.4.

functions $\hat{u}^\varepsilon$ *and* $\hat{u}_\varepsilon$, *given by* (1.12) *and* (1.15), *converge to the optimal cost* $\hat{u}$ *pointwise in* $\mathbb{R}\times[0, T]$.[2]

Proof. Denote by $y(s)$, $y'(s)$ the output corresponding to controls ν, ν' in $\mathcal{V}$ given by (1.2). Using Gronwall's inequality, we obtain

$$\int_t^T |y(s)-y'(s)|^m\,ds \leqq |\nu(0)-\nu'(0)|^m + C\int_0^{T-t} |\nu(s)-\nu'(s)|^m\,ds \tag{2.21}$$

for a constant C independent of ν, ν', x, and t.

Suppose an arbitrary control ν in $\mathcal{V}$ is given. We define

$$\nu_n(t) = \begin{cases} (1-nt)\nu(0) + n^2 t\displaystyle\int_0^{1/n} \nu(s)\,ds & \text{if } 0 \leqq t \leqq 1/n, \\ n\displaystyle\int_{t-1/n}^{t} \nu(s)\,ds & \text{otherwise} \end{cases} \tag{2.22}$$

and

$$\nu_-(t) = \begin{cases} \nu(0) & \text{if } t=0, \\ \lim\limits_{s\uparrow t} \nu(s) & \text{otherwise.} \end{cases} \tag{2.23}$$

Since $\nu(\cdot)$ is a cad-lag process, $\nu_n(s)$ converges, for any fixed ω, to $\nu_-(s)$ for every s, as n approaches infinity. Moreover, except for a countable set in s, we have $\nu_-(s)=\nu(s)$. This fact and the estimate (2.21) imply

$$J_{xt}(\nu_n) \to J_{xt}(\nu) \quad \text{as } n\to\infty. \tag{2.24}$$

Hence

$$\hat{u}(x, t) = \inf\{J_{xt}(\nu)\colon \nu \text{ in } \mathcal{V}_0\}, \tag{2.25}$$

because ν_n given by (2.22) belongs to $\mathcal{V}_0$.

Now, suppose ν is an arbitrary Lipschitz control in $\mathcal{V}_0$ and define

$$\nu_n(s) = \nu\left(\frac{i}{n}\right) \quad \text{if } \frac{i}{n} \leqq s < \frac{i+1}{n}, \quad i=0, 1, \cdots, \tag{2.26}$$

which is an impulse control in $\mathcal{V}_*$. Thus from (2.21) and (2.24) we deduce

$$\hat{u}(x, t) = \inf\{J_{xt}(\nu)\colon \nu \text{ in } \mathcal{V}_*\}. \tag{2.27}$$

To complete the proof, in view of Theorem 2.1, we only need to show that the optimal costs $\hat{u}^\varepsilon$ and $\hat{u}_\varepsilon$, given respectively by (1.12) and (1.15) satisfy for every (x, t) in $\mathbb{R}\times[0, T]$

$$\hat{u}^\varepsilon(x, t) \to \hat{u}(x, t) \quad \text{as } \varepsilon \downarrow 0, \tag{2.28}$$

$$\hat{u}_\varepsilon(x, t) \to \hat{u}(x, t) \quad \text{as } \varepsilon \downarrow 0, \tag{2.29}$$

where $\hat{u}$ is the optimal cost (1.4). The first convergence (2.28) is deduced from equalities (2.25) and

$$\mathcal{V}_0 = \bigcup\{\mathcal{V}_\varepsilon\colon \varepsilon > 0\}. \tag{2.30}$$

To prove the convergence (2.29), we use (2.27) and the fact that for every ν in $\mathcal{V}_*$

[2] The convergence is also uniform over every compact subset of $\mathbb{R}\times[0, T]$.

such that $J^{\varepsilon}_{xt}(\nu)$ is finite,

$$J^{\varepsilon}_{xt}(\nu) \to J_{xt}(\nu) \quad \text{as } \varepsilon \downarrow 0, \tag{2.31}$$

where the limit is decreasing. □

Remark 2.2. The estimates of Theorem 2.1 allow us to obtain a locally uniform convergence of the first derivative in x of the optimal cost $\hat{u}^{\varepsilon}(x, t)$ defined by (1.12). Moreover, some weak convergence of the first and second derivatives of $\hat{u}^{\varepsilon}$ and $\hat{u}_{\varepsilon}$ holds.

Remark 2.3. A similar result to Theorem 2.2 can be found in Menaldi, Quadrat and Rofman [40], Menaldi and Rofman [46].

2.2. Characterization of the optimal cost. Denote by V_m the function space,

v belongs to V_m if $v: \mathbb{R} \times [0, T] \to \mathbb{R}$ is such that

$$\begin{aligned} &|v(x, t)| + \left|\frac{\partial v}{\partial t}(x, t)\right| \leqq C(1+|x|^m), \\ &\left|\frac{\partial v}{\partial x}(x, t)\right| \leqq C(1+|x|^{m-1}), \end{aligned} \tag{2.32}$$

for almost every (x, t) and some constant C,

and by L^{∞}_{loc} the space of measurable real functions which are locally essentially bounded in $\mathbb{R} \times]0, T[$.

Consider the following partial differential equation:

(2.33) Find $\hat{u}^{\varepsilon}$ in V_m such that $\partial^2 \hat{u}^{\varepsilon}/\partial x^2$ belongs to L^{∞}_{loc}, $\hat{u}^{\varepsilon}(x, T) = 0$ for every x in $\mathbb{R}$ and $A\hat{u}^{\varepsilon} + (1/\varepsilon)(B\hat{u}^{\varepsilon})^{+} = f$, a.e. in $\mathbb{R} \times]0, T[$,

where operators A and B are defined by (2.1) and (2.2).

THEOREM 2.3. *Assume the hypotheses* (1.7), $\cdots$, (1.10) *hold. Then* (2.33) *has one and only one solution, which is given explicitly as the optimal cost* (1.12). *Moreover, the inequality* (2.9) *is valid and if*

$$x^*_{\varepsilon}(t) = \inf\left\{x: \frac{\partial \hat{u}^{\varepsilon}}{\partial x}(x, t) + c(t) > 0\right\}, \tag{2.34}$$

we have for every (x, t) *in* $\mathbb{R} \times [0, T]$

$$\begin{aligned} &A\hat{u}^{\varepsilon} = f \quad \text{and} \quad B\hat{u}^{\varepsilon} \leqq 0 && \text{if } x \geqq x^*_{\varepsilon}(t), \\ &A\hat{u}^{\varepsilon} + \frac{1}{\varepsilon} B\hat{u}^{\varepsilon} = f \quad \text{and} \quad B\hat{u}^{\varepsilon} \geqq 0 && \text{if } x \leqq x^*_{\varepsilon}(t). \end{aligned} \tag{2.35}$$

Proof. First we suppose that $\sigma(t)$ is nondegenerate, i.e.

$$\sigma^2(t) \geqq \mu > 0 \quad \text{for every } t \text{ in } [0, T]. \tag{2.36}$$

Then standard techniques in partial differential equations prove that problem (2.33) has a smooth solution. Moreover, classical arguments of stochastic control (e.g. Bensoussan and Lions [8], Fleming and Rishel [20], Krylov [32]) permit us to identify the unique solution of (2.33) with the optimal cost (1.12). Also, the dynamic programming principle holds. In particular (2.9) is true.

To study the degenerate case, i.e., dropping (2.36), we regularize the differential operator (2.1),

$$A_{\eta} = A - \frac{1}{2}\eta \frac{\partial^2}{\partial x^2}, \qquad \eta > 0. \tag{2.37}$$

Because the estimates (2.8) and (2.10) of Theorem 2.1 hold uniformly in η as η tends to zero, we can pass to the limit in η and obtain a solution of the problem (2.33). The uniqueness follows for instance, from the weak maximum principle for degenerate elliptic equations (e.g., Bony [11]). To show that the solution of (2.33) is the optimal cost (1.12), we observe that for every (x, t) in $\mathbb{R}\times[0, T]$, every control ν in $\mathcal{V}$ and some constant $C>0$,

$$E\{|y^{\eta}(s)-y(s)|^{m}\}\leqq C\eta^{m/2}, \qquad \eta \text{ positive}, \tag{2.38}$$

where $y^{\eta}(s)$ and $y(s)$ are the evolutions associated with $(\sigma^2+\eta)^{1/2}$ and σ respectively in the state equation (1.2).

As a consequence of convexity, we have

$$B\hat{u}^{\varepsilon}(x, t)\leqq B\hat{u}^{\varepsilon}(x', t) \quad \text{if } x\leqq x'$$

for every fixed t in $[0, T]$. This implies the last conditions (2.35). □

Remark 2.4. Using a convolution kernel it is possible to establish that for every function $u(x, t)$,

$$u \text{ in } V_m,\ Au=h \text{ in } \mathcal{D}'(\mathbb{R}\times[0, T[) \text{ with } h(x, t) \text{ continuous in } x \text{ and measurable in } t, \tag{2.39}$$

where $\mathcal{D}'(\mathbb{R}\times]0, T[)$ denotes the space of distributions on $\mathbb{R}\times]0, T[$, we can apply Itô's formula for every Lipschitz continuous control ν in $\mathcal{V}_0$, i.e.

$$\begin{aligned} &u(x, t)-E\left\{u(y(T), T)\exp\left(-\int_t^T \alpha(s)\,ds\right)\right\} \\ &\quad=E\left\{\int_t^T\left[h-\dot{\nu}(s)\frac{\partial u}{\partial x}\right](y(s), s)\exp\left(-\int_t^s \alpha(\lambda)\,d\lambda\right)ds\right\}, \end{aligned} \tag{2.40}$$

where $\dot{\nu}$ is the derivative of the Lipschitz control ν. □

Now, consider the problem:

Find u_{ε} in V_m such that

$$\begin{aligned} &u_{\varepsilon}(x, T)=0 \quad \text{for every } x \text{ in } \mathbb{R}, \\ &Au_{\varepsilon}\leqq f \quad \text{in } \mathcal{D}'(\mathbb{R}\times]0, T[), \\ &u_{\varepsilon}\leqq \varepsilon+Mu_{\varepsilon} \quad \text{in } \mathbb{R}\times[0, T], \end{aligned} \tag{2.41}$$

where M denotes the operator (2.6).

THEOREM 2.4. *Suppose the assumptions* (1.7), ⋯, (1.10) *hold. Then the quasi-variational inequality* (2.41) *has a maximum solution* $\hat{u}_{\varepsilon}$, *which is given explicitly as the optimal cost* (1.15). *Moreover, the inequality* (2.9) *is valid.*

Proof. First, for fixed ψ in V_m and $\eta>0$, consider the problem

Find u in V_m such that

$$\begin{aligned} &u(x, T)=0 \quad \text{for every } x \text{ in } \mathbb{R}, \\ &Au+\frac{1}{\eta}(u-\psi)^{+}=f \quad \text{in } \mathcal{D}'(\mathbb{R}\times]0, T[). \end{aligned} \tag{2.42}$$

It is clear that as in Theorem 2.3, we can show that the equation (2.42) has a unique

solution $u = u(x, t; \psi, \eta)$ which satisfies

$$u(\psi, \eta) = \inf\left\{ G(\delta) : \delta \text{ is adapted}, 0 \leqq \delta \leqq \frac{1}{\eta} \right\}, \tag{2.43}$$

where

$$G(\delta) = E\left\{ \int_t^T [f(y^0(s), s) + \delta(s)\psi(y^0(s), s)] \exp\left[-\int_t^s (\alpha(\lambda) + \delta(\lambda))\, d\lambda \right] ds \right\},$$

and $y^0(s)$ is given by (1.2) with $\nu = 0$.

Similar to Theorem 2.1 we can prove that $u(\psi, \eta)$ belongs to V_m uniformly as η tends to zero. Therefore,

$$u(\psi, \eta) \to u(\psi) \quad \text{as } \eta \to 0, \tag{2.44}$$

in a decreasing fashion and with a local uniformity in $\mathbb{R} \times [0, T]$. Moreover, the limit function $u = u(\psi)$ is the unique solution of the variational inequality:

$$\begin{aligned} &\text{Find } u \text{ in } V_m \text{ such that} \\ &\quad u(x, T) = 0 \quad \text{for every } x \text{ in } \mathbb{R}, \\ &\quad Au \leqq f \quad \text{in } \mathscr{D}'(\mathbb{R} \times]0, T[), \\ &\quad u \leqq \psi \quad \text{in } \mathbb{R} \times [0, T[, \\ &\quad Au = f \quad \text{in } \mathscr{D}'([u < \psi]), \end{aligned} \tag{2.45}$$

where $[u < \psi]$ denotes the set of points satisfying $u(x, t) < \psi(x, t)$, and also

$$u(\psi) = \inf\{F(\theta) : \theta \text{ is stopping time}, t \leqq \theta \leqq T\}, \tag{2.46}$$

with

$$\begin{aligned} F(\theta) = E\left\{ \int_t^\theta f(y^0(s), s) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda \right) ds \right. \\ \left. + \psi(y^0(\theta), \theta) \exp\left(-\int_t^\theta \alpha(s)\, ds \right) I(\theta < T) \right\}, \end{aligned}$$

and $I(\theta < T)$ is the characteristic function of the set $[\theta < T]$. We remark that (2.42) is referred to as the penalized problem associated to the variational inequality (2.45). Also the control problem (2.46) is called an optimal stopping time problem (e.g. Bensoussan and Lions [8], Friedman [21], Kinderlehrer and Stampacchia [31]). Notice that the running cost f is unbounded and the operator A could be degenerate (cf. [37], [41], [42] and [52]).

Now, observe that

$$\psi \leqq \varphi \quad \text{implies} \quad u(\varphi) \leqq u(\varphi). \tag{2.47}$$

We may define the decreasing sequence of function

$$u^n = u(\psi), \qquad \psi = \varepsilon + Mu^{n-1}, \qquad n = 1, 2, \cdots, \tag{2.48}$$

where u^0 is the unique solution in V_m of the equation

$$\begin{aligned} &Au^0 = f \quad \text{in } \mathscr{D}'(\mathbb{R} \times]0, T[), \\ &u^0(x, T) = 0 \quad \text{for every } x \text{ in } \mathbb{R}. \end{aligned} \tag{2.49}$$

Standard techniques (e.g. Bensoussan [7], Bensoussan and Lions [9], and [38], or [52]) show that $u^n = \hat{u}_\varepsilon^n$,

$$\hat{u}_\varepsilon^n = \inf\{J_{xt}^\varepsilon(\nu): \nu \text{ in } \mathcal{V}_*^n\}, \tag{2.50}$$

where $\mathcal{V}_*^n$ denotes the subset of impulse control $\mathcal{V}_*$ given by (1.13) such that $\theta_j = +\infty$ for every $j \geqq n$.

Thus, as in Theorem 2.1, we can prove that functions (2.50) remain in V_m uniformly as n approaches infinity. Hence the limit function

$$u_\varepsilon^* = \lim_n \hat{u}_\varepsilon^n \tag{2.51}$$

solves the quasi-variational inequality (2.41). It is clear that

$$u_\varepsilon^* \geqq \hat{u}_\varepsilon, \tag{2.52}$$

with $\hat{u}_\varepsilon$ denoting the optimal cost (1.15).

A crucial point is to deduce that

$$\hat{u}_\varepsilon \geqq u_\varepsilon \quad \text{for every solution } u_\varepsilon \text{ of (2.41).} \tag{2.53}$$

Indeed, let ν be any impulse control, i.e.

$$\nu(s) = \sum_{j=1}^{\infty} \xi_j I(\theta_j \leqq s),$$

which may satisfy

$$\int_0^{T-t} |\nu(s)|^m \, ds \leqq C(1+|x|^m), \tag{2.54}$$

for a suitable constant independent of (x, t), and

$$\theta_j = T \quad \text{for every } j \geqq N(\omega) \quad \text{some random index,} \tag{2.55}$$

without loss of generality. Since u_ε solves (2.41), we obtain

$$u_\varepsilon(x, t) \leqq J_{xt}^\varepsilon(\nu_n) + E\left\{ u_\varepsilon(y(\theta_n), \theta_n) \exp\left(-\int_t^{t+\theta_n} \alpha(s)\, ds \right) \right\}, \tag{2.56}$$

for

$$\nu_n(s) = \sum_{j=1}^{n} \xi_j I(\theta_j \leqq s).$$

As n tends to infinity in (2.56) and by virtue of (2.54), (2.55), we get

$$u_\varepsilon(x, t) \leqq J_{xt}^\varepsilon(\nu)$$

which implies (2.53). From this, the equality must hold in (2.52) and the optimal cost (1.15) is the maximum solution of (2.41). □

Remark 2.5. Under the same assumptions of Theorem 2.4, we can prove that the optimal cost (1.15) is the unique solution of problem (2.41) together with the condition

$$A\hat{u}_\varepsilon = f \quad \text{in } \mathcal{D}'([\hat{u}_\varepsilon < \varepsilon + M\hat{u}_\varepsilon]), \tag{2.57}$$

where $[\hat{u}_\varepsilon < \varepsilon + M\hat{u}_\varepsilon]$ is the set of all points satisfying $\hat{u}_\varepsilon(x, t) < \varepsilon + M\hat{u}_\varepsilon(x, t)$. For a complete treatment of impulse control problems of nondegenerate diffusion processes with bounded running cost, we refer to the book of Bensoussan and Lions [9]. Similar problems are studied in [37], [38] and [52], [53], and some discrete approximations

are described in Bensoussan and Robin [10], Capuzzo-Dolcetta and Matzeu [14] and in general in Kushner [33].

Going back to the initial problem (1.4), consider the set of conditions:

$$\begin{aligned} &\text{Find } u \text{ in } V_m \text{ such that} \\ &u(x, T) = 0 \quad \text{for every } x \text{ in } \mathbb{R}, \\ &Au \leqq f \quad \text{in } \mathscr{D}'(\mathbb{R} \times]0, T[), \\ &u \leqq Mu \quad \text{in } \mathbb{R} \times [0, T[. \end{aligned} \tag{2.58}$$

Notice that for every u in V_m,

$$u \leqq Mu \quad \text{in } \mathbb{R} \times [0, T[\tag{2.59}$$

is equivalent to

$$Bu \leqq 0 \quad \text{a.e. in } \mathbb{R} \times]0, T[, \tag{2.60}$$

where the operators A, B and M are defined by (2.1), (2.2) and (2.6).

THEOREM 2.5. *Let the assumptions* (1.7), $\cdots$, (1.10) *hold. Then problem* (2.58) *admits a maximum solution* $\hat{u}$, *which is given explicitly as the optimal cost* (1.4) *and satisfies* (2.8) *and* (2.10). *Moreover, defining*

$$x^*(t) = \inf \left\{ x : \frac{\partial \hat{u}}{\partial x}(x, t) + c(t) > 0 \right\}, \tag{2.61}$$

we have for almost every (x, t) *in* $\mathbb{R} \times]0, T[$

$$\begin{aligned} &A\hat{u} = f \quad \text{and} \quad B\hat{u} \leqq 0 \quad \text{if } x \geqq x^*(t), \\ &A\hat{u} \leqq f \quad \text{and} \quad B\hat{u} = 0 \quad \text{if } x \leqq x^*(t). \end{aligned} \tag{2.62}$$

Proof. The first part is obtained from Theorem 2.4 by letting ε tend to zero. It is clear that we also apply Theorem 2.2 and Corollary 2.1. Note that $\hat{u}^\varepsilon$ satisfies the dynamic programming principle (2.9).

In order to prove (2.62), we approximate the optimal cost (1.4) by the equation (2.33). Since the estimates (2.8) and (2.10) hold uniformly in $\varepsilon > 0$, for the solution $\hat{u}^\varepsilon$ of (2.33), we deduce

$$\lim_{\varepsilon \downarrow 0} B\hat{u}^\varepsilon \leqq 0, \tag{2.63}$$

which implies

$$B\hat{u} \leqq 0 \quad \text{in } \mathbb{R} \times [0, T[. \tag{2.64}$$

Let (x, t) be a point in $\mathbb{R} \times [0, T[$ at which $A\hat{u}$, $A\hat{u}^\varepsilon$, $0 < \varepsilon \leqq 1$ exist and are such that $x > x^*(t)$. Since $\hat{u}$ is convex, we have

$$B\hat{u} < 0 \quad \text{at } (x, t);$$

hence, for ε sufficiently small

$$B\hat{u}_\varepsilon < 0 \quad \text{at } (x, t),$$

and from the equation (2.33), we deduce

$$A\hat{u} = f \quad \text{at } (x, t).$$

This verifies (2.62) and the proof is completed. Note that the idea of this theorem can be traced back to [41]. □

3. The free boundary. Define the differential operator

$$A'u = -\frac{\partial u}{\partial t} - \frac{1}{2}\sigma^2(t)\frac{\partial^2 u}{\partial x^2} - (a(t)x + b(t))\frac{\partial u}{\partial x} + (\alpha(t) - a(t))u \tag{3.1}$$

and the substitutions

$$w = -\frac{\partial u}{\partial x} - c(t), \tag{3.2}$$

$$g = \frac{dc}{dt} - (\alpha(t) - a(t))c(t) - \frac{\partial f}{\partial x}, \tag{3.3}$$

for the given functions u and f.

If u solves (2.4), then by taking formal derivative with respect to the variable x, we can deduce the equation

$$\begin{aligned} &(A'w - g) \vee w = 0 \quad \text{in } \mathbb{R} \times [0, T[, \\ &w(\cdot, T) = 0 \quad \text{in } \mathbb{R}, \end{aligned} \tag{3.4}$$

to be satisfied by the optimal cost $\hat{u}$, defined in (1.4), through the transformation (3.2). It is clear that (3.4) represents a classical variational inequality in the unknown w (e.g. Bensoussan and Lions [8], Friedman [21], Kinderlehrer and Stampacchia [31]). In this connection with optimal stopping, we refer to Bather and Chernoff [4], Karatzas [28] and more recently to Karatzas and Shreve [29], [30]. Moreover, the solution w of (3.4) has a stochastic representation as the optimal cost of a stopping time problem, i.e.

$$w(x, t) = \inf\{S_{xt}(\theta)\colon t \leqq \theta \leqq T, \text{ stopping time}\}, \tag{3.5}$$

where

$$S_{xt}(\theta) = E\left\{\int_t^\theta g(y^0(s), s) \exp\left(-\int_t^s (\alpha(\lambda) - a(\lambda))\, d\lambda\right) ds\right\}, \tag{3.6}$$

and the process $y^0(s) = y^0_{xt}(s)$ is given by (1.2) with the control $\nu = 0$.

Then, with the function $w(x, t)$ we can define the moving boundary $x(t)$, $0 \leqq t < T$, by

$$x(t) = \inf\{x\colon w(x, t) < 0\}. \tag{3.7}$$

As in Bather and Chernoff [4], Benes et al. [6], Karatzas [27], Menaldi and Robin [41], the reflected diffusion process on the half-space $[x \geqq x(t)]$ will prove an optimal control for the original problem (1.4), (1.5). It is clear that the use of the variational inequality (3.4) will help us to obtain enough regularity of the free boundary (3.7) in order to be able to construct the reflected diffusion process.

First we consider the problem (3.4) and next the optimal control related to the free boundary (3.7).

3.1. Variational inequality. Consider the penalizing function

$$\beta(\lambda) = \begin{cases} 0 & \text{if } \lambda \leqq 0, \\ \lambda^2 & \text{if } 0 \leqq \lambda \leqq 1, \\ 2\lambda - 1 & \text{if } \lambda \geqq 1, \end{cases} \tag{3.8}$$

and the set of controls $\tilde{\mathcal{V}}_\varepsilon$, $\varepsilon > 0$, defined by

(3.9) (η, ξ) belongs to $\tilde{\mathcal{V}}_\varepsilon$ if $\eta(t)$, $\xi(t)$ are progressively measurable random processes from $[0, +\infty[$ into $\mathbb{R}$ such that for every $t \geqq 0$ and λ in $\mathbb{R}$,

$$\lambda\eta(t) - \frac{1}{\varepsilon}\beta(\lambda) \leqq \xi(t) \leqq \frac{1}{\varepsilon}.$$

Note that if (η, ξ) belongs to $\tilde{\mathcal{V}}_\varepsilon$, then by looking at the graph of $\beta(\lambda)$, we deduce for every $t \geqq 0$,

$$0 \leqq \eta(t) \leqq \frac{2}{\varepsilon}, \qquad 0 \leqq \xi(t) \leqq \frac{1}{\varepsilon}. \tag{3.10}$$

In order to be able to derive the variational inequality (3.4), we introduce another penalized problem,

$$\hat{u}^\varepsilon(x, t) = \inf\{\tilde{J}_{xt}(\eta, \xi) : (\eta, \xi) \text{ in } \tilde{\mathcal{V}}_\varepsilon\}, \tag{3.11}$$

$$\tilde{J}_{xt}(\eta, \xi) = E\left\{\int_t^T (f(y(s), s) + c(s)\eta(s) + \xi(s)) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) ds\right\}, \tag{3.12}$$

with $s \geqq t$,

$$y(s) = x + \int_t^s (a(\lambda)y(\lambda) + b(\lambda) + \eta(\lambda - t))\, d\lambda + \int_0^{s-t} \sigma(\lambda + t)\, dw(\lambda), \tag{3.13}$$

i.e., the equation (1.2) for

$$\nu(s) = \int_0^s \eta(\lambda)\, d\lambda, \qquad s \geqq 0. \tag{3.14}$$

The Hamilton-Jacobi-Bellman equation associated with the above penalized problem is precisely the following:

(3.15) Find $\hat{u}^\varepsilon$ in V_m such that $\partial^2\hat{u}^\varepsilon/\partial x^2$ belongs to L^∞_{loc},

$$\hat{u}^\varepsilon(x, T) = 0 \quad \text{for every } x \text{ in } \mathbb{R},$$

$$A\hat{u}^\varepsilon + \frac{1}{\varepsilon}\beta(B\hat{u}^\varepsilon) = f, \quad \text{a.e. in } \mathbb{R} \times]0, T[,$$

where the operators A, B are given by (2.1), (2.2) and the spaces V_m, L^∞_{loc} are defined in (2.32).

THEOREM 3.1. *Under the hypotheses* (1.7), $\cdots$, (1.10) *the optimal cost* $\hat{u}^\varepsilon$ *defined by* (3.11) *is a nonnegative continuous function such that for some constants* $0 < c \leqq C$, *the same* $m \geqq 1$ *for the assumption* (1.9), *and every* $0 < \varepsilon \leqq 1$, (x, t), (x', t') *in* $\mathbb{R} \times [0, T]$ *we have*

$$\begin{aligned}
&c|x^+|^m - C \leqq \hat{u}^\varepsilon(x, t) \leqq C(1 + |x|^m),\\
&|\hat{u}^\varepsilon(x, t) - \hat{u}^\varepsilon(x', t)| \leqq C(1 + |x|^{m-1} + |x'|^{m-1})|x - x'|,\\
&|\hat{u}^\varepsilon(x, t) - \hat{u}^\varepsilon(x, t')| \leqq C(1 + |x|^m)|t - t'|,\\
&0 \leqq \frac{\partial^2\hat{u}^\varepsilon}{\partial x^2}(x, t) \leqq C(1 + |x|^q), \qquad q = (m-2)^+,
\end{aligned} \tag{3.16}$$

so, $\hat{u}^\varepsilon$ *is convex in the first variable. Moreover, the partial differential equation* (3.15)

has one and only one solution, which is precisely the function $\hat{u}^\varepsilon$. *Furthermore,* $\hat{u}^\varepsilon$ *converges to the optimal cost* $\hat{u}$, *given by* (1.4), *as* ε *approaches zero.*

Proof. Use the same technique of Theorems 2.1, 2.2 and 2.3. □

Now, we differentiate (3.15) with respect to the variable x and let ε tend to zero, to obtain the variational inequality (3.4). Due to the lack of an a priori estimate of the mixed derivative of $\hat{u}^\varepsilon$ in x, t, we prefer to use a weak formulation of (3.4) in the sense of Mignot and Puel [48]. However, that estimate will be obtained later on by means of the interpretation (3.5).

Consider the weighted norm, $p > 2m+1$,

$$\|v\|_p = \left(\int_{\mathbb{R}} |v(x)|^2 (1+|x|^2)^{-p}\, dx \right)^{1/2} \tag{3.17}$$

and the Hilbert spaces

(3.18) H is the set of all real measurable functions v on $\mathbb{R}$ such that $\|v\|_p$ is finite,

(3.19) V is the set of all real measurable functions v on $\mathbb{R}$ with a derivative v' such that $\|v\|_p$ and $\|v'\|_{p-1}$ are finite.

Identifying H and its dual, we denoted by $\langle \cdot, \cdot \rangle$ the pairing between V', the dual, and V. The natural inner product in H is

$$(u, v) = \int_{\mathbb{R}} u(x) v(x) (1+|x|^2)^{-p}\, dx, \tag{3.20}$$

with the corresponding norm $|\cdot| = \|\cdot\|_p$, for a fixed p. Define the bilinear form, for t in $[0, T]$,

$$\begin{aligned} a(t, u, v) = \int_{\mathbb{R}} \Big[& \sigma^2(t) \left(\frac{\partial u}{\partial x}(x) \right) \left(\frac{\partial v}{\partial x}(x) - 2px(1+|x|^2)^{-1} v(x) \right) \\ & - (a(t)x + b(t)) \left(\frac{\partial u}{\partial x}(x) \right) v(x) + (\alpha(t) - a(t)) u(x) v(x) \Big] (1+|x|^2)^{-p}\, dx, \end{aligned} \tag{3.21}$$

which is continuous and strictly positive on V. Notice that for any smooth function $u(t) = u(t, x)$, we have

$$\left(-\frac{\partial u}{\partial t}(t), v \right) + a(t, u(t), v) = \langle A' u(t), v \rangle \tag{3.22}$$

for every v in V, with A' the differential operator (3.1).

Let $L^2(0, T; X)$ be the classical space of all square integrable functions on $]0, T[$ with values in a Hilbert space X. Introduce the problem:

Find w in $L^2(0, T; V)$, $w \leqq 0$ such that

$$\begin{aligned} & \int_0^T \left[\left\langle -\frac{\partial v}{\partial t}(t), v(t) - w(t) \right\rangle + a(t, w(t), v(t) - w(t)) \right] dt + \frac{1}{2} |v(T)|^2 \\ & \qquad \geqq \int_0^T (g(t), v(t) - w(t))\, dt, \end{aligned} \tag{3.23}$$

for every v in $L^2(0, T; V)$, with $\partial v/\partial t$ in $L^2(0, T; V')$ and $v \leqq 0$,

where the function $g(t) = g(t, x)$ is given by (3.3).

If $\hat{u}$ is the optimal cost (1.4), define

$$\hat{w} = -\frac{\partial \hat{u}}{\partial x} - c(t). \tag{3.24}$$

THEOREM 3.2. *Let the assumptions* (1.7), $\cdots$, (1.10) *hold. Suppose also that* $\sigma(t)$ *is nondegenerate, i.e.* (2.36). *Then the function* $\hat{w}$ *given by* (3.24) *is the maximum solution of the weak variational inequality* (3.23).

Proof. Note that from Mignot and Puel [48], we know that the problem (3.23) admits a maximum solution $\hat{w}$. This weak solution is actually a strong solution, i.e., it is smooth in t and satisfies (3.4) in a pointwise (a.e.) sense. However, the point is to identify that solution with (3.24).

Denote by $\beta'(\lambda)$ the derivative of the function (3.8),

$$\hat{w}^\varepsilon = -\frac{\partial \hat{u}^\varepsilon}{\partial x} - c(t), \tag{3.25}$$

with $\hat{u}^\varepsilon$ being the optimal cost (3.11). Since $\hat{u}^\varepsilon$ solves (3.15) and $\sigma(t)$ is nondegenerate, we are able to differentiate the equation (3.15) to obtain

$$\begin{aligned} &A'\hat{w}^\varepsilon - \frac{1}{\varepsilon}\beta'(\hat{w}^\varepsilon)\frac{\partial \hat{w}^\varepsilon}{\partial x} = g \quad \text{in } \mathbb{R}\times[0, T[, \\ &\hat{w}^\varepsilon(T, x) = 0 \quad \text{for every } x \text{ in } \mathbb{R}. \end{aligned} \tag{3.26}$$

The facts that $\hat{u}^\varepsilon(t, x)$ is convex in x and $\beta(\lambda)$ increasing, $\beta'(0) = 0$, imply that

$$\beta'(\hat{w}^\varepsilon) \geqq 0, \quad \hat{w}^\varepsilon \beta'(\hat{w}^\varepsilon) \geqq 0, \quad \frac{\partial \hat{w}^\varepsilon}{\partial x} \geqq 0 \quad \text{in } \mathbb{R}\times[0, T[.$$

Thus, an integration by parts in (3.26) gives

$$\begin{aligned} &\int_0^T \left[\left\langle -\frac{\partial v}{\partial t}(t), v(t) - \hat{w}^\varepsilon(t) \right\rangle + a(t, \hat{w}^\varepsilon(t), v(t) - \hat{w}^\varepsilon(t)) \right] dt + \frac{1}{2}|v(T)|^2 \\ &\qquad \geqq \int_0^T (g(t), v(t) - \hat{w}^\varepsilon(t))\, dt \end{aligned} \tag{3.27}$$

for every v in $L^2(0, T; V)$, with $\partial v/\partial t$ in $L^2(0, T; V')$, and $v \geqq 0$. Since the estimates (3.16) ensure that

$$\hat{w}^\varepsilon \to \hat{w} \text{ weakly in } L^2(0, T; V),$$

we have

$$\liminf_{\varepsilon \downarrow 0} \int_0^T a(t, \hat{w}^\varepsilon(t), \hat{w}^\varepsilon(t))\, dt \geqq \int_0^T a(t, \hat{w}(t), \hat{w}(t))\, dt.$$

Therefore, by means of the following bound, for some constant $C > 0$,

$$\beta(\hat{w}^\varepsilon) \leqq \varepsilon(f - A\hat{u}^\varepsilon) \leqq \varepsilon C(1 + |x|^m), \qquad \varepsilon > 0,$$

derived from Theorem 3.1, we take the limit in (3.27) as ε tends to zero in order to deduce that function $\hat{w}$, given by (3.24), is a solution of the weak variational inequality formulation (3.23).

Now, we prove that for any solution w of the problem (3.23)

$$w \leqq \hat{w}^\varepsilon \quad \text{for every } \varepsilon > 0. \tag{3.28}$$

Indeed, if $z = w - \hat{w}^\varepsilon$ from (3.23) and (3.26) we obtain

$$\int_0^T \left[\left\langle -\frac{\partial v}{\partial t}(t), v(t) - z(t) \right\rangle + a(t, z(t), v(t) - z(t)) \right] dt + \frac{1}{2}|v(T)|^2 \geqq \int_0^T (q(t), v(t) - z(t))\, dt, \tag{3.29}$$

for every v in $L^2(0, T; V)$, with $\partial v/\partial t$ in $L^2(0, T; V')$ and $v \geqq \hat{w}^\varepsilon$, where

$$q(x, t) = \frac{1}{\varepsilon} \beta'(\hat{w}^\varepsilon(x, t)) \frac{\partial \hat{w}^\varepsilon}{\partial x}(x, t). \tag{3.30}$$

Thus, by taking $v = \hat{w}^\varepsilon - \lambda\theta$, λ any arbitrary positive number, in (3.29), we may deduce

$$\int_0^T \left[\left\langle \frac{\partial \theta}{\partial t}(t), z(t) \right\rangle + a(t, z(t), \theta(t)) \right] dt \leqq \int_0^T (q(t), \theta(t))\, dt \tag{3.31}$$

for every $\theta = \theta(t)$ such that

$$\theta \text{ belongs to } L^2(0, T; V), \partial\theta/\partial t \text{ belongs to } L^2(0, T; V'), \theta(0) = 0 \text{ and } \theta \geqq 0. \tag{3.32}$$

Therefore, introducing θ_η as the solution of

$$\eta \frac{\partial \theta_\eta}{\partial t} + \theta_\eta = z^+, \qquad \theta_\eta(0) = 0, \tag{3.33}$$

we see that θ_η satisfies (3.32). Hence, from (3.31) with $\theta = \theta_\eta$, $\eta > 0$, we obtain

$$\int_0^T [a(t, z(t), \theta_\eta(t)) - (q(t), \theta_\eta(t))]\, dt \leqq 0.$$

Since $\theta_\eta \to z^+$ in $L^2(0, T; V)$, we have

$$\int_0^T [a(t, z(t), z^+(t)) - (q(t), z^+(t))]\, dt \leqq 0.$$

But, $z(t) \geqq 0$ implies $\hat{w}^\varepsilon \leqq w$, and note $w \leqq 0$. We have $q(t) = 0$. Thus

$$\int_0^T a(t, z(t), z^+(t))\, dt \leqq 0.$$

This means $z^+(t) = 0$, i.e. (3.28). This completes the proof. □

Recall the function space V_{m-1} as in (2.32), i.e.

v belongs to V_{m-1} if $v : \mathbb{R} \times [0, T] \to \mathbb{R}$ is locally Lipschitz continuous such that

$$|v(x, t)| + \left|\frac{\partial v}{\partial t}(x, t)\right| \leqq C(1 + |x|^{m-1}),$$
$$\left|\frac{\partial v}{\partial x}(x, v)\right| \leqq C(1 + |x|^q), \qquad q = (m-2)^+, \tag{3.34}$$

for every x, t and some constant C

and the space of distributions $\mathcal{D}'(\mathbb{R} \times]0, T[)$. Consider the variational inequality:

Find w in V_{m-1} such that

$$\begin{aligned} &w(x, T) = 0 \quad \text{for every } x \text{ in } \mathbb{R}, \\ &A'w \leqq g \quad \text{in } \mathscr{D}'(\mathbb{R}\times]0, T[), \\ &w \leqq 0 \quad \text{in } \mathbb{R}\times[0, T[. \end{aligned} \tag{3.35}$$

Suppose that

$$\left|\frac{\partial f}{\partial x}(x, t) - \frac{\partial f}{\partial x}(x, t')\right| \leqq C(1+|x|^{m-1})|t-t'|, \tag{3.36}$$

for every x, t and some constant C.

From this together with (1.9) we get $\partial f/\partial x$ belonging to V_{m-1}.

THEOREM 3.3. *Assume the hypotheses* (1.7), $\cdots$, (1.10) *and* (3.36) *hold. Then the variational inequality* (3.35) *admits a maximum solution* $\hat{w}$, *which is given explicitly as the optimal stopping cost* (3.5) *and* (3.24) *is true. Moreover, we have*

$$A'\hat{w} = g \quad \text{in } \mathscr{D}'([\hat{w} < 0]) \tag{3.37}$$

and $\hat{w}$ *is the unique solution of* (3.35) *and* (3.37) *simultaneously.*

Proof. First, suppose that $\sigma(t)$ is nondegenerate, i.e. (2.36). Then, as was described in the proof of Theorem 2.4, by applying the classical results we deduce that the function $\hat{w}$, defined by (3.5), solves the variational inequality (3.35), (3.37).

On the other hand, by means of the assumption (3.36) we can show that

$$|\hat{w}(x, t) - \hat{w}(x, t')| \leqq C(1+|x|^{m-1})|t-t'|, \tag{3.38}$$

for every x, t and some constant C.

Therefore, if w is a solution of the weak variational inequality (3.23), we claim that $w = \hat{w}$, the unique solution of problem (3.35), (3.37), i.e, the optimal cost (3.5). Indeed, an integration by parts in (3.35), (3.37) yields

$$\begin{aligned} &\int_0^T \left[\left(-\frac{\partial \hat{w}}{\partial t}(t), v(t) - \hat{w}(t)\right) + a(t, \hat{w}(t), v(t) - \hat{w}(t))\right] dt \\ &\qquad \geqq \int_0^T (g(t), v(t) - \hat{w}(t))\, dt, \end{aligned} \tag{3.39}$$

for every v in $L^2(0, T; V)$ such that $v \leqq 0$,

after using the property (3.22). Hence, adding (3.39) with $v = w$, to (3.23) with $v = \hat{w}$, we get

$$-\int_0^T a(t, w(t) - \hat{w}(t), w(t) - \hat{w}(t)) \geqq 0$$

which implies $w = \hat{w}$.

To study the degenerate case, i.e. to drop assumption (2.36), we regularize the problem by changing $\sigma(t)$ into

$$\sigma_\eta(t) = (\sigma^2(t) + \eta)^{1/2}, \qquad \eta > 0. \tag{3.40}$$

Since the estimate (3.38) is uniform in η and the expression (3.5) is stable as η tends to zero, we can complete the proof in a similar way as in [37], [38]. □

Remark 3.1. If $\hat{u}(x, t)$ is the optimal cost (1.4), then

$$\frac{\partial \hat{u}}{\partial x}(x, t) = \sup\{G_{xt}(\theta): 0 \leqq \theta \leqq T, \text{ stopping time}\}, \tag{3.41}$$

where

$$G_{xt}(\theta) = \left\{ \int_t^\theta \frac{\partial f}{\partial x}(y^0(s), s) \exp\left(-\int_t^s (\alpha(\lambda) - a(\lambda))\, d\lambda \right) ds \right. \\ \left. + c(\theta) \exp\left(-\int_t^\theta (\alpha(\lambda) - a(\lambda))\, d\lambda \right) \right\}, \tag{3.42}$$

and $y^0(s) = y^0_{xt}$ is the process (1.2) with $\nu = 0$.

Remark 3.2. The Theorem 3.2 holds without assuming the nondegeneracy condition (2.36). On the other hand, if we do not include regularity in t for the definition of the space V_{m-1}, i.e., $v(x, t)$ continuous in (x, t) but locally Lipschitz only in the variable x, then, the conclusion of Theorem 3.3 is true without the hypothesis (3.36) on f. Clearly, in that case, the optimal cost $\hat{w}(x, t)$ is continuous in (x, t) but locally Lipschitz only in the variable x.

3.2. Optimal decision. First we give an abstract result about the existence of an optimal policy.

THEOREM 3.4. *Under the assumptions* (1.7), $\cdots$, (1.10), *and* $m > 1$, *there exists an optimal control* $\hat{\nu}$ *in* $\mathcal{V}$ *of the initial problem* (1.4), (1.5).

Proof. Let t be fixed in $[0, T[$ and consider the norm

$$\|\nu\|_m = \left(E\left\{ \int_0^{T-t} |\nu(s)|^m\, ds \right\} \right)^{1/m}. \tag{3.43}$$

Noting that for $0 \leqq s < T - t$

$$(E\{|\nu(s)|^m\})^{1/m} \leqq (T - t - s)^{-1/m} \|\nu\|_m, \qquad 0 \leqq s < T - t, \tag{3.44}$$

and the linear character of the state equation (1.2) we have:

$$\text{If } (\nu_n, n = 0, 1, \cdots) \text{ is a sequence in } \mathcal{V} \text{ such that} \\ \|\nu_n - \nu_0\|_m \to 0, \quad E\{|\nu_n(0) - \nu_0(0)|^m\} \to 0, \quad |E\{\nu_n(T - t) - \nu_0(T - t)\}| \to 0, \\ \text{then } J_{xt}(\nu_n) \to J_{xt}(\nu_0) \text{ as } n \to \infty, \tag{3.45}$$

and

$$\text{the mapping } \nu \to J_{xt}(\nu) \text{ is convex from } \mathcal{V} \text{ into } \mathbb{R}, \tag{3.46}$$

where $J_{xt}(\nu)$ is the functional (1.3).

By means of the hypotheses (1.8), relative to $c(t)$, and (1.9) we deduce that

$$J_{xt}(\nu) \to +\infty \text{ as } \|\nu\|_m \to \infty \text{ and, unless } c(t) = 0 \text{ for every } t, \\ \text{also as } E\{|\nu(T - t)|\} \to \infty. \tag{3.47}$$

Thus, there is a sequence $(\nu'_n, n = 1, 2, \cdots)$ in $\mathcal{V}$ and a ν'_0 in $L^m(]0, T - t[\times \Omega)$, the space of m-integrable functions, such that as n goes to infinity

$$J_{x_t}(\nu'_n) \to \hat{u}(x, t), \ \nu'_n \to \nu'_0 \text{ weakly in } L^m, \text{ and } \|\nu'_n\|_m + |E\{\nu'_n(T - t)\}| \leqq C, \\ \text{for some constant } C. \tag{3.48}$$

Hence, we can define $(\nu_n,\ n=1,2,\cdots)$ in $\mathcal{V}$ as a convex combination of $(\nu'_n,\ n=1,2,\cdots)$,

$$\nu_n=\sum_{i=n}^{n+k}\alpha_i^n\nu'_i,\qquad \alpha_i^n \text{ in } [0,1] \text{ with } \sum_{i=n}^{n+k}\alpha_i^n=1 \tag{3.49}$$

and a nonnegative increasing function $q(s)$, $0\leqq s\leqq T-t$ satisfying

$$\nu_n\to\nu'_0 \text{ strongly in } L^m, \text{ and } E\{\nu_n(s)\}\to q(s) \text{ for every } s \text{ in } [0,\ T-t]. \tag{3.50}$$

Moreover, if N is a countable subset of $[0,\ T-t[$, a similar argument to the previous one and the inequality (3.44) allow us to show that $\nu_n(s)$ is strongly convergent in $L^m(\Omega)$ for every s in N; in particular we may assume that

$$\nu_n(s)\to\nu'_0(s) \text{ strongly in } L^m(\Omega) \text{ and almost surely in } \Omega, \text{ for every rational in } [0,\ T-t[.$$

Clearly, $\nu'_0(\cdot)$ is nonnegative, increasing and progressively measurable. Define

$$\nu_0(s)=\inf\{\nu'_0(s'):\ s'>t,\ s' \text{ rational}\} \tag{3.51}$$

which is right continuous having left-hand limits, adapted and $\nu_0=\nu'_0$ in $L^m(]0,\ T[\times\Omega)$, $\nu_0(0)=\nu'_0(0)$. Hence, for an eventual subsequence if necessary, from (3.50) we have

$$\nu_n\to\nu_0 \text{ strongly in } L^m,\ \nu_n(0)\to\nu_0(0) \text{ strongly in } L^m(\Omega), \text{ and } E\{\nu_n(T-t)\}\to q(T-t), \text{ as } n\to\infty, \tag{3.52}$$

and if

$$\nu_0(T-t)=\sup\{\nu_0(s):\ 0\leqq s<T-t\} \tag{3.53}$$

then

$$E\{\nu_0(s)\}=q(s)$$

provided both functions are continuous at s in $[0,\ T-t]$. Since

$$E\left\{\int_t^T c(s)\exp\left(-\int_t^s\alpha(\lambda)\,d\lambda\right)d\hat{\nu}(s-t)\right\}=\int_t^T c(s)\exp\left(-\int_t^s\alpha(\lambda)\,d\lambda\right)dq(s-t),$$

where

$$\hat{\nu}(s)=\begin{cases}\nu_0(s) & \text{if } 0\leqq s\leqq T-t,\\ \nu_0(T-t) & \text{if } s\geqq T-t,\end{cases} \tag{3.54}$$

we may deduce from (3.45) that $\hat{\nu}$ belongs to $\mathcal{V}$ and

$$J_{xt}(\nu_n)\to J_{xt}(\hat{\nu}). \tag{3.55}$$

But, based on the convexity properties (3.46), (3.49), we have

$$J_{xt}(\nu_n)\leqq\sum_{i=n}^{n+k}\alpha_i^n J_{xt}(\nu'_i)$$

and from (3.48), for every $\varepsilon>0$,

$$J_{xt}(\nu'_i)\leqq\hat{u}(x,t)+\varepsilon \quad \text{if } i\geqq n(\varepsilon)$$

which implies

$$J_{xt}(\nu_i)\leqq\hat{u}(x,t)+\varepsilon \quad \text{for every } i\geqq n(\varepsilon).$$

Therefore, we obtain with (3.55)

$$J_{xt}(\hat{\nu}) = \hat{u}(x, t)$$

and the proof is completed. □

Now we give a constructive approach of the optimal control through the free boundary (3.7).

Let $\hat{u}(x, t)$ be the optimal cost (1.4), for $0 \leqq t < T$ define

$$x^*(t) = \inf\left\{x: \frac{\partial \hat{u}}{\partial x}(x, t) + c(t) > 0\right\} \tag{3.56}$$

and suppose

(3.57) $x^*(t)$ is finite and can be extended to a continuous function on $[0, T]$.

Some sufficient conditions to ensure (3.57) will be given later on. Note that in order to determine the free boundary (3.56) we need only to know the function $\hat{w}(x, t)$, which is the unique solution of the variational inequalities (3.35), (3.37).

THEOREM 3.5. *Let the hypotheses* (1.7), $\cdots$, (1.10) *and* (3.57) *hold. Then there exists a control* $\hat{\nu}$ *in* $\mathcal{V}$ *whose associated state* $y(s) = y_{xt}(s, \hat{\nu})$, *defined by the stochastic equation* (1.2), *satisfies*

(3.58) $y(s) \geqq x^*(s)$, *for every* $t \leqq s \leqq T$, $\int_t^T I(y(s) > x^*(s))\, d\hat{\nu}(s-t) = 0$, $I(\cdot)$ *denotes the characteristic function, and* $\hat{\nu}(0) = (x^*(t) - x)^+$.

Moreover, the process $\hat{\nu}$ *is continuous, uniquely determined by the conditions* (1.2), (3.58) *and finally, the control* $\hat{\nu}$ *is optimal, i.e.,* (1.5) *is valid.*

Proof. It is clear that $y(s)$ is the reflected diffusion on the continuation set $[y \geqq x^*(s)]$ with initial value $x \vee x^*(t)$ at the time t. Since we assume $x^*(t)$, $0 \leqq t \leqq T$ to be only continuous, it is necessary to make precise the classical arguments about the existence of the reflected diffusion. Indeed, let $(x_\varepsilon(s),\ 0 < \varepsilon \leqq 1)$ be a smooth approximation of $x^*(s)$, i.e. $x_\varepsilon(s)$ has a continuous derivative $\dot{x}_\varepsilon(s)$, $x_\varepsilon(t) = x^*(t)$ and

$$x_\varepsilon(s) \to x^*(s), \quad \text{uniformly in } [t, T] \quad \text{as } \varepsilon \to 0. \tag{3.59}$$

We define the processes $(z_\varepsilon(s), \eta_\varepsilon(s), t \leqq s \leqq T)$, which are continuous and progressively measurable, as the unique solution of the stochastic equations, $t \leqq s \leqq T$

$$\begin{aligned} z_\varepsilon(s) &= (x - x^*(t))^+ + \int_t^s (a(\lambda) z_\varepsilon(\lambda) + b(\lambda))\, d\lambda + \int_t^s \sigma(\lambda)\, dw(\lambda - t) \\ &\quad + \int_t^s (a(\lambda) x_\varepsilon(\lambda) + \dot{x}_\varepsilon(\lambda))\, d\lambda + \int_t^s I(z_\varepsilon(\lambda) = 0)\, d\eta_\varepsilon(\lambda), \\ \eta_\varepsilon(0) &= 0, \qquad \eta_\varepsilon(s) - \eta_\varepsilon(\lambda) \geqq 0 \quad \text{for every } T \geqq s \geqq \lambda \geqq t, \\ z_\varepsilon(s) &\geqq 0 \quad \text{for every } T \geqq s \geqq t, \quad \text{and} \quad \int_t^T I(z_\varepsilon(s) > 0)\, d\eta_\varepsilon(s) = 0. \end{aligned} \tag{3.60}$$

Thus, if $y_\varepsilon(s) = z_\varepsilon(s) + x_\varepsilon(s)$, $t \leqq s \leqq T$, we have

$$
\begin{aligned}
y_\varepsilon(s) &= x \vee x^*(t) + \int_t^s (a(\lambda) y_\varepsilon(\lambda) + b(\lambda))\, d\lambda + \int_t^s \sigma(\lambda)\, dw(\lambda - t) \\
&\quad + \int_t^s I(y_\varepsilon(\lambda) = x_\varepsilon(\lambda))\, d\eta_\varepsilon(\lambda), \\
&y_\varepsilon(s) \geqq x_\varepsilon(s), \quad \text{for every } T \geqq s \geqq t, \\
&\int_t^T I(y_\varepsilon(s) > x_\varepsilon(s))\, d\eta_\varepsilon(s) = 0.
\end{aligned} \tag{3.61}
$$

Since

$$
y_\varepsilon(s) - y_{\varepsilon'}(s) = \int_t^s a(\lambda)(y_\varepsilon(\lambda) - y_{\varepsilon'}(\lambda))\, d\lambda + \eta_\varepsilon(s) - \eta_{\varepsilon'}(s),
$$

an integration by parts yields

$$
\begin{aligned}
|y_\varepsilon(s) - y_{\varepsilon'}(s)|^2 &= 2 \int_t^s a(\lambda) |y_\varepsilon(\lambda) - y_{\varepsilon'}(\lambda)|^2\, d\lambda \\
&\quad + 2 \int_t^s (y_\varepsilon(\lambda) - y_{\varepsilon'}(\lambda))\, d\eta_\varepsilon(\lambda) - 2 \int_t^s (y_\varepsilon(\lambda) - y_{\varepsilon'}(\lambda))\, d\eta_{\varepsilon'}(\lambda).
\end{aligned}
$$

But the last two terms are equal to

$$
\begin{aligned}
&2 \int_t^s (x_\varepsilon(\lambda) - y_{\varepsilon'}(\lambda))\, d\eta_\varepsilon(\lambda) + 2 \int_t^s (x_{\varepsilon'}(\lambda) - y_\varepsilon(\lambda))\, d\eta_{\varepsilon'}(\lambda) \\
&\quad \leqq 2 \int_t^s (x_\varepsilon(\lambda) - x_{\varepsilon'}(\lambda))\, d\eta_\varepsilon(\lambda) + 2 \int_t^s (x_{\varepsilon'}(\lambda) - x_\varepsilon(\lambda))\, d\eta_{\varepsilon'}(\lambda).
\end{aligned}
$$

Hence, by Gronwall's inequality we deduce

$$
|y_\varepsilon(s) - y_{\varepsilon'}(s)|^2 \leqq C(\eta_\varepsilon(T) + \eta_{\varepsilon'}(T)) \sup \{|x_\varepsilon(s) - x_{\varepsilon'}(s)|\colon t \leqq s \leqq T\}, \tag{3.62}
$$

for every $t \leqq s \leqq T$ and some deterministic constant C depending on T. Similarly, taking some $q \geqq 1 + x_\varepsilon(s)$, for every $t \leqq s \leqq T$, $0 < \varepsilon \leqq 1$, we obtain

$$
|y_\varepsilon(s) - q|^2 + \eta_\varepsilon(s) \leqq \exp\left(2 \int_t^T |a(\lambda)|\, d\lambda\right), \tag{3.63}
$$

for every s in $[t, T]$. Now, letting ε go to zero and using the estimates (3.62), (3.63), we get two continuous and progressively measurable $(y(s), \eta(s), t < s < T)$ such that

$$
\begin{aligned}
y(s) &= x \vee x^*(t) + \int_t^s (a(\lambda) y(\lambda) + b(\lambda))\, d\lambda + \int_t^s \sigma(\lambda)\, dw(\lambda - t) \\
&\quad + \int_t^s I(y(\lambda) = x^*(\lambda))\, d\eta(\lambda), \\
&\eta(0) = 0, \quad \eta(s) - \eta(\lambda) \geqq 0, \quad y(s) \geqq x^*(s) \\
&\text{for every } T \geqq s \geqq \lambda \geqq t, \text{ and } \int_t^T I(y(s) > x^*(s))\, d\eta(s) = 0.
\end{aligned} \tag{3.64}
$$

So the process $\hat{v}$ is defined by

$$\hat{v}(s) = \begin{cases} (x^*(t)-x)^+ + \eta(t+s) & \text{if } 0 \leqq s \leqq T-t, \\ (x^*(t)-x)^+ + \eta(T) & \text{if } s \geqq T-t. \end{cases} \tag{3.65}$$

It remains to prove that $\hat{v}$ is optimal. Indeed, let us assume that there is no degeneracy, i.e. (2.36); then the optimal cost (1.4) is smooth enough to apply Itô's formula for a semimartingale (cf. Meyer [47]) in order to get, for every v in $\mathcal{V}$

$$\begin{aligned} E\left\{\hat{u}(t, x+v(0)) - \hat{u}(T, y(T)) \exp\left(-\int_t^T \alpha(s)\, ds\right)\right\} \\ = E\left\{\int_t^T A\hat{u}(s, y(s)) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) ds \right. \\ - \int_t^T \frac{\partial \hat{u}}{\partial x}(s, y(s)) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) dv(s-t) \\ \left. - \sum_{t<s\leqq T} [\hat{u}(s, y(s)) - \hat{u}(s, y(s-))] \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right)\right\}, \end{aligned} \tag{3.66}$$

where $y(s-)$ denotes the limit from the left at s. Since $\hat{u}(T, \cdot) = 0$ in $\mathbb{R}$, $A\hat{u} \leqq f$, $-\partial\hat{u}/\partial x \leqq c(\cdot)$ in $\mathbb{R} \times [0, T[$ and

$$-[\hat{u}(s, y(s)) - \hat{u}(s, y(s-))] \leqq c(s)[y(s) - y(s-)] = c(s)[v(s) - v(s-)]$$

we deduce

$$\hat{u}(t, x) \leqq J_{xt}(v), \quad \text{for every } v \text{ in } \mathcal{V}. \tag{3.67}$$

Similarly, choosing $\hat{v}$ given by (3.65), we obtain from (3.66)

$$\hat{u}(t, x) = J_{xt}(\hat{v}), \tag{3.68}$$

after using the fact that

$$A\hat{u}(s, y) = f(s, y) \quad \text{if } y \geqq x^*(s), \quad T \geqq s \geqq 0,$$

$$\frac{\partial \hat{u}}{\partial x}(s, y) = -c(s) \quad \text{if } y \leqq x^*(s), \quad T \geqq s \geqq 0.$$

Until now, we have established the optimality of control $\hat{v}$ under the assumption (2.36). In order to remove the nondegeneracy (2.36), let us consider the function $(\hat{u}_\varepsilon, 0 < \varepsilon \leqq 1)$ given as the optimal cost (1.4) with a covariance

$$\sigma_\varepsilon(t) = (\sigma^2(t) + \varepsilon)^{1/2}$$

instead of $\sigma(t)$. We have, as ε tends to zero

(3.69) $\hat{u}_\varepsilon \to \hat{u}$, $\partial\hat{u}_\varepsilon/\partial x \to \partial\hat{u}/\partial x$ locally uniform in $\mathbb{R} \times [0, T]$, and $\partial^2\hat{u}_\varepsilon/\partial x^2$ locally bounded in $\mathbb{R} \times [0, T]$.

Since (3.67) holds for $\hat{u}_\varepsilon$, $0 < \varepsilon \leqq 1$, we obtain the same inequality as the limit when ε goes to zero. Now, the Itô's formula (3.66) for the control $q + \hat{v}$, $q > 0$, yields

$$\begin{aligned} \hat{u}_\varepsilon(t, x+\hat{v}(0)+q) = E\left\{\int_t^T A_\varepsilon \hat{u}_\varepsilon(s, y(s)+q) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) ds \right. \\ \left. - \int_t^T \frac{\partial \hat{u}_\varepsilon}{\partial x}(s, y(s)+q) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) d\hat{v}(s-t)\right\}. \end{aligned}$$

Because of $y(s) \geqq x^*(s)$ and

$$0 \leqq f(s, y+q) - A_\varepsilon \hat{u}_\varepsilon(s, y+q) \leqq C(1+|y|^m) I\left(\frac{\partial \hat{u}_\varepsilon}{\partial x}(s, x^*(s)+q) = 0\right),$$

where C is a constant, $I(\cdot)$ denotes the characteristic function, we deduce, by means of (3.69) as ε tends to zero

$$\hat{u}(t, x+\hat{\nu}(0)+q) = E\left\{\int_t^T f(s, y(s)+q) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) ds \right.$$
$$\left. - \int_t^T \frac{\partial \hat{u}}{\partial x}(s, x^*(s)+q) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) d\hat{\nu}(s-t)\right\}.$$

Thus the equality (3.68) follows when q becomes zero. Therefore, the proof is completed. □

Remark 3.3. A way to approximate the solution $(y(s), \eta(s), t \leqq s \leqq T)$ of (3.64) is by solving the Itô's equation

$$\begin{aligned} dy^\varepsilon(s) &= (a(s)y^\varepsilon(s) + b(s))\, ds + \sigma(s)\, dw(s-t) \\ &\quad + \frac{1}{\varepsilon}(x^*(s) - y^\varepsilon(s))^+ ds, \qquad T \geqq s \geqq t, \\ y^\varepsilon(t) &= x \vee x^*(t). \end{aligned} \tag{3.70}$$

Similar to [39], it can be proved that for every $1 \leqq p < \infty$,

$$\begin{aligned} &E(\sup\{|y^\varepsilon(s) - y(s)|^p : t \leqq s \leqq T\}) \to 0, \\ &E\left(\sup\left\{\left|\eta(s) - \frac{1}{\varepsilon}\int_t^s (x^*(\lambda) - y^\varepsilon(\lambda))^+ d\lambda\right|^p : t \leqq s \leqq T\right\}\right) \to 0 \end{aligned} \tag{3.71}$$

as ε tends to zero. This provides an approximation of the optimal control $\hat{\nu}$. □

Remark 3.4. Considering the solution $\hat{u}(t, x)$ of (2.62), i.e., the optimal cost (1.4), for $x^*(t) \pm \varepsilon$ and letting ε go to zero, we obtain

$$\begin{aligned} A\hat{u}(t, x^*(t)+) &= f(t, x^*(t)) \\ &\geqq -\frac{\partial \hat{u}}{\partial t}(t, x^*(t)-) + (a(t)x^*(t) + b(t))c(t) + \alpha(t)\hat{u}(t, x^*(t)) \end{aligned} \tag{3.72}$$

which implies

$$\frac{\partial \hat{u}}{\partial t}(t, x^*(t)+) - \frac{\partial \hat{u}}{\partial t}(t, x^*(t)-) \geqq -\sigma^2(t)\frac{\partial^2 \hat{u}}{\partial x^2}(t, x^*(t)+) \geqq 0. \tag{3.73}$$

So, the first derivative of $\hat{u}(t, x)$ with respect to t has a nonnegative jump at $x = x^*(t)$ and if that jump vanishes and $\sigma(t) \neq 0$, then the second derivative of $\hat{u}$ with respect to x is continuous throughout the free boundary $x^*(t)$. The last observation can be deduced also from the classical regularity on the function $\hat{w}$, a solution of the variational inequality (3.35), (3.37). □

Remark 3.5. Even under degeneracy, it can be proved (cf. [37]) that $A\hat{w}$ is locally bounded, more precisely as in Lewy and Stampacchia [34] we have

$$-g^- \leqq A\hat{w} \leqq g, \quad \text{a.e. in } \mathbb{R} \times]0, T[\tag{3.74}$$

where $\hat{w}$ is the solution of the variational inequality (3.35), (3.37), i.e., $\hat{w}$ is given by

either (3.5) or (3.24) and g by (3.3). This implies, using the standard regularity results for parabolic partial differential equations,

(3.75) $\partial\hat{w}/\partial t$, $\partial^2\hat{w}/\partial x^2$ are essentially locally bounded in (x, t) belonging to $\mathbb{R}\times[0, T]$ such that $\sigma(t)\neq 0$

and also

(3.76) $\partial\hat{w}/\partial t$ is essentially locally bounded in x belonging to $\mathbb{R}$ for almost every t such that $\sigma(t)\neq 0$.

Clearly, from (3.24), (3.75) and (3.76) we deduce that

(3.77) for almost every t in $[0, T]$ the function $\partial\hat{u}/\partial t$ is continuous in the variable x belonging to $\mathbb{R}$.

Note that (3.77) holds under the assumptions (1.7), $\cdots$, (1.10), and that (3.73) is actually an equality. □

Remark 3.6. Going through the proof of Theorem 3.5 we notice that the continuity of the free boundary $x^*(t)$, given by (3.56), at the end point $t=T$ is not really used. It suffices to suppose

(3.78) $x^*(t)$ is continuous and bounded from above on $[0, T[$

in lieu of (3.57). □

Remark 3.7. Define the function $(q(t), 0\leqq t\leqq T)$ by

$$q(t)=\sup\left\{x: \frac{\partial f}{\partial x}(x, t)\leqq\frac{dc}{dt}(t)-(\alpha(t)-a(t))c(t)\right\} \tag{3.79}$$

which is bounded in view of the hypotheses (1.8) and (1.9) if $m>1$. The function (3.79) will provide an upper bound for the free boundary (3.56), more precisely

(3.80) if $x^*(t)$ is continuous on $[0, T[$ then $x^*(t)\leqq q(t)$ for every t in $[0, T[$.

Indeed, fix (x, t) in $\mathbb{R}\times]0, T[$ such that $x<x^*(t)$. By continuity, there is $\delta>0$ such that $x'<x^*(t')$ for every $|t'-t|<\delta$, $|x'-x|<\delta$. Since $\hat{w}(x', t')=0$, by definition of the free boundary, we get $A\hat{w}=0$ at (x', t'). This fact and (3.35) yield $g(x', t')\geqq 0$. Hence, as x' approaches $x^*(t')$ we deduce $g(x^*(t'), t')\leqq 0$ for every $|t'-t|<\delta$. Clearly, this implies (3.80). □

4. Finite resources. In this section we study the case of a monotone follower problem with a constraint on the resources (1.6).

Let A be the differential operator (2.1) and define

$$B'v=\frac{\partial v}{\partial z}-\frac{\partial v}{\partial x}-c(t) \tag{4.1}$$

for a function $v(x, z, t)$, (x, z) in $\mathbb{R}\times[0, \infty[$, $0\leqq t\leqq T$. A heuristic application of the dynamic programming to the problem (1.16), $\cdots$, (1.19) yields the following Hamilton-Jacobi-Bellman equation

$$\begin{aligned}&(Av-f)\vee B'v=0 \quad \text{in } \mathbb{R}\times]0, \infty[\times[0, T[,\\ &v(\cdot, \cdot, T)=0 \quad \text{in } \mathbb{R}\times[0, \infty[,\\ &Av=f \quad \text{in } \mathbb{R}\times\{0\}\times[0, T[\end{aligned} \tag{4.2}$$

to be satisfied by the optimal cost $\hat{v}$ given by (1.6).

First of all, we need some a priori estimates.

THEOREM 4.1. *Assume* (1.7), $\cdots$, (1.10) *hold. Then the optimal cost* $\hat{v}$ *defined by* (1.6) *is a nonnegative continuous function such that for some constants* $0 < c \leqq C$, *the same* $m \geqq 1$ *of the hypothesis* (1.9), *and every* (x, z, t), (x', z', t') *in* $\mathbb{R} \times [0, \infty[\times [0, T]$ *we have*

$$
\begin{aligned}
&c|x^+|^m - C \leqq \hat{v}(x, z, t) \leqq C(1+|x|^m),\\
&|\hat{v}(x, z, t) - \hat{v}(x', z, t)| \leqq C(1+|x|^{m-1}+|x'|^{m-1})|x-x'|,\\
&|\hat{v}(x, z, t) - \hat{v}(x, z, t') \leqq C(1+|x|^m)|t-t'|,\\
&0 \leqq \frac{\partial^2 \hat{v}}{\partial x^2}(x, z, t) \leqq C(1+|x|^q), \qquad q = (m-2)^+
\end{aligned}
\tag{4.3}
$$

and

$$
\hat{v}(x, z, t) - \hat{v}(x, z', t) \leqq C(1+|x|^{m-1})(z'-z)^+, \tag{4.4}
$$

so, $\hat{v}$ *is convex in the first variable and decreasing in the second variable.*[3]

Proof. The estimate (4.3) is obtained by an analogy to Theorem 2.1. Let us prove (4.4). Indeed, notice that

$$
\hat{v}(x, z, t) - \hat{v}(x, z', t) \leqq \sup\{(J_{xt}(\nu) - J_{xt}(\nu')) : \nu' \text{ in } \mathcal{V} \text{ satisfying (2.9) and } \nu'(T-t) \leqq z'\}, \tag{4.5}
$$

where ν is chosen as any measurable function of ν', with $\nu(T-t) \leqq z$. In particular, we take

$$
\nu(s) = \begin{cases} \nu'(s) & \text{if } \nu'(s) \leqq z, \\ z & \text{if } \nu'(s) \geqq z. \end{cases}
$$

Hence, using the fact that for $y(s)$, $y'(s)$ denoting the processes associated with ν, ν', respectively,

$$
E\{|y(s) - y'(s)|^m\} \leqq C|(z'-z)^+|^m \quad \text{for every } s \text{ in } [t, T],
$$

for some constant C independent of x, t, z, z', ν and ν', we deduce, by virtue of (1.8), (1.9) and Hölder's inequality,

$$
J_{xt}(\nu) - J_{xt}(\nu') \leqq CE\left\{\int_t^T (1+|y(s)|^m + |y'(s)|^m)\, ds\right\} |(z'-z)^+|^m, \tag{4.6}
$$

for another constant C. Finally, since (2.11) is equivalent to

$$
E\left\{\int_0^{T-t} |\nu(s)|^m\, ds\right\} \leqq C(1+|x|^m), \tag{4.7}
$$

for an appropriate constant C, the expressions (4.5) and (4.6) imply (4.4). □

Denote by V_m the function space,

v belongs to V_m if $v : \mathbb{R} \times [0, \infty[\times [0, T] \to \mathbb{R}$ is locally Lipschitz continuous such that

$$
\begin{aligned}
&|v(x, z, t)| + \left|\frac{\partial v}{\partial t}(x, z, t)\right| \leqq C(1+|x|^m),\\
&\left|\frac{\partial v}{\partial x}(x, z, t)\right| + \left|\frac{\partial v}{\partial z}(x, z, t)\right| \leqq C(1+|x|^{m-1})
\end{aligned}
\tag{4.8}
$$

for almost every (x, z, t) and some constant C.

[3] Note that $\hat{v}$ satisfies the dynamic programming equation.

Note the change of notation with respect to the definition (2.34) in § 2.

Observe that function u^0, given by (1.18), is also the unique solution of the equation

$$\begin{aligned} &Au^0 = f \quad \text{in } \mathcal{D}'(\mathbb{R}\times[0, T[), \\ &u^0(\cdot, T) = 0 \quad \text{in } \mathbb{R}, \end{aligned} \tag{4.9}$$

under the regularity (2.34).

Consider the problem:

Find v in V_m such that

$$\begin{aligned} &v(\cdot, \cdot, T) = 0 \quad \text{in } \mathbb{R}\times[0, \infty[, \\ &v(\cdot, 0, \cdot) = u^0 \quad \text{in } \mathbb{R}\times[0, T], \\ &Av \leqq f \quad \text{in } \mathcal{D}'(\mathbb{R}\times]0, \infty[\times[0, T[), \\ &B'v \leqq 0 \text{ a.e.} \quad \text{in } \mathbb{R}\times]0, \infty[\times]0, T[. \end{aligned} \tag{4.10}$$

Notice that for every v in V_m,

$$B'v \leqq 0 \text{ a.e.} \quad \text{in } \mathbb{R}\times]0, \infty[\times]0, T[\tag{4.11}$$

is equivalent to

$$v \leqq M'v \quad \text{in } \mathbb{R}\times]0, \infty[\times[0, T[, \tag{4.12}$$

where the operator

$$Mv = \inf\{\xi c(t) + v(x+\xi, z-\xi, t) : 0 \leqq \xi \leqq z\}. \tag{4.13}$$

THEOREM 4.2. *Under the assumptions* (1.7), $\cdots$, (1.10) *the problem* (4.10) *possesses a maximum solution* $\hat{v}$, *which is given explicitly as the optimal cost* (1.6). *Moreover, we have the following decomposition*:

$$\hat{v}(x, z, t) = \hat{u}(x, t) + h(x+z, t) \quad \textit{for every } (x, z, t) \textit{ in } \mathbb{R}\times[0, \infty[\times[0, T], \tag{4.14}$$

where $\hat{u}$ *is the unlimited optimal cost* (1.4) *and*

$$h = u^0 - \hat{u} \quad \textit{in } \mathbb{R}\times[0, T], \tag{4.15}$$

with u^0 *being defined by* (4.9).

Proof. First of all, we remark the dynamic programming equation applies to both optimal control problems (1.4), (1.6), i.e. if

$$z(s) = z - \nu(s-t) \quad \text{for every } t \leqq s \leqq T, \tag{4.16}$$

$$\begin{aligned} J_{xt}(\nu, \theta) = E\left\{\int_t^\theta f(y(s), s) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) ds + c(t)\nu(0) I(t<\theta)\right. \\ \left. + \int_t^\theta c(s) \exp\left(-\int_t^s \alpha(\lambda)\, d\lambda\right) d\nu(s-t)\right\} \end{aligned} \tag{4.17}$$

and

$$\tau = \inf\{s \in [t, T] : z(s) < 0\} \tag{4.18}$$

then

$$\hat{u}(x, t) = \inf\left\{J_{xt}(\nu, \theta) + E\left[\exp\left(-\int_t^\theta \alpha(s)\, ds\right) \hat{u}(y(\theta-), \theta)\right] : \mathcal{A}\right\}, \tag{4.19}$$

$$\hat{v}(x, z, t) = \inf \Big\{ J_{xt}(\nu, \theta \wedge \tau) + E\left[\exp\left(-\int_t^{\theta\wedge\tau} \alpha(s)\, ds\right) \hat{v}(y(\theta\wedge\tau-), z(\theta\wedge\tau-), \theta\wedge\tau)\right] : \mathcal{A} \Big\}, \tag{4.20}$$

where $t \leqq \theta \leqq T$ is any stopping time associated with the system control $\mathcal{A}$, which includes the probability space $(\Omega, \mathcal{T}, P)$, the filtration, the Wiener process and the control $(\mathcal{T}^t, w(t), \nu(t), t \geqq 0)$.

Next, by virtue of the estimates (4.3), (4.4) we can prove as in § 2 that the optimal cost (1.6) is the maximum solution of the problem (4.10).

Finally, let us prove (4.14). Indeed, using either (1.19) or (4.20) with $\theta = \tau$, and the fact that

$$\hat{v}(\cdot, 0, \cdot) = \hat{u} + h,$$

we obtain

$$\hat{v}(x, z, t) = \inf \Big\{ J_{xt}(\nu, \tau) + E\left[\exp\left(-\int_t^{\tau} \alpha(s)\, ds\right) \hat{u}(y(\tau-), \tau)\right] + E\left[\exp\left(-\int_t^{\tau} \alpha(s)\, ds\right) h(y(\tau-), \tau)\right] : \mathcal{A} \Big\}. \tag{4.21}$$

Since we may assume that $\nu(\cdot)$ is continuous and because of

$$y(\tau-) = x + z + \int_t^{\tau} (a(s)y(s) + b(s))\, ds + \int_t^{\tau} \sigma(s)\, dw(s)$$

and

$$Ah = f - A\hat{u} \geqq 0$$

we get, by applying Itô's formula

$$E\left\{\exp\left(-\int_t^{\tau} \alpha(s)\, ds\right) h(y(\tau-), \tau)\right\} \leqq h(x+z, t). \tag{4.22}$$

Clearly, combining (4.19), (4.21) and (4.22), we deduce

$$\hat{v}(x, t) \leqq \hat{u}(x, t) + h(x+z, t). \tag{4.23}$$

On the other hand, denoting by $v(x, z, t)$ the right-hand side of (4.23), we have

$$Av(x, z, t) = A\hat{u}(x, t) + f(x+z, t) - A\hat{u}(x+z, t). \tag{4.24}$$

Denoting by $x^*(t)$ the free boundary (3.56), the equality (4.24) yields

$$Av(x, z, t) \leqq A\hat{u}(x, t) \leqq f(x, t) \quad \text{if } x+z \geqq x^*(t).$$

Because

$$A\hat{u}(x+z, t) - A\hat{u}(x, t) = \int_0^z A' \frac{\partial \hat{u}}{\partial x}(x+\lambda, t)\, d\lambda,$$

where A' is the operator (3.1), so from (4.24) we obtain

$$Av(x, z, t) = f(x, t) - \int_0^z g(x+\lambda, t)\, d\lambda \leqq f(x, t) \quad \text{if } x+z < x^*(t),$$

in view of Remark 3.7 and the definition (3.3). Hence

$$A\nu \leqq f \quad \text{in } \mathbb{R}\times]0,\infty[\times[0,T[\tag{4.25}$$

and also

$$B'\nu = B\hat{u} \leqq 0 \quad \text{in } \mathbb{R}\times]0,\infty[\times[0,T[. \tag{4.26}$$

This implies that ν solves the problem (4.10) and since $\hat{v}$ is the maximum solution, the equality must hold in (4.23). □

COROLLARY 4.1. *If the conditions* (1.7), $\cdots$, (1.10) *and* (3.78) *hold, then the control* $\hat{v}\wedge z$ *is optimal for the problem with the resource constraints* (1.6), *where* $\hat{v}$ *is the process defined in Theorem* 3.5.

Proof. The result is straightforward and follows from the decomposition (4.14), the technique of Theorem 3.5 and Remark 3.6. □

Remark 4.1. An equivalence to Theorem 3.4 can be stated for the problem with the resource constraints (1.6). Moreover, the fact that $f(t,x)$ approaches infinity as x goes to positive infinity is useless in the proof for existence of an optimal control relative to problem (1.6). □

Remark 4.2. From the expressions (1.4) and (1.6), it follows that

$$\hat{v}(x,z,t)\to\hat{u}(x,t) \quad \text{as } z\to+\infty \tag{4.27}$$

in a decreasing fashion and pointwise in $\mathbb{R}\times[0,T]$. Hence, the equalities (4.14) and (4.15) imply, for every t in $[0,T]$,

$$u^0(x,t)-\hat{u}(x,t)\to 0 \quad \text{as } x\to+\infty \tag{4.28}$$

in a decreasing fashion. This means that for a large initial state x, the optimal cost (1.4) is very close to the cost of the free-control evolution. Clearly, this agrees with the characteristics of the optimal control of Theorem 3.5. □

5. Optimal corrections. Now, we consider a model of an optimal correction control problem which will be reduced to a problem of the type presented in § 1.

Denote by $\mathcal{V}$ the set of controls $\nu(\cdot)$ which are progressively measurable random processes from $[0,+\infty]$ into $\mathbb{R}$, right continuous having left limit (cad-lag) and with locally bounded variation. Hence if $\mathcal{V}_+$ is the set of processes in $\mathcal{V}$ which are nonnegative and increasing, we have the following decomposition

$$\mathcal{V} = \mathcal{V}_+ \ominus \mathcal{V}_+, \tag{5.1}$$

i.e., for every $\nu(\cdot)$ in $\mathcal{V}$ there exist $\nu_1(\cdot)$, $\nu_2(\cdot)$ in $\mathcal{V}_+$ such that

$$\begin{aligned} &\nu(t)=\nu_1(t)-\nu_2(t), \qquad t\geqq 0,\\ &\nu_1(0)=(\nu(0))^+, \qquad \nu_2(0)=(\nu(0))^-. \end{aligned} \tag{5.2}$$

Note the change of notations used in § 1.

The state of the dynamic system is described by (1.2), i.e.,

$$\begin{aligned} y(s) = x+\nu(s-t)+\int_t^s (a(\lambda)y(\lambda)+b(\lambda))\,d\lambda\\ +\int_t^s \sigma(\lambda)\,dw(\lambda-t), \qquad s\geqq t, \end{aligned} \tag{5.3}$$

$y(s)=y_{xt}(s,\nu)$ being a cad-lag random process adapted to $(\mathcal{F}^{s-t}, s\geqq t)$. A cost

associated to each control ν in $\mathcal{V}$ is given by the payoff functional (1.3), i.e.

$$J_{xt}(\nu) = E\left\{\int_t^T f(y(s), s)\exp\left(-\int_t^s \alpha(\lambda)\,d\lambda\right)ds + c(t)|\nu(0)| + \int_t^T c(s)\exp\left(-\int_t^s \alpha(\lambda)\,d\lambda\right)d|\nu|(s-t)\right\}, \tag{5.4}$$

where $a(t)$, $b(t)$, $\sigma(t)$, $c(t)$, $\alpha(t)$, $f(x, t)$ and T satisfy (1.7), (1.8), (1.9), (1.10), and $|\nu|$ denotes the total variation of ν, i.e., $|\nu| = \nu_1 + \nu_2$ given by (5.2). Notice that a better notation could be $J_{xt}(\nu_1, \nu_2)$ in lieu of $J_{xt}(\nu)$, because ν_1, ν_2 are not uniquely determined by ν. However, we prefer to use (5.4).

Our purpose is to characterize the optimal cost

$$\hat{u}(x, t) = \inf\{J_{xt}(\nu): \nu \text{ in } \mathcal{V}\} \tag{5.5}$$

and to construct an optimal control $\hat{\nu}$ in $\mathcal{V}$.

In the first part of this section we treat the problem just stated and then offer some general comments about other extensions of these results.

5.1. Reduction. Let us suppose that $f(x, t)$ is symmetric in the following sense.

$$f(x, t) = f(2x_0(t) - x, t),\ (x, t) \text{ in } \mathbb{R}\times[0, T] \text{ with } x_0(t) \text{ being Lipschitz continuous in } [0, T] \text{ and satisfying } \dot{x}_0(t) = a(t)x_0(t) + b(t),\ t \text{ in } [0, T], \tag{5.6}$$

where $\dot{x}_0(t)$ denotes the derivative of $x_0(t)$. From (5.6) we have

$$\frac{\partial f}{\partial x} = 0 \quad \text{at } (x_0(t), t) \text{ for every } t \text{ in } [0, T]. \tag{5.7}$$

Therefore, the function $f(x, t)$ is completely determined by the restriction of $f(x, t)$ on the half-line $x \geqq x_0(t)$ for every t in $[0, T]$. The assumptions (1.9) and (5.6) imply

$$c|x|^m - C \leqq f(x, t) \leqq C(1 + |x|^m) \quad \text{in } \mathbb{R}\times[0, T], \tag{5.8}$$

for some constants $C \geqq c > 0$, $m \geqq 1$. Observe that $x_0(t)$ represents the minimal trajectory of the system.

THEOREM 5.1. *Let the assumptions* (1.7), $\cdots$, (1.10) *and* (5.6) *hold. Then, if* $\hat{u}(x, t)$ *denotes the optimal cost* (5.5), *we have*

$$\hat{u}(x, t) = \hat{u}(2x_0(t) - x, t) \quad \textit{for every } (x, t) \quad \textit{in } \mathbb{R}\times[0, T], \tag{5.9}$$

where $x_0(t)$ *is given in* (5.6).

Proof. Let ν be an arbitrary control in $\mathcal{V}$ and (x, t) be any point in $\mathbb{R}\times[0, T]$. From (5.3) we have for $t \leqq s \leqq T$

$$y_{xt}(s, \nu) = 2x_0(s) + y_{xt}(s, \nu - 2q) \quad \text{with } q(s) = x_0(s) - \int_t^s a(s)x_0(s)\,ds.$$

Since

$$q(s) = \int_t^s b(s)\,ds + x_0(t),$$

we have

$$y_{xt}(s, \nu) = 2x_0(s) - \check{y}_z(s, -\nu), \qquad z = 2x_0(t) - x, \tag{5.10}$$

where $\check{y}(s)$ solves an equation similar to (5.3) with a new Wiener process $\check{w}(s-t)=-w(s-t)$ in lieu of $w(s-t)$. Hence

$$\check{y}_{zt}(s,-\nu)=y_{zt}(s,-\nu) \quad \text{in law.}$$

Thereby, we obtain by virtue of (5.6)

$$J_{xt}(\nu)=J_{zt}(-\nu), \tag{5.11}$$

where z is given by (5.10).

Thus, the assertion (5.9) is deduced from (5.11) by taking the infimum over ν in $\mathcal{V}$ □

Remark 5.1. As in Theorem 2.1, we can prove that under the hypotheses (1.7), $\cdots$, (1.10) and (5.6), there exist constants $C \geqq c > 0$, such that for the same $m \geqq 1$ of the assumption (1.9) and every (x, t), (x', t') in $\mathbb{R}\times[0, T]$ we have

$$\begin{aligned}
&0 \leqq \hat{u}(x,t) \leqq C(1+|x|^m),\\
&|\hat{u}(x,t)-\hat{u}(x',t)| \leqq C(1+|x|^{m-1}+|x'|^{m-1})|x-x'|,\\
&|\hat{u}(x,t)-\hat{u}(x,t')| \leqq C(1+|x|^m)|t-t'|,\\
&0 \leqq \frac{\partial^2 u}{\partial x^2}(x,t) \leqq C(1+|x|^q), \qquad q=(m-2)^+,
\end{aligned} \tag{5.12}$$

so $\hat{u}$ is convex in the first variable. Actually, $m=1$ in (5.12) even if $m>1$ in the assumption (5.8).

Remark 5.2. From Theorem 5.1 we deduce that

$$\frac{\partial \hat{u}}{\partial x}=0 \quad \text{at } (x_0(t), t) \text{ for every } t \text{ in } [0, T], \tag{5.13}$$

which represents a Neumann boundary condition for the corresponding Hamilton-Jacobi-Bellman equation, i.e. the optimal cost $\hat{u}$ is the solution of the equation

$$\begin{aligned}
&(A\hat{u}-f) \vee B\hat{u}=0 \quad \text{if } x \leqq x_0(t),\ 0 \leqq t \leqq T,\\
&\hat{u}(\cdot, T)=0 \quad \text{in }]-\infty, x_0(T)],
\end{aligned} \tag{5.14}$$

with the boundary condition (5.13). This implies that the restriction of the optimal cost $\hat{u}(x, t)$ to the half-line $x \leqq x_0(t)$, $0 \leqq t \leqq T$, is actually the solution of a quasi-variational inequality with Neumann boundary condition, associated with an optimal impulse control problem where the state of the system is a reflected diffusion process (cf. Bensoussan and Lions [9], and [37], [52]). On the other hand, notice that $\hat{u}=f(x_0)$ if $c=0$ and f is time-independent.

The whole § 3 can be adapted to this case. For instance, define the differential operator

$$A'u=-\frac{\partial u}{\partial t}-\sigma^2(t)\frac{\partial^2 u}{\partial x^2}-(a(t)x+b(t)+\dot{x}_0(t))\frac{\partial u}{\partial x}+(\alpha(t)-a(t))u \tag{5.15}$$

and the substitutions

$$\hat{w}(x,t)=-\frac{\partial \hat{u}}{\partial x}(x-x_0(t), t)-c(t), \tag{5.16}$$

$$g(x,t)=\frac{dc}{dt}(t)-(\alpha(t)-a(t))c(t)-\frac{\partial f}{\partial x}(x-x_0(t), t), \tag{5.17}$$

for the given functions $\hat{u}$ and f.

Then, the following equation is satisfied by the optimal cost (5.5) through (5.16) and (5.17),

$$
\begin{aligned}
&(A'\hat{w} - g) \vee \hat{w} = 0 \quad \text{in }]-\infty, 0] \times [0, T[, \\
&\hat{w}(\cdot, T) = 0 \quad \text{in }]-\infty, 0], \\
&\hat{w}(0, \cdot) = 0 \quad \text{in } [0, T].
\end{aligned} \tag{5.18}
$$

Moreover, the solution $\hat{w}$ of (5.18) admits a stochastic representation as the optimal cost of a stopping time problem, i.e.,

$$
\hat{w}(x, t) = \inf \{S_{xt}(\theta) : t \leqq \theta \leqq T, \text{ stopping time}\}, \tag{5.19}
$$

where

$$
\begin{aligned}
S_{xt}(\theta) &= E\left\{\int_t^{\theta \wedge \tau} g(y^0(s), s) \exp\left(-\int_t^s (\alpha(\lambda) - a(\lambda))\, d\lambda\right) ds\right\}, \\
\tau &= \inf \{s \geqq t : y^0(s) \geqq 0\},
\end{aligned} \tag{5.20}
$$

and the process $y^0(s) = y^0_{xt}(s)$ is given by (5.3) with the control $\nu = 0$.

Next, with the function $\hat{w}(x, t)$ we can define the moving boundary $x^*(t)$, $0 \leqq t < T$, by

$$
x^*(t) = \inf \{x \leqq 0 : \hat{w}(x, t) < 0\} \tag{5.21}
$$

which induces an optimal control.

The precise variational inequality is exactly (3.23) with the space

(5.22) V is the set of all real measurable functions v on $[0, \infty[$ with a derivative v' such that $\|v\|_p$ and $\|v'\|_{p-1}$ are finite, and $v(0) = 0$,

where $\|\cdot\|_p$ and $(\cdot, \cdot)$ are the norm and the inner product on $[0, \infty[$ instead of $\mathbb{R}$. The bilinear form $a(t, u, v)$ is defined as in (3.21) but the integration is over $[0, \infty[$ in lieu of $\mathbb{R}$, where a term is added in order to use the new definition (5.15) of the operator A'. In a similar way, if the space V_{m-1} is given by (3.34) restricted to $[0, \infty[$, we can state a strong formulation of the variational inequality as follows:

Find w in V_{m-1} such that

$$
\begin{aligned}
&w(x, T) = w(0, t) = 0 \quad \text{for every } (x, t) \text{ in }]-\infty, 0] \times [0, T], \\
&A'w \leqq g \quad \text{in } \mathcal{D}'(]-\infty, 0[\times]0, T[), \qquad w \leqq 0 \quad \text{in }]-\infty, 0] \times [0, T[.
\end{aligned} \tag{5.23}
$$

As in Theorems 3.2 and 3.3 we can prove

THEOREM 5.2. *Under the hypotheses* (1.7), $\cdots$, (1.10) *the function* (5.19) *is the maximum solution of the weak variational inequality* (3.23) *with the changes* (5.22). *Moreover, if we also suppose* (3.36) *is true and*

(5.24) *the derivative of* $x_0(t)$ *is Lipschitz continuous in* $[0, T]$,

then the strong version (5.23) *of the variational inequality admits a maximum solution, which is precisely the optimal cost* (5.19) *and the equality*

$$
A'\hat{w} = g \quad \text{in } \mathcal{D}([\hat{w} < 0]) \tag{5.25}
$$

holds.

Remark 5.3. Similar results to Theorems 3.4 and 3.5 can be proved. For instance, assuming (1.7), $\cdots$, (1.10) and (3.78), there exists an optimal control $\hat{v}$ in $\mathcal{V}$ which is

continuous and uniquely determined by (5.3) and the conditions

$$\hat{\nu}(s) = \hat{\nu}_+(s) - \hat{\nu}_-(s) \quad \text{with } \hat{\nu}_+, \hat{\nu}_- \text{ in } \mathcal{V}_+, \tag{5.26}$$

if $z_+(s) = x_0(s) + x^*(s)$ and $z_-(s) = x_0(s) - x^*(s)$, then we impose

$$\begin{aligned} &\hat{\nu}_+(0) = (z_+(t) - x)^+, \qquad \hat{\nu}_-(0) = (z_-(t) - x)^-, \\ &z_+(s) \leqq y(s) \leqq z_-(s) \quad \text{for every} \quad t \leqq s \leqq T, \end{aligned} \tag{5.27}$$

$$\begin{aligned} &\int_t^T I(y(s) > z_+(s))\, d\hat{\nu}_+(s-t) = 0, \\ &\int_t^T I(y(s) < z_-(s))\, d\hat{\nu}_-(s-t) = 0, \end{aligned} \tag{5.28}$$

where $I(\cdot)$ denotes the characteristic function, $y(s)$ the associated state and $x_0(t)$, $x^*(t)$ are given by (5.6), (5.21) respectively, i.e. $\hat{\nu}$ reproduces the reflected diffusion of $y(s)$ on the interval $[z_+, z_-]$.

5.2. General comments. Most of the results presented herein can be extended to more general situations. Let us mention the following examples:

Extension to multidimensional model. This includes all of § 2 about the dynamic programming equation, the second part of § 3, i.e. § 3.2, about the optimal decision process, all of § 4 about the case of finite resources, the first part of this section, i.e., § 5.1, about the optimal correction problem. Let us mention that one of the main difficulties of the multidimensional case is the smoothness of the free boundary, which is for us an open question.

Extension to partially observed system. Since the model-equation is linear and the system may be degenerate, we can treat a multidimensional model with incomplete information on the state of the system. In particular, a separation principle result can be obtained (cf. [44]).

Extension to nonconvex data. In all of §§ 2, 4 and in the first part of this section, i.e., § 5.2, we may allow the coefficients of the stochastic equation (1.2) to be nonlinear in x, i.e., $\sigma = \sigma(x, t)$, $g = g(x, t)$ in lieu of $ax + b$, and also $\alpha = \alpha(x, t)$, $c = c(x, t)$ and $f = f(x, t)$ to not necessarily be convex in x. In that case, the optimal cost $\hat{u}(x, t)$ is no longer convex in x and the technique of [41] applies.

Extension to diffusion with jumps. All results herein may be extended to a model in which a Poisson integral is added to the stochastic equation (1.2). The technique is similar to that used in [42].

Extension to long term average criterion. When the horizon is infinite, we may consider a model with a long term average cost instead of the cost (2). (See, e.g. [43].)

Nonsymmetric case. It is possible to treat cases in which the reduction (5.9) does not hold. This is the case, for instance, if $f(x, t)$ is not symmetric or the cost $J_{xt}(\nu)$ involves $c_1(\cdot)\nu_1(\cdot)$ and $c_2(\cdot)\nu_2(\cdot)$ with $\nu = \nu_1 - \nu_2$.

To conclude, let us mention that decomposable models and problems with the long run average criterion may be treated. Also, a combined version of §§ 4 and 5 can be developed.

6. Examples. To illustrate the results obtained in the previous sections, we shall consider some examples. We assume that the coefficients a, b, α, σ in (1.2) and (1.3) are constant, and the running cost $f(x)$ is time-independent and satisfies the condition (1.9). In addition, let $c(t) \equiv 0$, i.e., the cost for control is negligible. As mentioned in the introduction, for $a < 0$ and $b > 0$, the equation (1.2) may be interpreted as an

automatic cruise control problem. Probabilistically it pertains to the control of the motion of a Brownian particle with viscous damping, or an Ornstein-Uhlenbeck process [56]. In the case that $a>0$ and $b<0$, it becomes a simple model for the control of the population of a renewable resource. In either case, the unperturbed equilibrium state is $x_0=(-b/a)>0$. We wish to construct the optimal control, in particular, to find the free boundary, so that the mean-square deviation from the equilibrium value x_0 is minimum.

6.1. Unlimited resources. Under the above assumptions, the average cost (1.3) yields

$$J_{xt}(\nu)=E\left\{\int_t^T f(y(s))\,e^{-\alpha s}\,ds\right\}. \tag{6.1}$$

By Theorem 2.5, the optimal cost $\hat{u}$ (1.4) must satisfy

$$\begin{aligned} &A_0\hat{u}=f \quad \text{and} \quad \frac{\partial \hat{u}}{\partial x}\geqq 0 \quad \text{if } x\geqq x^*(t),\\ &A_0\hat{u}\leqq f \quad \text{and} \quad \frac{\partial \hat{u}}{\partial x}=0 \quad \text{if } x\leqq x^*(t),\ 0\leqq t\leqq T, \end{aligned} \tag{6.2}$$

where

$$A_0u=-\frac{\partial u}{\partial t}-\frac{1}{2}\sigma^2\frac{\partial^2 u}{\partial x^2}-(ax+b)\frac{\partial u}{\partial x}+\alpha u, \tag{6.3}$$

$$x^*(t)=\inf\left\{x:\frac{\partial \hat{u}}{\partial x}(x,t)>0\right\}. \tag{6.4}$$

To construct the solution $\hat{u}$ for $x\geqq x^*(t)$, we let $s=(T-t)$ so that (5.2) gives the following free-boundary problem

$$\begin{aligned} &v(x,s)\equiv \hat{u}(x,T-s),\\ &Lv=\frac{\partial v}{\partial s}-\frac{1}{2}\sigma^2\frac{\partial^2 v}{\partial x^2}-(ax+b)\frac{\partial v}{\partial x}+\alpha v=f(x),\\ &\frac{\partial v}{\partial x}\geqq 0, \quad \text{for } x>x^*(T-s),\ 0\leqq s\leqq T,\\ &v(x,0)=0,\\ &\left.\frac{\partial v}{\partial x}\right|_{x=x^*(T-s)}=0, \end{aligned} \tag{6.5}$$

where $v(x,s)=\hat{u}(x,T-s)$.

Introduce the following change of variables:

$$\begin{aligned} &\tau=\frac{e^{2as}-1}{2a}, \qquad 0\leqq s\leqq T,\\ &\xi=\frac{1}{\sigma}(x-x_0)e^{as}, \qquad x_0=-\frac{b}{a},\\ &\omega=ve^{\alpha s},\\ &\xi^*(\tau)=\frac{1}{\sigma}(1+2a\tau)^{1/2}\left\{x^*\left[T-\frac{1}{2a}\ln(1+2a\tau)\right]-x_0\right\}. \end{aligned} \tag{6.6}$$

In terms of the above variables, it is easy to check that (6.5) reduces to a standard free-boundary problem for a heat equation.

$$Mw = \frac{\partial w}{\partial \tau} - \frac{1}{2}\frac{\partial^2 w}{\partial \xi^2} = g(\xi, \tau),$$

$$\begin{aligned} &\frac{\partial w}{\partial \xi} \geqq 0, \quad \text{for } \xi \geqq \xi^*(\tau), \quad 0 \leqq \tau \leqq \tau_1 = (e^{2aT} - 1)/2a, \\ &w(\xi, 0) = 0, \\ &\left.\frac{\partial w}{\partial \xi}\right|_{\xi = \xi^*(\tau)} = 0, \end{aligned} \tag{6.7}$$

where

$$\begin{aligned} &g(\xi, \tau) = (1 + 2a\tau)^{\beta} \cdot f\{\sigma\xi(1 + 2a\tau)^{-1/2} + x_0\}, \\ &\beta = \frac{\alpha - 2a}{2a}. \end{aligned} \tag{6.8}$$

To solve (6.7) we seek a similarity solution of the form

$$\begin{aligned} &w(\xi, \tau) = [\theta(\tau)]^n \varphi(\eta) \quad \text{for some } n \in \mathbb{R}^+, \\ &\eta = \frac{\xi}{\theta(\tau)}, \qquad \tau > 0. \end{aligned} \tag{6.9}$$

By a straightforward computation, we get

$$Mw = \theta^{n-1}\dot{\theta}(n\varphi - \eta\varphi') - \theta^{n-2}\varphi'' = g(\xi^*\eta, \tau) \tag{6.10}$$

or

$$\theta\dot{\theta}(n\varphi - \eta\varphi') - \varphi'' = g(\xi^*\eta, \tau)/\theta^{n-2}. \tag{6.11}$$

Now, suppose that f is symmetric about x_0 such that

$$f(x + x_0) = h(x) = |r|^m h(rx) \quad \text{for every } r \in \mathbb{R} - \{0\}. \tag{6.12}$$

That is, h is positive and homogeneous of degree m. Then the system (6.7) is reducible to a one-dimensional problem, if we choose

$$\theta\dot{\theta} = \tfrac{1}{2}, \qquad \theta(0) = 0, \tag{6.13}$$

so that the free boundary is given by

$$\xi^*(\tau) = \delta\theta(\tau) = \delta\left(\frac{\tau}{2}\right)^{1/2} \quad \text{for some } \beta \in \mathbb{R}, \quad 0 \leqq \tau < \tau_1. \tag{6.14}$$

In view of (6.8), (6.12)-(6.14), the equation becomes an ordinary differential equation

$$-\tfrac{1}{2}\varphi'' + \tfrac{1}{2}(n\varphi - \eta\varphi') = \sigma^m h(\eta), \tag{6.15}$$

provided that

$$n = m + 2, \qquad \beta = \frac{m}{2}. \tag{6.16}$$

Let us summarize the above results:

THEOREM 6.1. *In* (1.2) *and* (1.3), *we assume the following*:

(6.17) *a, b, α, σ are constant and $c(t) \equiv 0$, the conditions* (6.12) *and* (6.16) *are satisfied.*

Then, under the transformations (6.6) *and* (6.9), *the free boundary problem* (6.5) *is reducible to*

$$
\begin{aligned}
&-\tfrac{1}{2}\varphi'' + \tfrac{1}{2}(n\varphi - \eta\varphi') = \sigma^m h(\eta) \quad \text{for } \eta \geqq \delta, \\
&\varphi'(\delta) = 0, \\
&\varphi''(\delta) = 0, \\
&\varphi(\eta) = O(\eta^m) \quad \text{as } \eta \to \infty.
\end{aligned} \tag{6.18}
$$

Remark 6.1. The last two conditions in (6.18) follow from Theorem 2.1. The reduced problem (6.18) is a free boundary value problem in one dimension where δ is to be determined in the process of constructing the solution. A special case, to be considered in what follows, has been solved by Benes, Shepp and Witsenhausen [6].

As a special case, let $m = 2$. By (6.16), we get

$$\beta = 1, \qquad n = 4. \tag{6.19}$$

Then, setting $\sigma = 1$, (6.18) may be written as

$$
\begin{aligned}
&-\tfrac{1}{2}\varphi'' + \tfrac{1}{2}(4\varphi - \eta\varphi') = \eta^2, \qquad \eta \geqq \delta, \\
&\varphi(\delta) = \tfrac{1}{2}\delta^2, \\
&\varphi'(\delta) = 0, \\
&\varphi(\eta) = O(\eta^2) \quad \text{as } \eta \to \infty.
\end{aligned} \tag{6.20}
$$

Similar to [6, Problem 2] (with η replaced by $-x$), the solution of (6.20) is given by

$$\varphi(\eta) = \varphi_0(\eta) + b(\delta)\varphi_1(\eta) \int_\eta^\infty [\varphi_1(\lambda)]^{-2} e^{-\lambda^2/2}\, d\lambda \tag{6.21}$$

where

$$
\begin{aligned}
&\varphi_0(\eta) = (\eta^2 + \tfrac{1}{2}), \\
&\varphi_1(\eta) = (\eta^4 + 6\eta^2 + 3), \\
&b(\delta) = \varphi_0'(\delta) \Big/ \left\{ [\varphi_1(\delta)]^{-1} e^{-\delta^2/2} - \varphi_1'(\delta) \int_\delta^\infty [\varphi_1(\lambda)]^{-2} e^{-\lambda^2/2}\, d\lambda \right\}.
\end{aligned} \tag{6.22}
$$

The parameter δ is determined by the equation

$$\delta^2 + 1 = \frac{4\delta\varphi_1(\delta) \int_\eta^\infty [\varphi_1(\lambda)]^{-2} e^{-(\lambda^2 - \delta^2)/2}\, d\lambda}{\varphi_1(\delta)\varphi_1'(\delta) \int_\eta^\infty [\varphi_1(\lambda)]^{-2} e^{-\lambda^2/2}\, d\lambda - 1} \tag{6.23}$$

which may be solved numerically to yield $\delta = -0.6388 \cdots$. In view of (6.5), (6.6), (6.9) and (6.22), the problem (6.2) is solved and the associated free boundary is given by

$$x = \frac{\delta}{2}\left[\frac{1 - e^{-2a(T-t)}}{a}\right]^{1/2} - \frac{b}{a}, \qquad 0 \leqq t \leqq T. \tag{6.24}$$

6.2. Finite resources. In the previous case 6.1, suppose the resource ν for control is finite so that $0 \leqq \nu(T) \leqq z$. The optimal cost $\hat{v}(x, z, t)$ defined by (1.19) can be

decomposed, according to Theorem 4.2, into two simple problems. That is, noting (4.13) and (4.14),

$$\hat{v}(x, z, t) = u^0(x+z, t) - [\hat{u}(x+z, t) - \hat{u}(x, t)] \tag{6.25}$$

where $\hat{u}(x, t)$ is the optimal cost without resource constraint, while $u^0(x, t)$ is the cost of free evolution defined by (1.18). Therefore it must satisfy

$$\begin{aligned} &A_0 u^0 = f, \quad 0 \leqq t < T, \quad x \in \mathbb{R}, \\ &u^0(x, T) = 0, \\ &u^0(x, t) = O(|x|^m) \quad \text{as } |x| \to \infty, \end{aligned} \tag{6.26}$$

where A_0 is defined by (6.3). By the transformation (6.6), (6.26) may be solved to give

$$\begin{aligned} u^0(x, t) = e^{-\alpha(T-t)} \int_0^{(T-t)} \int_{\mathbb{R}} & \frac{\exp\left([\xi(x, t) - \rho]^2 / 2[\tau(t) - \lambda] + 2a\beta\lambda\right)}{2\pi[\tau(t) - \lambda]} \\ & \times (1 + 2a\lambda)^{\beta} f\left[\sigma(1+2a\lambda)^{-1/2}\rho - \frac{b}{a}\right] d\lambda \, d\rho, \\ \xi(x, t) = \frac{1}{\sigma}\left(x + \frac{b}{a}\right) e^{a(T-t)}, & \\ \tau(t) = (2a)^{-1}[e^{2a(T-t)} - 1]. & \end{aligned} \tag{6.27}$$

Thus, as a consequence of Theorems 4.2 and 6.1, we have

COROLLARY 6.1. *If, in addition to the hypotheses* (6.17), *we assume* $\nu \leqq z$, *then, in view of* (6.27), *the solution of* (6.26) *is reducible to a one-dimensional problem* (6.18).

Remark 6.2. Note that the free boundary, given by (6.14), remains unchanged. In particular, for $m = 2$, this problem may be solved explicitly.

We wish to point out that, for the optimal correction problems, the case of vanishing cost, $c = 0$, is less interesting. In this case the optimal policy would be to counteract the noise as long as the resources remain available so that $f(y(t), t)$ is kept to the minimum. However, for $c \neq 0$, the method of similarity transformations (6.6) and (6.9) is no longer applicable. This, of course, is true also for the one-sided control problems. Consequently one must deal with the genuine free-boundary problems for which the analytical solutions are difficult to obtain.

REFERENCES

[1] E. N. BARRON AND R. JENSEN, *Optimal control problems with no turning back*, J. Differential Equations, 36(1980), pp. 223-248.

[2] J. A. BATHER, *A continuous time inventory model*, J. Appl. Prob., 3(1966), pp. 538-549.

[3] ———, *A diffusion model for the control of a dam*, J. Appl. Prob., 5(1968), pp. 55-71.

[4] J. A. BATHER AND H. CHERNOFF, *Sequential decisions in the control of a spaceship*, in Proc. Fifth Berkeley Symposium of Mathematical Statistics and Probability, Berkeley, Univ. California Press, 1967, Vol. 3, pp. 181-207.

[5] ———, *Sequential decisions in the control of a spaceship (finite fuel)*, J. Appl. Prob., 4(1967), pp. 584-604.

[6] V. E. BENES, L. A. SHEPP AND H. S. WITSENHAUSEN, *Some solvable stochastic control problems*, Stochastics, 4(1980), pp. 39-83.

[7] A. BENSOUSSAN, *Inéquations quasi-variationnelles avec données non bornées et interpretation probabiliste*, C.R. Acad. Sc. Paris, Serie. I, 292(1981), pp. 751-754.

[8] A. BENSOUSSAN AND J. L. LIONS, *Applications des inéquations variationnelles en contrôle stochastique*, Dunod, Paris, 1978.

[9] ———, *Contrôle impulsionnel et inéquations quasi-variationnelles*, Dunod, Paris, 1982.

[10] A. BENSOUSSAN AND M. ROBIN, *On the convergence of the discrete time dynamic programming equation for general semi-groups*, this Journal, 20(1982), pp. 722-746.

[11] J. M. BONY, *Principe du maximum, inégalite de Harnack et unicité du problème de Cauchy pour les opérateurs elliptiques dégénérés*, Ann. Inst. Fourier Grenoble, 19(1967), pp. 277-304, see also C.R. Acad. Sc. Paris, 265(1967), pp. 333-336.

[12] M. I. BORODOWSKI, A. S. BRATUS AND F. L. CHERNOUSKO, *Optimal impulse correction under random perturbation*, Appl. Math. Mech. (PMM), 39(1975), pp. 797-805.

[13] A. S. BRATUS, *Solution of certain optimal correction problems with error of execution of the control action*, Appl. Math. Mech. (PMM), 38(1974), pp. 433-440.

[14] I. CAPUZZO-DOLCETTA AND M. MATZEU, *On the dynamic programming inequalities associated with the deterministic optimal stopping problem in discrete and continuous time*, Numerical Funct. Anal. Optim., 3(1981), pp. 425-450.

[15] H. CHERNOFF, *Optimal stochastic control*, Sankhyā, Ser. A, 30(1968), pp. 221-252.

[16] H. CHERNOFF AND A. J. PETKAU, *Optimal control of a Brownian motion*, SIAM J. Appl. Math., 34(1978), pp. 717-731.

[17] F. L. CHERNOUSKO, *Optimum correction under active disturbances*, Appl. Math. Mech. (PMM) 32(1968), pp. 203-208.

[18] ———, *Self-similar solutions of the Bellman equation for optimal correction of random disturbances*, Appl. Math. Mech. (PMM), 35(1971), pp. 333-342.

[19] M. J. FADDY, *Optimal control of finite dams: continuous output procedure*, Adv. Appl. Prob., 6(1974), pp. 689-710.

[20] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[21] A. FRIEDMAN, *Variational Principles and Free-Boundary Problems*, John Wiley, New York, 1982.

[22] V. K. GORBUNOV, *Minimax impulsive correction of perturbations of a linear damped oscillator*, Appl. Math. Mech. (PPM), 40(1976), pp. 252-259.

[23] J. M. HARRISON, T. M. SELLKE AND A. J. TAYLOR, *Impulse control of Brownian motion*, Math. Oper. Res., 8(1983), pp. 454-466.

[24] J. M. HARRISON AND M. I. TAKSAR, *Instantaneous control of a Brownian motion*, Math. Oper. Res., 8(1983), pp. 439-453.

[25] J. H. HARRISON AND A. J. TAYLOR, *Optimal control of a Brownian storage system*, Stoch. Proc. Appl., 6(1978), pp. 179-194.

[26] S. D. JACKA, *A finite fuel stochastic control problem*, Stochastics, 10(1983), pp. 103-113.

[27] I. KARATZAS, *The monotone follower problem in stochastic decision theory*, Appl. Math. Optim., 7(1981), pp. 175-189.

[28] ———, *A class of singular stochastic control problems*, Adv. Appl. Prob., 15(1983), pp. 225-254.

[29] I. KARATZAS AND S. E. SHREVE, *Connection between optimal stopping and singular stochastic control* I. *Monotone follower problems*, this Journal, 22(1984), pp. 856-877.

[30] ———, *Connection between optimal stopping and singular stochastic control* II. *Reflected follower problems*, this Journal, 23(1985), pp. 433-451.

[31] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[32] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, New York, 1980.

[33] H. KUSHNER, *Probability Methods for Approximations in Stochastic Control, and for Elliptic Equations*, Academic Press, New York, 1977.

[34] H. LEWY AND G. STAMPACCHIA, *On the smoothness of superharmonics which solve a minimum problem*, J. Anal. Math., 23(1970), pp. 227-236.

[35] P. L. LIONS AND J. L. MENALDI, *Optimal control of stochastic integrals and Hamilton-Jacobi-Bellman equations*, I, this Journal. 20(1982), pp. 58-81, pp. 82-95, see also C.R. Acad. Sc. Paris, Serie A, 287(1978), pp. 409-412.

[36] D. LUDWIG, *Optimal harvesting of randomly fluctuating resource* I, *application of perturbation methods*, SIAM J. Appl. Math., 37(1979), pp. 166-184.

[37] J. L. MENALDI, *On the optimal stopping time problem for degenerate diffusions*, this Journal., 18(1980), pp. 697-721, see also C.R. Acad. Sc. Paris, Serie A, 284(1977), pp. 1443-1446.

[38] ———, *On the optimal impulse control problem for degenerate diffusions*, this Journal 18(1980), pp. 722-739, see also C.R. Acad. Sc. Paris, Serie A, 284(1977), pp. 1499-1502.

[39] ———, *Stochastic variational inequality for reflected diffusion*, Indiana Univ. Math. J., 32(1983), pp. 733-744.

[40] J. L. MENALDI, J. P. QUADRAT AND E. ROFMAN, *On the role of the impulse fixed cost in stochastic optimal control: an application to the management of energy production*, in Proc. Tenth IFIP Conference on System Modelling and Optimization, Lecture Notes in Control and Information Sciences 38, Springer-Verlag, New York, 1982, pp. 671-679.

[41] J. L. MENALDI AND M. ROBIN, *On some cheap control problems for diffusion processes*, Trans. Am. Math. Soc., 278(1983), pp. 771-802, see also C.R. Acad. Sc. Paris, Serie I, 294(1982), pp. 541-544.

[42] ———, *On singular stochastic control problems for diffusions with jumps*, IEEE Trans. Automat. Control, AC-29(1984), pp. 991-1004. See also Proc. 1983 Automatic Control Conference, San Francisco, CA, June 1983, pp. 1186-1192.

[43] ———, *Some Singular Control Problems with Long Term Average Criterion*, in Proceeding of the Eleventh IFIP Conference on System Modelling and Optimization, Copenhagen, Denmark, July 1983, Lecture Notes in Control and Information Sciences, Springer-Verlag, New York, 1984, pp. 424-432.

[44] ———, *On optimal correction problems with partial information*, Stochastic Analysis and Applications, 3 (1985), to appear.

[45] J. L. MENALDI AND E. ROFMAN, *A continuous multi-echelon inventory problem*, in Proc. Fourth IFAC-IFIP Symposium on Information Control Problems in Manufacturing Technology, Gaithersburg, MD, October 1982, pp. 41-49.

[46] ———, *On stochastic control problems with impulse cost vanishing*, Proc. International Symposium on Semi-Infinite Programming and Applications, Lecture Notes in Economics and Mathematical Systems 215, Springer-Verlag, New York, 1983, pp. 281-294.

[47] P. A. MEYER, *Cours sur les intégrales stochastiques*, Lecture Notes in Mathematics 511, Springer-Verlag, New York, 1976, pp. 245-400.

[48] F. MIGNOT AND J. P. PUEL, *Inéquations d'évolution paraboliques avec convexes dépendant du temps. Applications aux inéquations quasi-variationnelles d'évolution*, Arch. Rat. Mech. Anal., 64(1977), pp. 59-91, see also C.R. Acad. Sc. Paris, Serie A, 280(1975), pp. 259-262.

[49] M. NISIO, *On a non-linear semi-group attached to stochastic optimal control*, Publ. RIMS, Kyoto Univ., 13(1976), pp. 513-537.

[50] J. H. RATH, *Controlled queues in heavy traffic*, Adv. Appl. Prob., 7(1975), pp. 656-671.

[51] ———, *The optimal policy for a controlled Brownian motion process*, SIAM J. Appl. Math., 32(1977), pp. 115-125.

[52] M. ROBIN, *Contrôle impulsionnel des processus de Markov*, These d'Etat, INRIA, Le Chesnay, France, 1978; see also C.R. Acad. Sc. Paris, Serie A, 282(1976), pp. 631-642.

[53] ———, *On some impulse control problems with long run average cost*, this Journal, 19(1981), pp. 333-358.

[54] H. SCARF, *The optimality of (S, s) policies in the dynamic inventory problem*, in Proc. First Stanford Symposium on Mathematical Methods in the Social Sciences, Stanford, CA, Stanford Univ. Press, 1960, pp. 196-202.

[55] S. E. SHREVE, J. P. LEHOCZKY AND D. P. GAVER, *Optimal consumption for general diffusions with absorbing and reflecting barriers*, this Journal, 22(1984), pp. 55-75.

[56] G. E. UHLENBECK AND L. S. ORNSTEIN, *On the theory of Brownian motion*, Physics Rev., 36(1930), pp. 823-841.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 6, November 1985

THE PARAMETER ESTIMATION PROBLEM FOR PARABOLIC EQUATIONS AND DISCONTINUOUS OBSERVATION OPERATORS*

K. KUNISCH† AND L. WHITE‡

Abstract. Parameter estimation problems are studied for a class of linear autonomous parabolic partial differential equations with various fit-to-data criteria, which may be discontinuous with respect to the state variable. We analyze the convergence of Galerkin schemes approximating the optimization problems and generalize results to higher dimensional problems. An example is then presented for the case of point observation fit-to-data criteria in higher dimensions. Finally, discretization of coefficients is discussed for identification problems with variable coefficients.

Key words. parameter estimation of parabolic distributed systems, Galerkin approximations, spline functions, point observations

1. Introduction and statement of the problem. In this paper we study the parameter estimation problem for a class of linear autonomous parabolic partial differential equations. We suppose that information $\mathfrak{z}$ of a physical system is available for which a priori knowledge guides us to choose a mathematical model from a certain class of equations depending on parameters q. The mathematical problem consists in determining q so that some observed part $\mathcal{C}u$ of the state u depending on q best approximates $\mathfrak{z}$ or in very special cases even equals $\mathfrak{z}$. We therefore study

$$\min_{q\in\tilde{Q}} \|\mathcal{C}u(q)-\mathfrak{z}\|^2 \tag{1.1}$$

where $\tilde{Q}$ is some admissible set of parameters, $u(q)$ satisfies the differential equations and $\|\cdot\|$ stands for an appropriately chosen norm. The choice of a quadratic criterion is made on the basis of its widespread use in the applications [16], [23] but other criteria are feasible and could be treated with the methods described below.

The optimization problem that we just described is an infinite dimensional problem, and its approximation by sequences of finite dimensional problems has recently received much attention (see [3]-[6], [8], [23] et al.). In [3]-[6], [17] the convergence problem of approximations to (1.1) is addressed. It is generally assumed that the fit-to-data criterion is continuous in the state variable u of the differential equation, although in some special cases [5] this condition is not used. Moreover, in all the numerical examples of parabolic equations studied in [3]-[6] the fit-to-data criterion was not continuous as a map from the state space to $\mathbb{R}$. It is the purpose of this paper to give a rather complete analysis of the convergence of Galerkin schemes applied to the optimization problem (1.1) without the benefit of this continuity assumption. At the same time results of ([3]-[6], [17] et al.) are generalized to the multi-dimensional case. Much of the motivation for our work is provided by the study of parameter estimation problems in transport equations arising in biological modeling as described in the

* Received by the editors April 12, 1983, and in final revised form September 27, 1984.

† Institut für Mathematik, Technische Universität Graz, A-8010 Graz, Austria and Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019. This author gratefully acknowledges support from the Max Kade Foundation. His research was supported in part by the Fonds zur Förderung der Wissenschaftlichen Forschung, Austria, No. P4534, and by the Steiermärkischen Wissenschafts- und Forschungsförderungsfonds.

‡ Department of Mathematics, University of Oklahoma, Norman, Oklahoma 73019. The research of this author was supported in part by the National Science Foundation grant MCS-7902037 and by the Cooperative Institute for Mesoscale Meteorological Studies.

work of [4]. In those problems the available information of the unknown system usually has the form of point observations, see (M1) below.

We now describe the problem that is studied. Let us consider the equation

$$
\begin{aligned}
&\frac{\partial u}{\partial t} = \sum_{i,j=1}^{n} D^i(a_{ij}(x)D^j u) + \sum_{i=1}^{n} b_i(x)D^i u + c(x)u, \quad \text{for } t>0, \quad x\in\Omega, \\
&u(0,x) = \varphi(x) \quad \text{in } \Omega, \\
&u(t,\cdot)|\partial\Omega = 0,
\end{aligned} \tag{1.2}
$$

where Ω is a bounded region in $\mathbb{R}^n$ with boundary $\partial\Omega$ and D^i stands for differentiation with respect to the ith space variable. We assume that $\varphi \in H^0(\Omega)$ and $u(t,x)\in\mathbb{R}$, although systems of equations could be treated along the same lines. The precise conditions on the coefficients and on $\partial\Omega$ will be given in the next section. In this section we assume sufficient smoothness so that the solution u of (1.2) exists and that the subsequent operations make sense. We shall denote u also by $u(t)$, $u(t,x)$ and $u(t,x;q)$ as it is dictated by the context. Here q stands for the (unknown) parameter vector given by

$$q = ((a_{ij}), (b_i), c, \varphi), \qquad q\in Q,$$

where Q is a subset of an appropriately chosen function space and $\tilde{Q}\subset Q$. In the notation for q we let (a_{ij}) stand for the matrix with element a_{ij}, and similarly (b_i) denotes a vector in $\mathbb{R}^n$, with b_i as ith coordinate.

Next we describe several fit-to-data criteria, corresponding to the general quadratic criterion cited in (1.1). Throughout we direct our attention to a finite time interval $[0, T]$ during which observations can take place. Let μ be a real valued, monotone increasing function and let $I\subset[0,T]$ be measurable with respect to the Lebesgue-Stieltjes measure $d\mu$ and with meas $I\neq 0$. Further assume that $\tilde{\Omega}$ is a Lebesgue measurable subset of Ω with meas $\tilde{\Omega}\neq 0$.

We start with four specific examples corresponding to four different types of observations $\mathcal{Y}$. Let $t_i\in(0,T]$ and $x_j\in\Omega$.

(M1) $\mathcal{Y} = \{z(t_i,x_j)\}\in\mathcal{Z}_1$, $i=1,\cdots,r$, $j=1,\cdots,l$ where $\mathcal{Z}_1=\mathbb{R}^{r+l}$ endowed with the Euclidean norm, (discrete-discrete case),

(M2) $\mathcal{Y} = \{z(\cdot,x_j)\}\in\mathcal{Z}_2$, $j=1,\cdots,l$, $\mathcal{Z}_2=\prod_{j=1}^{l} L^2_\mu(I)$ (continuous-discrete case),

(M3) $\mathcal{Y} = \{z(t_i,\cdot)\}\in\mathcal{Z}_3$, $i=1,\cdots,r$, $\mathcal{Z}_3=\prod_{i=1}^{r} L^2(\tilde{\Omega})$ (discrete-continuous case),

(M4) $\mathcal{Y} = \{z(\cdot,\cdot)\}\in\mathcal{Z}_4$, $\mathcal{Z}_4 = L^2_\mu(I, L^2(\tilde{\Omega}))$, (continuous-continuous case).

Here $L^2_\mu(I,H)$ stands for the usual Sobolev space of H-valued functions, and I is endowed with the Lebesgue-Stieltjes measure $d\mu$; if $H=\mathbb{R}$ then its notation is suppressed. Moreover, $\prod_{i=1}^{r} H$ denotes the product of r copies of the Hilbert space H endowed with the Hilbert-space product norm. Corresponding to these types of observations we consider four quadratic fit-to-data criteria:

$$\tilde{J}_1(q) = \sum_{i,j} |u(t_i,x_j;q) - z(t_i,x_j)|^2,$$

$$\tilde{J}_2(q) = \sum_{j} \int_I |u(t,x_j;q) - z(t,x_j)|^2 \, d\mu,$$

$$\tilde{J}_3(q) = \sum_{i} \int_{\tilde{\Omega}} |u(t_i,x;q) - z(t_i,x)|^2 \, dx,$$

$$\tilde{J}_4(q) = \int_I \int_{\tilde{\Omega}} |u(t, x; q) - z(t, x)|^2 \, dx \, d\mu.$$

These fit-to-data criteria involve point evaluations which will be seen to be justified for $t > 0$ due to the smoothing properties of (1.2). We shall study (1.2) in its usual setting in $H^0(\Omega)$. Then, note that for fixed q, $\tilde{J}_1(q)$ and $\tilde{J}_3(q)$ are not continuous functions of the state variable if the latter is considered as element in $H^0(\Omega)$. We next consider fit-to-data criteria in a more general setting.

Let Z_i, $i = 1, \cdots, 4$, be Hilbert spaces (observation spaces) and let

$$\mathscr{C}_1 \colon C(I, C(\Omega)) \to Z_1, \qquad I \subset (0, T],$$
$$\mathscr{C}_2 \colon L^2_\mu(0, T; C(\Omega)) \to Z_2,$$
$$\mathscr{C}_3 \colon C(I; H^0(\Omega)) \to Z_3, \qquad I \subset (0, T],$$
$$\mathscr{C}_4 \colon L^2_\mu(0, T; H^0(\Omega)) \to Z_4$$

be continuous linear observation operators. Further define $F_i \colon Z_i \to \mathbb{R}$ by

$$F_i \varphi = |\mathscr{C}_i \varphi - \hat{z}|^2_{Z_i},$$

where $\hat{z} \in Z_i$ is the observation, which is assumed to be known (exactly). It is now easily seen that $\tilde{J}_i(q)$ is a special case of

$$J_i(q) = F_i \mathscr{C}_i u(\cdot, \cdot; q).$$

In fact, the observation operator involved to define $\tilde{J}_1$, for example, is given by

$$\tilde{\mathscr{C}}_1 \varphi = \{\varphi(t_i, x_j)\} \in \mathbb{R}^{k+l}.$$

Given observations in a space Z_i so that a continuous observation operator $\mathscr{C}_i$ can be associated and a set of admissible parameters $\tilde{Q}$, we shall investigate approximation schemes of the *parameter estimation problem*, phrased as the optimization problem:

($\mathscr{P}$) minimize $J_i(q) = F_i \mathscr{C}_i u(\cdot, \cdot; q)$, over $q \in \tilde{Q}$, subject to u satisfying (1.2).

The concept of solution of (1.2) is given by the semigroup solution associated with the abstract differential equation

$$\frac{du(t)}{dt} = A(q)u(t), \qquad u(0) = \varphi, \tag{1.3}$$

where $A(q)$ is the elliptic operator associated with the right-hand side of (1.2), as precisely defined in § 2.

We point out that the use of the Lebesgue-Stieltjes measure in (M2), (M4), or analogously in $\mathscr{C}_2$, $\mathscr{C}_4$ is related to the well-known behavior at $t = 0$ of the analytic semigroup associated with (1.2). Recall that if $T(t)$ is an analytic semigroup generated by A in the Hilbert space H with operator norm $\|\cdot\|$, then $\|AT(t)\| \leqq C/t$ and similar estimates will be relevant in the approximation results given in this paper. Therefore for *certain* types of estimates, it will be desirable to have fit-to-data criteria that compensate for this singularity of $t = 0$. Moreover, if there is no observation taken in a neighborhood of $t = 0$, then $\mu(t) = t$ is an acceptable choice for all practical purposes.

To approximate ($\mathscr{P}$) by a sequence of finite dimensional problems let H^N be a sequence of subspaces of H and denote by $P^N \colon H \to H^N$ the orthogonal projections $N = 1, 2, \cdots$. We shall soon describe an operator-theoretic form of Galerkin approximations $A^N(q) \colon H^N \to H^N$ to $A(q)$. The sequence of approximating Cauchy problems

is given by

$$\frac{du^N(t)}{dt} = A^N(q)u^N(t), \qquad u^N(0) = P^N\varphi. \tag{1.4}$$

Note that (1.4) is an equation in H^N. Let $J_i^N(q)$ be given by $J_i(q)$ with $u(\cdot, \cdot; q)$ replaced by the solution $u^N(\cdot, \cdot; q)$ of (1.4).

The *approximation of the parameter estimation problem* then becomes to

(a) solve

$$(\mathcal{P}^N) \qquad \underset{q \in \tilde{Q}}{\text{minimize}}\, J_i^N(q) \text{ subject to } u^N \text{ satisfying (1.4)}$$

and

(b) to establish conditions such that the solutions $\bar{q}^N$ of $(\mathcal{P}^N)$ (or a subsequence thereof) converge to some $q^* \in \tilde{Q}$ with q^* a solution of $(\mathcal{P})$ and

$$u^N(t, \cdot; \bar{q}^N) \to u(t, \cdot; q^*) \text{ in } H^0(\Omega), \qquad t \in [0, T], \tag{1.5}$$

$$J_i^N(\bar{q}^N) \to J_i(q^*). \tag{1.6}$$

DEFINITION 1.1. A sequence $(H^N, A^N(q), \mathcal{C}_i)$ is called *parameter estimation convergent* (PEC) scheme for (1.2), if $(\mathcal{P}^N)$ has a solution $\bar{q}^N$ for $N = 1, 2, \cdots$, if there exists a convergent subsequence $\bar{q}^{N_k} \to q^*$ with q^* a solution of $(\mathcal{P})$ and if (1.5), (1.6) hold for the subsequence. If there exists more than one convergent subsequence, then all of the limits must satisfy $(\mathcal{P})$ with (1.5), (1.6) holding.

Clearly (1.5) implies (1.6) with J_i taken as $\tilde{J}_3$; if (1.5) is replaced by a convergence assumption in $L^2_\mu(0, T; H^0(\Omega))$, then (1.6) holds for J_4. In these cases it is also quite standard to show that a scheme is parameter estimation convergent (see [4]-[6], [17]). The major technical step that has to be accomplished to establish PEC in the general case is to show that the Galerkin approximations (1.4) converge uniformly in the parameter q to the original problem (1.3) in a norm that is finer than the $H^0(\Omega)$-norm, in spite of the fact that the projections are taken with respect to the $H^0(\Omega)$ inner product. One remedy to this difficulty in the case of (spatial) point observations would be to treat the whole problem in a finer topology, say $H^1(\Omega)$ for one-dimensional domains, or $H^2(\Omega)$ for two- and three-dimensional domains. We will not follow this idea, however, since it restricts the possible choices for subspaces H^N requiring more smoothness of the elements in H^N and since the projections become more difficult to calculate (due to the more complicated inner product structure). Instead we make precise use of the location of the spectrum of the stationary equation associated with (1.2). Here the strict ellipticity assumption is essential. For our convergence analysis we use a technique developed in [11], where finite element approximations of the state of (1.1) (q fixed) are studied, by first considering the (rate of) convergence of the resolvents of A^N in $H^0(\Omega)$ and subsequently employing the representation of the solution semigroup by the Dunford integral formula.

Numerical experiments for the choice of H^N as cubic spline functions are documented in [4], [5] for constant and in [3] for variable coefficients. After completion of this manuscript preliminary numerical studies were carried out comparing the use of the $\tilde{J}_3(q)$ criterion to the $\tilde{J}_1(q)$ criterion to identify spatially and temporally varying diffusion coefficients in a one-dimensional parabolic equation [32]. For the temporally varying coefficient there is no essential difference in the numerical results obtained when using the $\tilde{J}_1$ and the $\tilde{J}_3$ criterion. For the spatially varying coefficient, however, examples could be found where the $\tilde{J}_1$ criterion gives faster convergence than the $\tilde{J}_3$

criterion as well as "better" estimates; in some cases the $\tilde{J}_3$ criterion would even fail to lead to convergence of the minima of the approximating problems.

In § 2 we present the main parameter dependent convergence theorems of this paper. These results are subsequently applied to establish PEC for various fit-to-data criteria. Many of the proofs of this section are given in Appendix A. The techniques developed in § 2 do not directly apply to verify PEC in the presence of point observations as in (M1) and (M2) if the dimension of the domain is higher than one. In § 3 we take up the problem of showing PEC in the presence of point observations if the dimension of the domain is higher than one. As a general technique to handle this case we propose to first show convergence (uniform in q) in the $H^1(\Omega)$-norm with a sufficiently high rate and then to use the inverse assumption [2, p. 89] together with Sobolev's embedding theorem (in $H^2(\Omega)$ for dimensions two and three). This is demonstrated by means of an example. The proofs of the technical lemmas of § 3 are postponed until Appendix B. To avoid some cumbersome notation, we do not discretize the coefficients in (1.1) in §§ 2 and 3. A simultaneous approximation of the coefficients is carried out in § 4. We summarize our results in the section "Conclusions". The notation that is used is standard. Norms are denoted by $|\cdot|$ throughout and we generally use a subscript to denote the space in which the norm is taken. Operator norms are denoted by $\|\cdot\|$. The subscript for the norm in $H^0(\Omega)$ is dropped whenever this seems appropriate. The inner product in $H^0(\Omega)$ is designated by $(\cdot,\cdot)$. In our notation for Sobolev spaces we generally follow [1], and we specify preimage and image space, unless the latter is $\mathbb{R}^j$ for some $j\in\mathbb{N}$. For a linear operator A the resolvent set is denoted by $\rho(A)$. Throughout we frequently use a generic constant C in our estimates.

2. Parameter estimation convergence. In this section we present the main approximation results of this paper. It is convenient to repeat the equation under study:

$$
\begin{aligned}
&\frac{du}{dt} = A_0(q)u(t,\cdot) \quad \text{in } (0,T]\times\Omega\subset\mathbb{R}^{n+1},\\
&u(0,x)=\varphi(x) \quad \text{in } \Omega,\\
&u(t,\cdot)|_{\partial\Omega}=0, \quad \text{for } t>0,
\end{aligned}
\tag{2.1}
$$

where $A_0(q)u$ is formally defined by

$$
A_0(q)u=\sum_{i,j=1}^{n} D^i(a_{ij}(x)D^j u)+\sum_{i=1}^{n} b_i(x)D^i u + c(x)u,
$$

with

$$
q=((a_{ij}),(b_i),c,\varphi)\in Q\subset (W^{1,p}(\Omega,\mathbb{R}^{n\times n})\times L^p(\Omega,\mathbb{R}^n)\times L^{\hat{p}}(\Omega,\mathbb{R})\times H^0(\Omega,\mathbb{R})),
$$

with $p>n$, $\hat{p}=\max(p,4)/2$. Here (a_{ij}) denotes a symmetric matrix with elements a_{ij} and similarly (b_i) denotes a vector in $\mathbb{R}^n$. The domain Ω is assumed to be bounded and either a parallelepiped or with a C^2-boundary $\partial\Omega$. In the case $n=1$, we take $\Omega=(0,1)$. Throughout we frequently use the convention to write $c\in Q$ if for some $(a_{ij}),(b_i),\varphi$ we have $((a_{ij}),(b_i),c,\varphi)\in Q$, and similarly for other combinations of variables. Further $P=\{((a_{ij}),(b_i),c)$: for some φ we have $((a_{ij}),(b_i),c,\varphi)\in Q\}$, and Q and P are endowed with their natural product topologies.

Let us briefly outline the contents of this section. After stating the technical assumptions we summarize properties of $A_0(q)$ and their consequences in a series of lemmas. These results depend on the location of the spectrum and consequently on the strong uniform (in q) ellipticity assumption on $A_0(q)$ and the parabolicity of (2.1).

Next approximation of the resolvents is discussed in several propositions. The proofs of the lemmas and the propositions, some of which are generalizations from [11], can be found in Appendix A. Parameter dependent approximation of the solution semigroup associated with (2.1) is obtained in theorems that follow. These finally can be used to verify PEC for Galerkin approximations of $(\mathcal{P})$ for various choices of observation operators (Theorems 2.6–2.8).

We summarize conditions that will be needed as the theory is developed.

(H1) There exists $\nu > 0$ such that $\nu \sum_{i=1}^n \xi_i^2 \leqq \sum_{i,j=1}^n a_{ij}\xi_i\xi_j$, for all $(a_{ij}) \in Q$ and $(\xi_i) \in \mathbb{R}^n$.

(H2) There exist constants $\mu > 0$, $\tilde{\mu} > 0$ such that Q is a closed convex subset of

$$\{((a_{ij}), (b_i), c, \varphi): |a_{ij}|_{W^{1,p}(\Omega)} \leqq \mu, |b_i|_{L^p(\Omega)} \leqq \mu,$$
$$|c|_{L^{\hat{p}}} \leqq \mu, c(x) \leqq -\tilde{\mu} \text{ for almost every } x \in \Omega, |\varphi|_{H^0(\Omega)} \leqq \mu\}$$

where $p > n$, $\hat{p} = \max(p, 4)/2$.

Note that by Sobolev's embedding theorem it follows that for some constant μ_1,

$$\sum_{i,j=1}^n a_{ij}\xi_i\xi_j \leqq \mu_1 \sum_{i=1}^n \xi_i^2$$

for all $(a_{ij}) \in Q$ and $(\xi_i) \in \mathbb{R}^n$.

(H3) The set $\tilde{Q} \subset Q$ is a compact subset of $W^{1,p}(\Omega, \mathbb{R}^{n\times n}) \times L^p(\Omega, \mathbb{R}^n) \times L^{\hat{p}/2}(\Omega, \mathbb{R}) \times H^0(\Omega, \mathbb{R})$.

(H3*) The set $\tilde{Q} \subset Q$ is a compact subset of $W^{1,p}(\Omega, \mathbb{R}^{n\times n}) \times L^p(\Omega, \mathbb{R}^n) \times L^{\hat{p}/2}(\Omega, \mathbb{R}) \times H_0^1(\Omega, \mathbb{R})$.

(H4, k) The set $\tilde{Q} \subset Q$ is a closed and bounded (by a constant κ) subset of $C^k(\bar{\Omega}, \mathbb{R}^{n\times n}) \times C^k(\bar{\Omega}, \mathbb{R}^n) \times C^{k-1}(\bar{\Omega}, \mathbb{R}) \times H^0(\Omega)$.

(H4*, k) The set $\tilde{Q} \subset Q$ is a closed and bounded (by a constant κ) subset of $C^k(\bar{\Omega}, \mathbb{R}^{n\times n}) \times C^k(\bar{\Omega}, \mathbb{R}^n) \times C^{k-1}(\bar{\Omega}, \mathbb{R}) \times H^1(\Omega)$.

Conditions (H1) and (H2) are common conditions in elliptic operator theory; here they are assumed to hold uniformly in Q. The condition on c to be negative is not a stringent one. We will soon associate with (2.1) a semigroup $e^{tA(q)}$ and subtracting a multiple of the identity αI from $A_0(q)$ only changes this semigroup to $e^{-\alpha t}e^{tA(q)}$. We next discuss some of the consequences of (H1) and (H2) in some detail.

Let $B(q): H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{C}$ denote the sesquilinear form

$$B(q)\langle u, v\rangle = \int_\Omega \left\{ \sum_{i=1}^n \sum_{j=1}^n a_{ij} D^j u D^i \bar{v} - \left(\sum_{i=1}^n b_i D^i u + cu \right) \bar{v} \right\} dx. \tag{2.2}$$

Certainly from (H1) and (H2), it follows that

$$|B(q)\langle u, v\rangle| \leqq C_1 |u|_{H^1(\Omega)} |v|_{H^1(\Omega)}, \tag{2.3}$$

for some constant $C_1 = C_1(\mu, \Omega, n)$ (see [18, 3.4.18]).

Given conditions (H1) and (H2) we have (cf. [18, pp. 45, 144] for any $v \in H_0^1(\Omega)$

$$|v|_{H^1(\Omega)}^2 + \frac{4\tilde{\mu}}{\nu} |v|_{H^0(\Omega)}^2 \leqq \frac{4}{\nu} \operatorname{Re} B(q)\langle v, v\rangle + \delta |v|_{H^0(\Omega)}^2,$$

where $\delta = \delta(p, \mu, n, \nu, \Omega)$ can be calculated explicitly. If $4\tilde{\mu} \geqq \nu\delta$, then

$$\operatorname{Re} B(q)\langle v, v\rangle \geqq C_2 |v|_{H^1(\Omega)}^2 \tag{2.4}$$

where $C_2 = C_2(n, \mu, \tilde{\mu}, \nu, p, \Omega)$. Here and throughout we take $q \in Q$ unless otherwise specified.

The bilinear form $B(q)$ may be used to define operators $A(q)$ in $H^0(\Omega)$. An element $u \in H_0^1(\Omega)$ belongs to dom $A(q)$ if the map $v \to B(q)\langle u, v\rangle$ is continuous from $H_0^1(\Omega)$ to $\mathbb{C}$ with respect to the $H^0(\Omega)$ topology. Since $H_0^1(\Omega)$ is dense in $H^0(\Omega)$, the map $v \to B(q)\langle u, v\rangle$ may be continuously extended to $H^0(\Omega)$. Hence, there exists a unique $w \in H^0(\Omega)$ such that

$$-B(q)\langle u, v\rangle = (w, v)$$

for all $v \in H_0^1(\Omega)$. Set $A(q)u = w$ for $u \in \text{dom } A(q)$. A consequence of the Lax-Milgram theorem is that dom $A(q)$ is dense in $H^0(\Omega)$ for every $q \in Q$ and that

$$A(q)(\text{dom } A(q)) = H^0(\Omega)$$

([10, Chap. 4]). By (2.4) we have for all $u \in \text{dom } A(q)$

$$\text{Re}\,(A(q)u, u) \leqq -C_2|u|^2_{H^1(\Omega)}. \tag{2.5}$$

The Lumer-Phillips theorem ([22, p. 17]) implies that $A(q)$ generates a C_0-semigroup of contractions $T(t; q)$ on $H^0(\Omega)$ with $|T(t; q)|_{L^2(\Omega)} \leqq e^{-C_2 t}$. For the results so far, the conditions on $\partial\Omega$ were stronger than necessary. They will be needed for the following properties of $A(q)$, however. It will be useful to know that

$$\text{dom } A(q) = H_0^1(\Omega) \cap H^2(\Omega)$$

(see [18, Thm. 3.9.1 and § 3.4.10]). We also make use of the following a priori estimate

$$|v|_{H^2(\Omega)} \leqq C|A(q)v| \tag{2.6}$$

for all $v \in H_0^1(\Omega) \cap H^2(\Omega)$, [18] where $C = C(\mu, \Omega, n, \nu)$ but independent of $q \in Q$. From (2.6) we deduce that for $\lambda \in \mathbb{C}$ with $\text{Re}\,\lambda \geqq 0$

$$|(\lambda - A(q))^{-1}\nu|_{H^2(\Omega)} \leqq 2C|v|_{H^0(\Omega)} \tag{2.7}$$

holds for all $v \in H^0(\Omega)$. We shall also need estimates similar to (2.7) for stronger norms. This necessitates the stronger smoothness assumptions (H4, k) on the coefficients and on the boundary $\partial\Omega$. We summarize from [12, p. 177].

If $\partial\Omega$ is of class $C^{(k+2)}$, (H1), (H2), (H4, $k+1$) hold and $f \in H^k(\Omega)$, then $u \in H^{k+2}(\Omega)$, where $A(q)u = f$ and

$$|A(q)^{-1}f|_{H^{k+2}(\Omega)} \leqq C(|f|_{H^k(\Omega)} + |A(q)^{-1}f|_{H^0(\Omega)}) \tag{2.8}$$

with $C = C(n, \nu, \Omega, \kappa)$. In Remark 2.1 below we verify that (2.8) holds in fact for $A(q)$ replaced by $\lambda - A(q)$ for all λ outside a certain sector in $\mathbb{C}$ with C independent of λ.

The sequence of operators $A^N(q)$ is defined next. First we make the following assumption, which will be used for fixed $s \geqq 2$.

(H5, s) H^N is a finite dimensional linear space with $H^N \subset W_0^{1,\infty}(\Omega) \cap C(\bar{\Omega})$. Moreover, there exists $\rho(N)$ with $\lim_{N\to\infty} \rho(N) = 0$, such that for every $\varphi \in \text{dom } A^{s/2}$, when s is even, ($\varphi \in \text{dom } (A(q)^{(s-1)/2})$ with $A(q)^{(s-1)/2}\varphi \in H_0^1(\Omega)$, when s is odd), and every $N = 1, 2, 3, \cdots$, there exists an element $\hat{\varphi}^N \in H^N$ such that

$$|\varphi - \hat{\varphi}^N|_{H^1(\Omega)} \leqq \rho(N)|\varphi|_{H^s(\Omega)},$$

and

$$|\varphi - \hat{\varphi}^N|_{H^0(\Omega)} \leqq \rho(N)|\varphi|_{H^{s-1}(\Omega)}.$$

Let $P^N: H^0(\Omega) \to H^N$ denote the orthogonal projection. The restriction of $B(q)$ to $H^N \times H^N$ defines uniquely a bounded linear operator $A^N(q): H^N \to H^N$ by

$$B(q)\langle u, v\rangle = (-A^N(q)u, v) \quad \text{for all } u, v \in H^N. \tag{2.9}$$

Certainly, $A^N(q)$ generates a C_0 semigroup $T^N(t; q) = \sum_{i=0}^{\infty} (A^N(q)^i / i!)$ on H^N.

Let $\{B_i^N\}_{i=1}^{k^N}$ be a basis for H^N. Then the solutions $u^N(t; q)$ of (1.4) are characterized by

$$\begin{aligned} &\frac{d}{dt}(u^N(t; q), B_i^N) = -B(q)\langle u^N(t; q), B_i^N\rangle, \\ &(u^N(0; q), B_i^N) = (\varphi, B_i^N) \end{aligned} \tag{2.10}$$

for all $i = 1, \cdots, k^N$. Here $(\cdot, \cdot)$ denotes the inner product in $H^0(\Omega)$. Let us discuss two special cases that have been frequently used in approximation and parameter estimation problems (see [3]-[6], [17], for example).

Example 2.1. If H^N is a finite dimensional subspace of dom A, then $A^N(q)$ as defined in (2.9) is given by $P^N A(q)$. Approximations of this form with H^N chosen as a subspace of spline functions or eigenfunctions were studied in [3]-[6].

Example 2.2. Let us assume (H1), (H2), that $H^N \subset W_0^{1,\infty}(\Omega)$ and that $A(q)$ can be expressed as $A(q) = -T^*(q)T(q) + C(q)$ where $C(q)$ and $T(q)$ are bounded linear operators from $H_0^1(\Omega)$ into $H^0(\Omega)$. For $u, v \in \text{dom } A$ we have $(A(q)u, v) = -(T(q)u, T(q)v) + (C(q)u, v)$ and by assumption the sesquilinear form

$$\sigma_q(u, v) = -(T(q)u, T(q)v) + (C(q)u, v)$$

can be uniquely extended to $H_0^1(\Omega) \times H_0^1(\Omega)$. Therefore, it coincides with $B(q)$ on this set. Since $H^N \subset H_0^1(\Omega)$, we find that the approximating operators $A^N(q)$ in (2.9) can be expressed as $A^N(q) = -(T(q)P^N)^* T(q) + P^N C(q)$. For further details on these schemes we refer to [29], for approximation of the state and to [17] for approximations of the parameter estimation problem in $H^0(\Omega)$.

We now summarize some technical results in a series of lemmas. The following notation will be needed. Let

$$S_R(\theta, \gamma) = \{z \in \mathbb{C}: |\arg(z - \gamma)| \leqq \theta\}$$

and

$$S_L(\theta, \gamma) = \{z \in \mathbb{C}: \pi - \theta \leqq |\arg(z - \gamma)| \leqq \pi\},$$

where $\gamma \in \mathbb{R}$ and $\theta \in [0, \pi/2)$. By $S_L^C(\vartheta, \gamma)$ we denote the complement of $S_L(\vartheta, \gamma)$ in $\mathbb{C}$.

LEMMA 2.1. *Let* (H1)-(H2) *hold. Then*

(a) $\{(A(q)u, u): u \in \text{dom } A(q), |u|_{H^1(\Omega)} = 1, q \in Q\} \subset S_L(\arctan(2C_1/C_2), -C_2/2)$,

(b) $|\text{Im}(A(q)u, u)| \leqq -(2C_1/C_2)(\text{Re}(A(q)u, u) + (C_2/2)|u|^2_{H^1(\Omega)})$, *for* $u \in \text{dom } A(q)$, $q \in Q$,

(c) $\{B(q)\langle u, u\rangle: u \in H_0^1(\Omega)\} \subset S_R(\arctan(2C_1/C_2), C_2/2)$,

(d) $\rho(A(q)) \supset S_L^C(\theta, -C_2/2)$, *for any* $\theta > \arctan(2C_1/C_2)$,

(e) $A(q)$ *generates an analytic semigroup* $T(t; q)$, *given by*

$$T(t; q) = \frac{1}{2\pi i} \int_\Gamma e^{\lambda t}(\lambda - A(q))^{-1}\, d\lambda$$

where Γ *is any positively oriented contour in* $\rho(A(q))$ *with* $\arg \lambda \to \pm\theta$ *as* $|\lambda| \to \infty$ *with* $\theta \in (\pi/2, \pi)$.

(f) *There exists a constant* C *independent of* $\alpha \in [0,1]$, $q \in Q$ *and* $\lambda \in S_L^C(\arctan(2C_1/C_2), 0)$ *such that*

$$\|(-A(q))^{\alpha}(\lambda - A(q))^{-1}\| \leqq \frac{C}{(|\lambda|+1)^{1-\alpha}},$$

(g) $\|(\lambda - A(q))^{-1}\| \leqq 1/(\operatorname{Re}\lambda + C_2)$ *for all* λ *with* $\operatorname{Re}\lambda > C_2$.

(h) *For all* $\alpha \in [0,\infty)$ *there exists a constant* C_α *(not depending on* $q \in Q$*) such that* $|(-A(q))^{\alpha}T(t;q)| \leqq C_\alpha t^{-\alpha} e^{-tC_2}$.

LEMMA 2.2. *Let* (H1)-(H2) *hold and let* $\theta_1 \in (\arctan(2C_1/C_2), \pi/2)$. *Then there exists a constant* $C = C(C_1, C_2, \theta_1)$ *such that*

$$|\lambda||u|^2 + |u|^2_{H^1(\Omega)} \leqq C|\lambda|u|^2 + B(q)\langle u, u\rangle|$$

for all $q \in Q$, $u \in H_0^1(\Omega)$ *and* $\lambda \in S_L^C(\theta_1, 0)$.

COROLLARY 2.1. *If* (H1) *and* (H2) *hold then*

$$|(\lambda - A(q))^{-1}u|_{H^1(\Omega)} \leqq \frac{C}{\sqrt{|\lambda|+1}}|u|$$

for all $q \in Q$, $u \in H^0(\Omega)$ *and* $\lambda \in S_L^C(\theta_1, 0)$.

Proof. By Lemma 2.2 we find for $u \in \operatorname{dom} A(q)$ that $(|\lambda|+1)|u| \leqq C|(\lambda - A(q))u|$ and $|u|^2_{H^1(\Omega)} \leqq C|(\lambda - A(q))u||u|$. From these two estimates we obtain the corollary with the same constant C as in Lemma 2.2.

LEMMA 2.3. *Let* (H1) *and* (H2) *hold. Then* $\operatorname{dom}(-(A(q))^{1/2}) = H_0^1(\Omega)$ *and*

$$\frac{1}{C}|u|^2_{H^1(\Omega)} \leqq |A^{1/2}(q)u|^2_{H^0(\Omega)} \leqq C|u|^2_{H^1(\Omega)} \tag{2.11}$$

for all $u \in H_0^1(\Omega)$ *with* C *independent of* u *and* q.

Remark 2.1. The estimates developed so far allow us to generalize (2.8). Let $\partial\Omega$ be of class $C^{(k+2)}$, let (H1), (H2), (H4, $k+1$) hold and let $f \in \operatorname{dom}(A^{k/2})$ for k even and $f \in \operatorname{dom}(A^{(k-1)/2})$ with $A^{(k-1)/2}f \in H_0^1(\Omega)$ for k odd. Then

$$|(\lambda - A(q))^{-1}f|_{H^{k+2}(\Omega)} \leqq C(|f|_{H^k(\Omega)} + |(\lambda - A(q))^{-1}f|_{H^0(\Omega)}), \tag{2.8*}$$

where $C = C(n, \nu, \Omega, \kappa)$ is independent of $\lambda \in S_L^C(\arctan(2C_1/C_2), 0)$. For the proof we refer to Appendix A.

LEMMA 2.4. *Let* (H1) *and* (H2) *hold and let* $\theta_1 \in (\arctan(2C_1/C_2), \pi/2)$. *Then for all* $\lambda \in S_L^C(\theta_1, 0)$, $q \in Q$ *and* $u \in H^N$, $N = 1, 2, \cdots$, *we have*

(a) $|\lambda||u|^2_{H^0(\Omega)} + |u|^2_{H^1(\Omega)} \leqq C|\lambda|u|^2 - (A^N(q)u, u)|$,

(b) $S_L^C(\arctan(2C_1/C_2), 0) \subset \rho(A^N(q))$,

(c) $|(\lambda - A^N(q))^{-1}u|_{H^0(\Omega)} \leqq C/(|\lambda|+1)|u|_{H^0(\Omega)}$,

(d) $|(\lambda - A^N(q))^{-1}u|_{H^1(\Omega)} \leqq C/(\sqrt{|\lambda|+1})|u|_{H^0(\Omega)}$,

(e) $|(\lambda - A^N(q))^{-1}u|_{H^0(\Omega)} \leqq 1/(\operatorname{Re}\lambda + C_2)$ *for all* λ *with* $\operatorname{Re}\lambda > 0$.

In the following three propositions we study the convergence properties of the resolvent. We first introduce some additional notation. We recall the definition of P and denote by $|\cdot|_P$ the norm induced from $W^{1,p}(\Omega, \mathbb{R}^n \times \mathbb{R}^n) \times L^p(\Omega, \mathbb{R}^n) \times L^{\hat{p}}(\Omega, \mathbb{R})$.

PROPOSITION 2.1. *Let* (H1) *and* (H2) *hold and let* $q^N \to q^0$ *in* P. *Then there exists a constant* C *(independent of* φ, λ, q *and* N*) such that for all* $\lambda \in S_L^C(\theta_1, 0)$

(a) $|(\lambda - A(q^N))^{-1}\varphi - (\lambda - A(q^0))^{-1}\varphi|_{H^0(\Omega)} \leqq (C/|\lambda|)|q^N - q^0|_P|\varphi|_{H^0(\Omega)}$,

(b) $|(\lambda - A(q^N))^{-1}\varphi - (\lambda - A(q^0))^{-1}\varphi|_{H^1(\Omega)} \leqq (C/\sqrt{|\lambda|+1})|q^N - q^0|_P|\varphi|_{H^0(\Omega)}$,

(c) $|(\lambda - A(q^N))^{-1}\varphi - (\lambda - A(q^0))^{-1}\varphi|_{H^2(\Omega)} \leqq C|q^N - q^0|_P|\varphi|_{H^0(\Omega)}$, *for* $\varphi \in H^0(\Omega)$.

If in addition $\varphi \in H_0^1(\Omega)$, *then*

(d) $|(\lambda - A(q^N))^{-1}\varphi - (\lambda - A(q^0))^{-1}\varphi|_{H^1(\Omega)} \leqq (C/(|\lambda|+1))|q^N - q^0|_P |\varphi|_{H^1(\Omega)}$.

In the following estimates we will not distinguish between various functions ρ with $\lim_N \rho(N) = 0$, and which differ from ρ used in (H5, s) only by a constant multiple $C = C(C_1, C_2, \kappa, \mu, \vartheta_1, \Omega, n)$ independent of λ, q and φ. We further set for $\lambda \in S_L^C(\theta_1, 0)$

$$e_q^N(\lambda) = (\lambda - A(q))^{-1}\varphi - (\lambda - A^N(q))^{-1}P^N\varphi.$$

PROPOSITION 2.2. *Let* (H1), (H2) *and* (H5, s) *hold, and let* $\varphi \in H^0(\Omega)$, $\lambda \in S_L^C(\theta_1, 0)$, *and* $q \in Q$. *Then*

(a) $$|e_q^N(\lambda)|_{H^1(\Omega)} \leqq \rho(N)|\varphi|_{H^0(\Omega)}, \qquad |e_q^N(\lambda)|_{H^0(\Omega)} \leqq \frac{\rho(N)}{\sqrt{|\lambda|+1}}|\varphi|_{H^0(\Omega)}.$$

(b) *More generally, if* s *is even,* $s \geqq 2$, *if* (H4, $s-1$) *and* (H5, s) *hold, if* $\partial\Omega$ *is of class* C^s, *and* $\varphi \in \operatorname{dom}(A^{(s/2)-1})$ *then*

$$|e_q^N(\lambda)|_{H^1(\Omega)} \leqq \rho(N)|\varphi|_{H^{s-2}(\Omega)}$$

and

$$|e_q^N(\lambda)|_{H^0(\Omega)} \leqq \frac{\rho(N)}{\sqrt{|\lambda|}}|\varphi|_{H^{s-2}(\Omega)}.$$

(c) *If* s *is odd,* (H4, $s-1$), (H5, s) *hold and if* $\partial\Omega$ *is of class* C^2, *and* $\varphi \in \operatorname{dom}(A^{((s-2)/2)-1})$, *with* $A^{((s-1)/2)-1}\varphi \in H_0^1(\Omega)$, *then the estimates in* (b) *hold.*

PROPOSITION 2.3. *Let* (H1), (H2), (H4, $s-1$), (H5, s) *hold and assume* $\partial\Omega$ *to be of class* C^s. *Further let* $\varphi \in \operatorname{dom}(A^{(s/2)-1})$ *with* $A^{(s/2)-1}\varphi \in H_0^1(\Omega)$ *if* s *is even and* $\varphi \in \operatorname{dom}(A^{(s-1)/2})$ *if* s *is odd,* $s \geqq 2$. *Then*

$$|e_q^N(\lambda)|_{H^1(\Omega)} \leqq \frac{\rho(N)}{\sqrt{|\lambda|+1}}|\varphi|_{H^{s-1}(\Omega)},$$

and

$$|e_q^N(\lambda)|_{H^0(\Omega)} \leqq \frac{\rho(N)}{|\lambda|+1}|\varphi|_{H^{s-1}(\Omega)}.$$

For $s = 2$ *it suffices that* (H1), (H2), (H5, s) *hold, and that* $\varphi \in H_0^1(\Omega)$.

Finally we are prepared to study the parameter dependent convergence of the semigroups. Our first method of proof will be that of integral representation of the approximating as well as the original semigroups.

THEOREM 2.1. *Let* (H1), (H2), *and* (H5, 2) *hold and let* $q^N \to q^0$ *in* Q. *Then*

(a) $|T(t; q^0)\varphi^0 - T^N(t; q^N)P^N\varphi^N| \leqq C(|q^0 - q^N|_P + \rho(N)(1/\sqrt{t})|\varphi^0| + |\varphi^N - \varphi^0|$,

(b) $|T(t; q^0)\varphi^0 - T^N(t; q^N)P^N\varphi^N|_{H^1(\Omega)}$
$\leqq C((1/\sqrt{t})|q^0 - q^N|_P + \rho(N)(1/t))|\varphi^0| + (C/\sqrt{t})|\varphi^N - \varphi^0|$.

If in addition $q^N \to q^0$ *in* $P \times H_0^1(\Omega)$, *then*

(c) $|T(t; q^0)\varphi^0 - T^N(t; q^N)P^N\varphi^N| \leqq C|q^0 - q^N|_P|\varphi^0| + \rho(N)|\varphi^0|_{H^1(\Omega)} + |\varphi^N - \varphi^0|$,

(d) $|T(t; q^0)\varphi^0 - T^N(t; q^N)P^N\varphi^N|_{H^1(\Omega)}$
$\leqq C(|q^0 - q^N|_P + (\rho(N)/\sqrt{t})|\varphi^0|_{H^1(\Omega)} + C|\varphi^0 - \varphi^N|_{H^1(\Omega)}$.

Remark 2.2. In Lemma 2.5 below we will demonstrate that under additional assumptions $\|T^N(t; q)\|_{H^1(\Omega)} \leqq e^{Ct}$. If in addition $\|P^N\|_{H_0^1}$ (i.e., the $H^0(\Omega)$-projection restricted to $H^1(\Omega)$) is uniformly bounded, we can replace the bound in (d) by the

following estimate:

$$|T(t;q^0)\varphi^0 - T^N(t;q^N)P^N\varphi^N|_{H^1(\Omega)}$$
$$\leqq C(|q^0-q^N|_P+\rho(N)/\sqrt{t})|\varphi^0|+C|\varphi^0-\varphi^N|_{H^1(\Omega)} \quad \text{for } q^N\to q^0 \text{ in } Q.$$

Proof of Theorem 2.1. Let $\Gamma_1=\cup_{i=1}^3\Gamma_1^i$, with $\Gamma_1^1=\{\rho e^{-i\vartheta_2}\colon \rho\geqq 1\}$, $\Gamma_1^2=\{e^{i\rho}\colon -\vartheta_2\leqq\rho\leqq\vartheta_2\}$, and $\Gamma_1^3=\{\rho e^{i\vartheta_2}\colon \rho\geqq 1\}$, where $\vartheta_2=\pi-\vartheta_1$. Let us observe that for an appropriately defined $\tilde{C}$ we have

$$\int_{\Gamma_1}|e^{\lambda t}||d\lambda|\leqq\tilde{C}t^{-1},$$
$$\int_{\Gamma_1}\frac{|e^{\lambda t}|}{\sqrt{|\lambda|+1}}|d\lambda|\leqq\tilde{C}t^{-1/2},$$
$$\int_{\Gamma_1}\frac{|e^{\lambda t}|}{|\lambda|+1}\leqq\tilde{C} \quad \text{for } t\geqq 0.$$

To verify (a) we note that by Lemma 2.1(e) and the Hille-Yosida theorem we have $|T^N(t;q)\varphi|\leqq|\varphi|$ for all $\varphi\in H^N$ and $q\in Q$, and consequently $\|T^N(t;q)P^N\|\leqq 1$ for all $q\in Q$. Further, by Lemma 2.1(e)

$$|T(t;q^0)\varphi^0 - T^N(t;q^N)P^N\varphi^N|$$
$$\leqq\frac{1}{2\pi}\int_{\Gamma_1}|e^{\lambda t}[(\lambda-A(q^0))^{-1}-(\lambda-A(q^N))^{-1}]\varphi^0||d\lambda|$$
$$+\frac{1}{2\pi}\int_{\Gamma_1}|e^{\lambda t}[(\lambda-A(q^N))^{-1}\varphi^0-(\lambda-A^N(q^N))^{-1}P^N\varphi^0]||d\lambda|$$
$$+|T^N(t;q^N)P^N(\varphi^0-\varphi^N)|$$
$$\leqq C|q^N-q^0|_P|\varphi^0|\int_{\Gamma_1}|e^{\lambda t}|\frac{|d\lambda|}{\sqrt{|\lambda|+1}}+\rho(N)\int_{\Gamma_1}|e^{\lambda t}|\frac{|d\lambda|}{\sqrt{|\lambda|+1}}+|\varphi^0-\varphi^N|.$$

For the last inequality we have used Proposition 2.1(a) and Proposition 2.2(a). The above estimates imply (a). To verify (b), we note that

$$|T(t;q^0)\varphi - T^N(t;q^N)P^N\varphi^N|_{H^1(\Omega)}$$
$$\leqq|T(t;q^0)\varphi^0-T(t;q^N)\varphi^0|_{H^1(\Omega)}+|T(t;q^N)\varphi^0-T^N(t;q^N)P^N\varphi^0|_{H^1(\Omega)}$$
$$+|T^N(t;q^N)P^N\varphi^0-T^N(t;q^N)P^N\varphi^N|_{H^1(\Omega)}$$
$$=\mathrm{I}+\mathrm{II}+\mathrm{III}.$$

From Lemma 2.4(d) it follows that $|T^N(t;q^N)P^N\varphi-T^N(t;q^N)P^N\varphi^N|_{H^1(\Omega)}\leqq Ct^{-1/2}|\varphi^0-\varphi^N|_{H^1(\Omega)}$. Estimates on I and II are obtained by Proposition 2.1(b) and Proposition 2.2(a). The remaining estimates (c) and (d) are proved by using Proposition 2.1(a) and Proposition 2.3 (with $s=2$) for (c) and Proposition 2.3 for (d).

If the initial state φ is more regular, we can expect a higher rate of convergence in condition (H5, s). For simplicity of presentation we assume that φ is known in our next result.

THEOREM 2.2. *Let* (H1), (H2), (H4, $s-1$) *and* (H5, s) *hold, let* $\partial\Omega$ *be of class* C^s, *and let* $q^N\to q^0$ *in* $C^{s-1}(\bar{\Omega};\mathbb{R}^{n\times n})\times C^{s-1}(\bar{\Omega};\mathbb{R}^n)\times C^{s-2}(\bar{\Omega};\mathbb{R})$, $s\geqq 2$. *Then*

(a) $|T(t;q^0)\varphi-T^N(t;q^N)P^N\varphi|\leqq C|q^0-q^N|_P|\varphi|+\rho(N)t^{-1/2}|\varphi|_{H^{s-2}(\Omega)}$,

(b) $|T(t;q^0)\varphi - T^N(t;q^N)P^N\varphi|_{H^1(\Omega)} \leqq Ct^{-1/2}|q^0-q^N|_P|\varphi| + \rho(N)t^{-1}|\varphi|_{H^{s-2}(\Omega)}$ *for* $\varphi \in \operatorname{dom} A^{(s-2)/1}$ *if* s *is even, and for* $\varphi \in \operatorname{dom} A^{((s-1)/2)-1}$ *with* $A^{((s-1)/2)-1}\varphi \in H_0^1(\Omega)$ *if* s *is odd,*

(c) $|T(t;q^0)\varphi - T^N(t;q^N)P^N\varphi| \leqq C|q^0-q^N|_P|\varphi| + \rho(N)|\varphi|_{H^{s-2}(\Omega)}$,

(d) $|T(t;q^0)\varphi - T^N(t;q^N)P^N\varphi|_{H^1(\Omega)} \leqq C|q^0-q^N|_P|\varphi|_{H^1(\Omega)} + \rho(N)t^{-1/2}|\varphi|_{H^{s-1}(\Omega)}$ *for* $\varphi \in \operatorname{dom} A^{(s/2)-1}$ *with* $A^{(s/2)-1}\varphi \in H_0^1(\Omega)$ *if* s *is even and for* $\varphi \in \operatorname{dom}(A^{(s-1)/2})$ *if* s *is odd.*

The proof of this theorem is an immediate consequence of previous results and we will not include it. Next we turn to estimates that hold for all $\varphi \in H^0(\Omega)$.

THEOREM 2.3. *If*

(a) (H1), (H2) *and* (H5, 2) *hold and* $q^N \to q^0$ *in* Q

or if

(b) (H1), (H2), (H4, $s-1$), (H5, s) *hold, if* $\partial\Omega$ *is of class* C^2 *and* $q^N \to q^0$ *in* $C^{s-1}(\bar{\Omega};\mathbb{R}^{n\times n}) \times C^{s-1}(\bar{\Omega};\mathbb{R}^n) \times C^{s-2}(\bar{\Omega};\mathbb{R}) \times H^0(\Omega)$, $s \geqq 2$,

then

$$T^N(t;q^N)P^N\varphi^N \to T(t,q^0)\varphi \quad \text{in } H^0(\Omega)$$

uniformly in compact t-intervals.

Proof. As in the proof of Theorem 2.1 we have $|T^N(t;q)P^N\varphi| \leqq |\varphi|$, for all $q \in Q$, $\varphi \in H^0(\Omega)$. From density of $\bigcap_{j=1}^{\infty} \operatorname{dom} A^j$ in $H^0(\Omega)$ [22], and Theorem 2.1(c), respectively Theorem 2.2(c), it follows that $T^N(t;q)P^N\varphi^0 \to T(t;q^0)\varphi^0$ in $H^0(\Omega)$ uniformly in compact t-intervals. Using $\|T^N(t;q)P^N\| \leqq 1$ once again, an additional triangle inequality implies the result.

We now prepare some technicalities for a different technique of convergence proof given by the Trotter-Kato theorem.

LEMMA 2.5. *Let* (H1) *and* (H2) *hold.*

(a) *The restriction of the family of operators* $T(t;q)$ *to* $H_0^1(\Omega)$ *is a* C_0*-semigroup with* $\|T(t;q)\|_{H^1(\Omega)} \leqq M$ *for some* M *independent of* $q \in Q$ *and* $t \geqq 0$.

(b) *If moreover* $H^N \subset H_0^1(\Omega) \cap H^2(\Omega)$, *then*

$$\|T^N(t;q)\|_{H_0^1(\Omega)} \leqq e^{Ct}$$

uniformly in N, $t \geqq 0$ *and* $q \in Q$.

THEOREM 2.4. *If* (H1), (H2) *and* (H5, 2) *hold and if* $q^N \to q^0$ *in* $P \times H_0^1(\Omega)$, $H^N \subset H_0^1(\Omega) \cap H^2(\Omega)$ *and* $P^N \to 1$ *strongly in* $H_0^1(\Omega)$, *then*

$$T^N(t;q^N)P^N\varphi \to T(t;q^0)\varphi^0 \quad \text{in } H^1(\Omega)$$

for every $\varphi \in H_0^1(\Omega)$.

Proof of Theorem 2.4. The proof will be given by employing the Trotter-Kato theorem as stated in [30], for instance. In view of Lemma 2.5 it suffices to show consistency. For $v \in H_0^1$ we have

$$\begin{aligned}
&|P^N(I-A(q^0))^{-1}v - (I-A(q^N))P^Nv|_{H^1(\Omega)} \\
&\quad \leqq |(P^N-I)(I-A(q^0))^{-1}v|_{H^1(\Omega)} \\
&\qquad + |(I-A(q^0))^{-1}v - (I-A(q^N))^{-1}v|_{H^1(\Omega)} \\
&\qquad + |(I-A(q^N))^{-1}v - (I-A^N(q^N))^{-1}P^Nv|_{H^1(\Omega)}.
\end{aligned}$$

The first term converges to 0 by assumption; the second and third terms go to 0 as $N \to \infty$ by Proposition 2.1(b) and Proposition 2.2(b). Consequently $\lim |P^NT(t;q^0)\varphi^0 - T^N(t;q^N)P^N\varphi^0|_{H^1(\Omega)} = 0$, uniformly in t as t varies in compact intervals. Another application of the assumption that $P^N \to I$ in $H_0^1(\Omega)$ implies $T^N(t;q^N)P^N\varphi^0 \to$

$T(t; q^0)\varphi^0$ in $H_0^1(\Omega)$. By the uniform boundedness principle $\|P^N\|_{H_0^1(\Omega,\mathbb{R})}$ is uniformly bounded in N, and consequently, by Lemma 2.5, $\|T^N(t; q)P^N\|_{H_0^1(\Omega,\mathbb{R})}$ is bounded uniformly in N and as t ranges in compact subsets of $[0, \infty)$. An additional triangle inequality and the fact that $\varphi^N \to \varphi^0$ in $H_0^1(\Omega)$ imply the final convergence claim.

The assumptions (H5, s) are comparatively easy to satisfy for certain subspaces of spline functions, see [11] for linear splines and the estimates in [26] for higher order splines. The additional assumption that $P^N \to I$ strongly in $H_0^1(\Omega)$ appears to be more technical to verify. We discuss these assumptions for several examples.

Example 2.3. We consider the one-dimensional case in which the state space is $H^0 = H^0(0, 1)$. It is well-known that $\{e_i\}_{i=1}^{\infty}$ with $e_i = \sqrt{2}\sin(i\pi x)$ is a complete orthonormal set in H^0. Let us take subspaces H^N as $H^N = \operatorname{span}\{e_i\}_{i=1}^N$. We will show that (H5, 2) with $\hat{\varphi}^N = P^N\varphi$ is satisfied and that $P^N \to I$ strongly in $H_0^1(0, 1)$. Here again $P^N: H^0 \to H^N$ stands for the orthogonal projection given by $P^N\varphi = \sum_{i=1}^N \alpha_i e_i$ where $\alpha_i = (\varphi, e_i)$ with $(\cdot, \cdot)$ the inner product in H^0.

Let $\varphi \in H^2(0, 1) \cap H_0^1(0, 1)$ and let D denote the differentiation operator. Then

$$
\begin{aligned}
|P^N\varphi - \varphi|^2 &= \left| \sum_{i=N+1}^{\infty} (\varphi, e_i)e_i \right|^2 \\
&= \sum_{i=N+1}^{\infty} |(\varphi, e_i)|^2 \qquad (2.12) \\
&= \sum_{i=N+1}^{\infty} \frac{1}{\pi^2 i^2} |(D\varphi, f_i)|^2 \leqq \frac{1}{N^2\pi^2} |D\varphi|^2,
\end{aligned}
$$

where in the last step we used Parseval's formula and put $f_i(x) = \sqrt{2}\cos(i\pi x)$. Similarly, we have

$$
\begin{aligned}
|P^N\varphi - \varphi|^2_{H^1(0,1)} &= \left| \sum_{i=N+1}^{\infty} (\varphi, e_i)e_i \right|^2_{H^1(0,1)} \\
&\leqq \frac{1}{N^2\pi^2} |D\varphi|^2 + \left| \sum_{i=N+1}^{\infty} i\pi(\varphi, e_i) f_i \right|^2 \\
&= \frac{1}{N^2\pi^2} |D\varphi|^2 + \sum_{i=N+1}^{\infty} |i\pi(\varphi, e_i)|^2 \qquad (2.13) \\
&= \frac{1}{N^2\pi^2} |D\varphi|^2 + \sum_{i=N+1}^{\infty} \left| \frac{1}{i\pi} (D^2\varphi, e_i)^2 \right| \\
&\leqq \frac{1}{N^2\pi^2} (|D\varphi|^2 + |D^2\varphi|^2).
\end{aligned}
$$

From (2.12) and (2.13) it follows that (H5, 2) holds for this choice of subspaces with $\rho(N) = N^{-2}$. Similar estimates can be given to show that (H5, s) holds for $s > 2$. Finally, we point out that the estimates leading to (2.13) also imply that $P^N\varphi \to \varphi$ in $H^1(0, 1)$ for $\varphi \in H_0^1(0, 1)$.

Example 2.4. Let $\Omega \subset \mathbb{R}^3$ be a parallelepiped and let H^N be the set of all functions in $H_0^1(\Omega)$ which are linear with respect to some triangulation of Ω, with the diameter of the triangles bounded by $1/N$. For $\varphi \in H^2(\Omega)$ let $I^N\varphi$ denote the function in H^N which coincides with φ in the nodal points of the triangulation. Note that $I^N\varphi$ is well defined since $H^2(\Omega)$ is continuously embedded in $C(\Omega)$. It is recalled in [11] that

$$
|\varphi - I^N\varphi| \leqq \frac{C}{N^2} |\varphi|_{H^2(\Omega)}
$$

and

$$|\varphi - I^N\varphi|_{H^1(\Omega)} \leqq \frac{c}{N}|\varphi|_{H^2(\Omega)}.$$

These estimates imply $|\varphi - P_1^N\varphi| \leqq (C/N)|\varphi|_{H^1(\Omega)}$ for $\varphi \in H_0^1(\Omega)$ and $|\varphi - P_1^N\varphi|_{H^1(\Omega)} \leqq (C/N)|\varphi|_{H^2(\Omega)}$ for $\varphi \in \operatorname{dom} A$ (cf. [11]). Here $P_1^N: H_0^1(\Omega) \to H^N$ denotes the projection with respect to the $H^1(\Omega)$-norm. In particular (H5, 2) is satisfied with $\hat{\varphi}^N = P_1^N\varphi$. If in addition $|\varphi^N|_{H^1(\Omega)} \leqq (C_1/N)|\varphi^N|_{H^0(\Omega)}$ for some C_1 independent of $\varphi^N \in H^N$ (inverse assumption), then it can easily be shown that $P^N \to I$ strongly in $H^1(\Omega)$.

Example 2.5. We consider another one-dimensional choice of subspaces. Let Δ^N be the partition of $[0, 1]$ given by $\{i/N\}_{i=0}^N$ and put

$$H^N = H_3^N(\Delta^N) = \{p \in C^1(0, 1): p \text{ is a cubic polynomial on each subinterval } [i/N, i+1/N], \text{ and } p(0) = p(1) = 0\}.$$

Alternatively, $H_3^N(\Delta^N)$ is the space of cubic Hermite splines modified so that $H^N(\Delta^N) \subset H_0^1(0, 1) \cap H^2(0, 1)$, see ([26, p. 24]). For $\varphi \in C(0, 1)$ we let $I_H\varphi$ denote the interpolating spline function in $H_3^N(\Delta^N)$, so that $I_H\varphi(t_i^N) = \varphi(t_i^N)$ and $DI_H\varphi(t_i^N) = D\varphi(t_i^N)$, for $0 \leqq i \leqq N$. Thus, we have

$$|\varphi - I_H\varphi|_{L^2(0,1)} \leqq \frac{C}{N}|D\varphi|_{L^2(0,1)} \quad \text{for } \varphi \in H^1(0, 1),$$

and

$$|\varphi - I_H\varphi|_{H^1(0,1)} \leqq \frac{C}{N}|D^2\varphi|_{L^2(0,1)} \quad \text{for } \varphi \in H^2(0, 1),$$

see ([26, p. 40]). In particular, (H5, 2) is satisfied in this case with $\hat{\varphi}^N = I_H\varphi$. Next we show that $P^N \to I$ strongly in $H_0^1(\Omega)$, where P^N is the projection with respect to the $H^0(0, 1)$-norm. Let $\mathcal{Q}^N$ denote the quasi-interpolation operator from $C(0, 1) \to S_3(\Delta^N)$, the space of cubic B-splines with respect to the partition Δ^N, as defined in [27, pp. 108, 136]. In [27, p. 230] it is proved that

$$|D\mathcal{Q}^N\varphi|_{L^2(0,1)} \leqq C|\varphi|_{H^1(0,1)} \quad \text{for } \varphi \in H^1(0, 1).$$

Therefore, it follows that

$$\begin{aligned}|P^N\varphi|_{H^1(0,1)} &\leqq |P^N\varphi - \mathcal{Q}^N\varphi|_{H^1(0,1)} + |\mathcal{Q}^N\varphi - \varphi|_{H^1(0,1)} + |\varphi|_{H^1(0,1)} \\ &\leqq CN|P^N\varphi - \mathcal{Q}^N\varphi|_{L^2(0,1)} + C|\varphi|_{H^1(0,1)},\end{aligned}$$

where in the first estimate we have used Schmidt's inequality as stated in [26]. Further, $|P^N\varphi - \mathcal{Q}^N\varphi|_{L^2(0,1)} \leqq |P^N\varphi - \varphi|_{L^2(0,1)} + |\mathcal{Q}^N\varphi - \varphi|_{L^2(0,1)} \leqq (C/N)|\varphi|_{H^1(0,1)}$, follows from [27, Thm. 6.25], and [26, p. 40]. We have therefore established that $\|P^N\|_{H_0^1(\Omega)}$ is uniformly bounded in N. Note that $|P^N\varphi - \varphi|_{H^1(0,1)} \leqq (C/N)|\varphi|_{H^2(0,1)}$ from estimates on $\varphi - I_H\varphi$ given above and another application of the Schmidt inequality. These estimates together with a density argument imply that $P^N \to I$ strongly in $H_0^1(0, 1)$.

THEOREM 2.5. *Let* (H1) *and* (H2) *hold, and assume* $H^N \subset W_0^{1,\infty}(\Omega) \cap C(\bar{\Omega})$ *for* $N = 1, 2, \cdots$. *Then* $(\mathcal{P}^N)$ *with* $\tilde{Q}$ *replaced by* Q *has a solution* $\bar{q}^N \in Q$ *for any of the fit-to-data criteria* J_i, $i = 1, \cdots, 4$. *Recall that in this case* $J_i^N(q) = |C_iu^N(\cdot, \cdot; q) - \hat{z}|_{Z_i}^2$ *with* u^N *a solution of* (2.10). *If in addition* (H3) (*respectively* (H3*)) *holds, then* $(\mathcal{P}^N)$ *has a solution* $\bar{q}^N \in \tilde{Q}$.

Proof. The solution of (2.10) can be expressed as $u^N(t, x; q) = \sum_{i=1}^{k^N} \alpha_j^N(t; q)B_j^N(x)$ with $\alpha^N(t; q) = \operatorname{col}(\alpha_1^N(t; q), \cdots, \alpha_{k^N}^N(t; q)) \in \mathbb{R}^{k^N}$. It follows

from (2.10) that $\alpha^N(t; q)$ is the unique solution of

$$R^N \frac{d}{dt}\alpha^N(t; q) = G^N(q)\alpha^N(t; q), \qquad R^N\alpha^N(0; q) = h^N(q),$$

where R^N and G^N are real $k^N \times k^N$ matrices with elements

$$(R^N)_{ij} = (B_i^N, B_j^N)_{H^0(\Omega)}, \qquad (G^N(q))_{ij} = -B(q)\langle B_i^N, B_j^N\rangle$$

and $h^N(q)$ is a k^N-vector with $(h^N(q))_i = (\varphi, B_i^N)_{H^0(\Omega)}$. The map $q \to G^N(q)$ is continuous from the weak topology of Q to $\mathbb{R}^{k^N \times k^N}$. Similarly, $q \to h^N(q)$ is weakly continuous. Consequently, if $q^l \to q$ weakly in Q as $l \to \infty$, then $\alpha^N(t; q^l) \to \alpha^N(t; q)$ uniformly in compact t-intervals. Since $H^N \subset C(\bar{\Omega}, \mathbb{R})$, $\alpha^N(t; q^l) \to \alpha^N(t; q)$ implies $u^N(\cdot, \cdot; q^l) \to u^N(\cdot, \cdot; q)$ in $C(0, T; C(\bar{\Omega}, \mathbb{R}))$ for any $T > 0$. The special form of the fit-to-data criterion, together with (H2) and the above discussion imply the existence of a solution $\bar{q}^N \in Q$. The remaining assertions are simple to verify.

Remark 2.3. The smoothness requirement $H^N \subset W_0^{1,\infty}(\Omega)$ rather than $H^N \subset H_0^1(\Omega)$ was not used before Theorem 2.5 and it can be relaxed if the coefficients are more regular. For example if $((a_{ij}), (b_i), c) \in W^{1,\infty}(\Omega; \mathbb{R}^{n\times n}) \times L^\infty(\Omega; \mathbb{R}^n) \times L^\infty(\Omega; \mathbb{R})$ then $H^N \subset H_0^1(\Omega)$ is sufficient for all our results.

Theorem 2.5 guarantees existence of solutions $\bar{q}^N \in \tilde{Q}$ of $(\mathcal{P}^N)$. If (H3) (resp. (H3*)) holds, then there exists a subsequence of $\bar{q}^N$, again denoted by $\bar{q}^N$, and $q^* \in Q$ such that $\lim_N \bar{q}^N = q^*$ in Q (resp. $\lim_N \bar{q}^N = q^*$ in $P \times H_0^1$).

Here we choose to present the final parameter estimation results for the case that the parameters $\bar{q}^N$ converge strongly. Similar but technically different results are obtained if one builds upon a weakly convergent sequence of optimal parameters $\bar{q}^N$ of $(\mathcal{P}^N)$ (compare Proposition 2.1). In the following theorems we establish PEC of schemes $(H^N, A^N(q), C_i)$ where $A^N(q)$ is defined in (2.9)-(2.10) under various conditions on H^N.

THEOREM 2.6. *Let* (H1), (H2), (H3) *and* (H5, 2) *hold. Then*

(a) $(H^N, A^N(q), \mathcal{C}_3)$ *is* PEC,

(b) $(H^N, A^N(q), \mathcal{C}_4)$ *with* $\mu(t) = t^\varepsilon$ *is* PEC, *if* $\varepsilon > 1$,

(c) *if in addition* (H3*) *holds, then*

$$(H^N, A^N(q), \mathcal{C}_4) \text{ with } \mu(t) = t^\varepsilon \text{ is PEC if } \varepsilon > 0,$$

(d) *if* Ω *is one-dimensional, then* $(H^N, A^N(q), \mathcal{C}_1)$ *is* PEC,

(e) *if* Ω *is one-dimensional, then* $(H^N, A^N(q), \mathcal{C}_2)$ *with* $\mu(t) = t^\varepsilon$ *is* PEC *if* $\varepsilon > 2$,

(f) *if* Ω *is one-dimensional and* (H3*) *holds, then* $(H^N, A^N(q), \mathcal{C}_2)$ *with* $\mu(t) = t^\varepsilon$ *is* PEC, *if* $\varepsilon > 1$.

Proof. We only verify (a) of the theorem and point out the changes necessary to justify the remaining assertions. By Theorem 2.5 there exists a solution $\bar{q}^N$ of $(\mathcal{P}^N)$, where $(\mathcal{P}^N)$ is considered with observation $\mathcal{C}_3$, so that

$$J_3^N(\bar{q}^N) \leqq J_3^N(q) \quad \text{for all } q \in \tilde{Q}. \tag{2.14}$$

By (H3) there exists a subsequence of $\bar{q}^N$ again denoted by $\bar{q}^N$ such that $\lim \bar{q}^N = q^*$ in Q, with $q^* \in Q$. Theorem 2.1(a) implies that $J_3^N(\bar{q}^N) \to J_3(q^*)$ and $J_3^N(q) \to J_3(q)$ for every $q \in \tilde{Q}$. Therefore, we find by (2.14) that $J_3(q^*) \leqq J_3(q)$ for all $q \in Q$ and q^* is a solution of $(\mathcal{P})$. The limits in (1.5) and (1.6) have been established already and consequently $(H^N, A^N(q), \mathcal{C}_3)$ is PEC. Similarly (b) and (c) follow from Theorem 2.1(a) and (c) respectively. Finally, (d), (e) and (f) are verified by using Theorem 2.1(b), (b) and (d) respectively and the Sobolev embedding theorem.

Remark 2.4. Under stronger smoothness assumptions on the coefficients, the boundary and the initial data, one can employ Theorem 2.2 to derive a theorem on PEC of $(H^N, A^N(q), \mathscr{C}_i)$ analogous to Theorem 2.6. In this case $\lim_N \rho(N)$ will generally converge to zero faster than for less smooth data; since again we can only use compactness (cf. (H4)) to extract a subsequence of convergent parameters, it can, however, not be guaranteed that the overall convergence will be improved. The advantages of Theorem 2.2 will be exploited in § 3.

THEOREM 2.7. *Let* (H1), (H2), (H3) *and* (H5, 2) *hold. Then* $(H^N, A^N(q), \mathscr{C}_3)$ *with* $I \in [0, T]$ *and* $(H^N, A^N(q), \mathscr{C}_4)$ *with* $\mu(t) = t$ *are* PEC.

The proof of this theorem uses Theorem 2.5 and Theorem 2.3(a) and is otherwise analogous to the proof of Theorem 2.6.

THEOREM 2.8. *Let* (H1), (H2), (H3*) *and* (H5, 2) *hold and assume that* Ω *is one-dimensional. Then* $(H^N, A^N(q), \mathscr{C}_i)$ *with* $I \subset [0, T]$ *and* $\mu(t) = t$, $i = 1, \cdots, 4$ *are* PEC. *Again this result is a direct consequence of Theorem* 2.4, *Theorem* 2.5, *and Sobolev's embedding theorem.*

3. Point observation operators in multidimensional domains. In the case of observation operators $\mathscr{C}_1$ and $\mathscr{C}_2$ we have established PEC only for one-dimensional domains in § 2. In this section we show, by means of an example, how Theorem 2.2 can be used to guarantee PEC in domains of dimension greater than one. Convergence in a higher order Sobolev space will be shown to hold and continuous embedding into $C(\bar{\Omega}, \mathbb{R})$ will be used. At this point we note that state convergence results (for fixed parameters) in $C(\bar{\Omega}, \mathbb{R})$ do not seem to be readily available for parabolic equations in the generality considered in this paper [25], [29], [30].

Let us consider (2.1) with Ω an annulus in $\mathbb{R}^2$ with inner radius 1 and outer radius 2. Then $f: [0, 1] \times [0, 1] \to \Omega$ given by

$$f(r, \vartheta) = \operatorname{col}((r+1)\cos 2\pi\vartheta, (r+1)\sin 2\pi\vartheta)$$

maps the square $R = [0, 1] \times [0, 1]$ onto Ω. Moreover, f is infinitely differentiable and, restricted to $[0, 1] \times [0, 1)$ it is injective. Let f^{-1} denote the inverse of f; f^{-1} is infinitely differentiable from Ω to R if the upper and lower side of the square are identified with each other. Further we put $F\varphi = \varphi(f)$ and $F^{-1}\varphi = \varphi(f^{-1})$, and note that $F\varphi$ and $F^{-1}\varphi$ are defined on the rectangle and on the annulus respectively.

We next choose an equidistant grid on R with grid points (r_i^N, ϑ_j^N), where $r_i^N = iN$, $\vartheta_j^N = i/N$, $i = 0, \cdots, N$; $j = 0, \cdots, N$. Let S^N be the $(N+3) \times (N+3)$ dimensional vector space of bicubic spline functions, (see e.g. [24, p. 131]) so that each $s^N \in S^N$ has the representation

$$s^N(r, \vartheta) = \sum_{i=-1}^{N+1} \sum_{j=-1}^{N+1} \beta_{ij} B_i^N(r) B_j^N(\vartheta),$$

with B_i^N the usual B-spline basis elements. Next let $\tilde{S}^N$ be the subset of S^N defined by

$$\tilde{S}^N = \{s^N \in S^N : s^N(0, \vartheta) = s^N(1, \vartheta) = 0 \text{ for } 0 \leqq \vartheta \leqq 1, \text{ and } s^N(r, 0) = s^N(r, 1), \text{ for } 0 \leqq r \leqq 1\}.$$

It is not difficult to construct an $(N+1) \times (N+1)$ dimensional basis for $\tilde{S}^N$ (compare [24, p. 209]). Note that $F^{-1}(\tilde{s}^N) \in C(\bar{\Omega}, \mathbb{R}) \cap W_0^{1,\infty}(\Omega, \mathbb{R})$ for $\tilde{s}^N \in \tilde{S}^N$ and $F^{-1}(\tilde{s}^N)|\partial\Omega = 0$.

We prepare the following theorem by first stating two lemmas, which are proved in Appendix B.

LEMMA 3.1. *Let $s \in S^N$. Then there exists a constant c independent of N and s such that*

$$|s|_{H^2(\Omega)} \leqq cN|s|_{H^1(\Omega)}.$$

We need to introduce some more notation. For a function $\varphi:[0,1]\to\mathbb{R}$ we let $I^N\varphi \in C^2(0,1;\mathbb{R})$ denote the cubic spline function with respect to the partition $t_i^N = i/N$, $i=0,\cdots,N$ and with $(I^N\varphi)(t_i^N) = \varphi(t_i^N)$ and $D(I^N\varphi)(0) = D\varphi(0)$, $D(I^N\varphi)(1) = D\varphi(1)$. Further, for $\varphi:\mathbb{R}\to\mathbb{R}$, we put $(I_r^N\varphi)(\cdot,\vartheta) = I^N\varphi(\cdot,\vartheta)$ and $(I_\vartheta^N\varphi)(r,\cdot) = I^N\varphi(r,\cdot)$.

LEMMA 3.2. (a) *Let $\varphi \in H^3(\mathbb{R})$ with $\varphi(0,\vartheta) = \varphi(1,\vartheta) = 0$, $\varphi(r,0) = \varphi(r,1)$, $0 \leqq r \leqq 1$, $0 \leqq \vartheta \leqq 1$, and let $\tilde{P}^N: H^0(\mathbb{R}) \to \tilde{S}^N$ be the canonical projection. Then*

$$|\varphi - \tilde{P}^N\varphi|_{H^i(\mathbb{R})} \leqq \frac{c}{N^{3-i}}|\varphi|_{H^3(\mathbb{R})}, \qquad i=0,1,2. \tag{3.1}$$

(b) *If $\varphi \in H^2(\mathbb{R})$, satisfying the same boundary conditions as in* (a), *then*

$$|\varphi - \tilde{P}^N\varphi|_{H^0(\mathbb{R})} \leqq \frac{c}{N^2}|\varphi|_{H^2(\mathbb{R})}. \tag{3.2}$$

Here c is a constant independent of N and φ.

COROLLARY 3.1. (a) *Let $\Phi \in H^3(\Omega) \cap H_0^1(\Omega)$. Then*

$$|\Phi - F^{-1}\tilde{P}^N F\Phi|_{H^i(\Omega)} \leqq \frac{c}{N^{3-i}}|\Phi|_{H^3(\Omega)}, \qquad i=0,1,2. \tag{3.3}$$

(b) *If $\Phi \in H^2(\Omega) \cap H_0^1(\Omega)$, then*

$$|\Phi - F^{-1}\tilde{P}^N F\Phi|_{H^0(\Omega)} \leqq \frac{c}{N^2}|\Phi|_{H^2(\Omega)}, \tag{3.4}$$

with c independent of N and Φ, but dependent on F.

LEMMA 3.3. *Let* (H1), (H2) *hold, and let $q^N \to q^0$ in P, $\varphi \in H^0(\Omega)$. Then $|T(t;q^0)\varphi - T(t;q^N)\varphi|_{H^2(\Omega)} \leqq Ct^{-1}|q^N - q^0|_P|\varphi|$.*

Proof. This convergence result is a direct consequence of Proposition 2.1(c) and the technique of proof that was employed to verify Theorem 2.1.

Having established the above estimates we can obtain the following convergence result in $H^2(\Omega)$ by some simple inequalities.

THEOREM 3.1. *Let us consider* (2.1) *with Ω an annulus with inner radius* 1 *and outer radius* 2, *and let* (H1), (H2) *and* (H4, 2) *hold. Put*

$$H^N = F^{-1}\tilde{S}^N$$

and let $q^N \to q^0$ in $C^2(\bar{\Omega};\mathbb{R}^{n\times n}) \times C^2(\bar{\Omega};\mathbb{R}^n) \times C^1(\bar{\Omega};\mathbb{R}) \times H_0^1(\Omega)$. Then

$$|T(t;q^0)\varphi^0 - T^N(t;q^N)P^N\varphi^N|_{H^2(\Omega)} \leqq \frac{C}{t}\psi\left(|q^0 - q^N|_P + \frac{c}{N}\right) + c\sqrt{t}|\varphi^N - \varphi^0|_{H^1(\Omega)}$$

where $\psi = \sup_N |\varphi^N|_{H^1(\Omega)}$.

Proof. From the construction of $\tilde{S}^N$, it follows that $H^N \subset W_0^{1,\infty}(\Omega) \cap C(\bar{\Omega})$. Theorem 2.2(b) implies

$$|T(t;q^N)\varphi^N - T^N(t;q^N)P^N\varphi^N|_{H^1(\Omega)} \leqq Ct^{-1}N^{-2}|\varphi^N|_{H^1(\Omega)}, \tag{3.5}$$

if only (H5, 3) is satisfied with $\rho(N) = C/N^2$. But this follows from Corollary 3.1 with $\hat{\Phi}^N = F^{-1}\tilde{P}^N F\Phi$. (Here $\hat{\Phi}^N$ replaces $\hat{\varphi}^N$ in the notation of (H5, 3).)

Since $\varphi^N \in H_0^1(\Omega)$ we have by Lemma 2.1(e) and (2.8*) that

$$|T(t;q)\varphi^N|_{H^3(\Omega)} = \left|\frac{1}{2\pi i}\int_{\Gamma_1} e^{\lambda t}(\lambda - A(q))^{-1}\varphi^N\, d\lambda\right|_{H^3(\Omega)} \leqq Ct^{-1}|\varphi^N|_{H^1(\Omega)}, \tag{3.6}$$

where Γ_1 is as in the proof of Theorem 2.1 and c is independent of $q \in Q$. By Lemma 3.1 and Corollary 3.1, (3.3), (3.5) and (3.6) we have

$$\begin{aligned}
&|F^{-1}\tilde{P}^N FT(t;q^N)\varphi^N - T^N(t;q^N)P^N\varphi^N|_{H^2(\Omega)}\\
&\quad \leqq CN|F^{-1}\tilde{P}^N FT(t;q^N)\varphi^N - T^N(t;q^N)P^N\varphi^N|_{H^1(\Omega)}\\
&\quad \leqq CN|F^{-1}\tilde{P}^N FT(t;q^N)\varphi^N - T(t;q^N)\varphi^N|_{H^1(\Omega)} \qquad (3.7)\\
&\qquad + CN|T(t;q^N)\varphi^N - T^N(t;q^N)P^N\varphi^N|_{H^1(\Omega)}\\
&\quad \leqq CN^{-1}|T(t;q^N)\varphi^N|_{H^3(\Omega)} + CN^{-1}t^{-1}|\varphi^N|_{H^1(\Omega)} \leqq CN^{-1}\frac{1}{t}|\varphi^N|_{H^1(\Omega)}.
\end{aligned}$$

Consequently, Lemma 3.3, (3.3), (3.6), (3.7) imply

$$\begin{aligned}
&|T(t;q^0)\varphi^N - T^N(t;q^N)P^N\varphi^N|_{H^2(\Omega)}\\
&\quad \leqq |T(t;q^0)\varphi^N - T(t;q^N)\varphi^N|_{H^2(\Omega)}\\
&\qquad + |T(t;q^N)\varphi^N - F^{-1}\tilde{P}^N FT(t;q^N)\varphi^N|_{H^2(\Omega)} \qquad (3.8)\\
&\qquad + |F^{-1}\tilde{P}^N FT(t;q^N)\varphi^N - T^N(t;q^N)P^N\varphi^N|_{H^2(\Omega)}\\
&\quad \leqq \frac{C}{t}|q^N - q^0|_P|\varphi^N| + \frac{C}{Nt}|\varphi^N|_{H^1(\Omega)} + \frac{C}{Nt}|\varphi^N|_{H^1(\Omega)}\\
&\quad = \frac{C}{t}|\varphi^N|_{H^1(\Omega)}\left(|q^0 - q^N|_P + \frac{C}{N}\right).
\end{aligned}$$

Finally, by Lemma 2.1(h), Lemma 2.3 and (2.6) we have

$$|T(t;q^0)(\varphi^0 - \varphi^N)|_{H^2(\Omega)} \leqq C\sqrt{t}|\varphi^N - \varphi^0|_{H^1(\Omega)}. \tag{3.9}$$

Estimates (3.8) and (3.9) imply the result.

COROLLARY 3.2. *Let the assumptions of Theorem* 3.1 *hold. If* $\tilde{Q} \subset Q$ *is a compact subset of* $C^2(\bar{\Omega}, \mathbb{R}^{n\times n}) \times C^2(\bar{\Omega}, \mathbb{R}^n) \times C^1(\bar{\Omega}, \mathbb{R}) \times H_0^1(\Omega)$, *then*

(a) $(H^N, A^N(q), \mathscr{C}_1)$ *is* PEC,

(b) $(H^N, A^N(q), \mathscr{C}_2)$ *with* $\mu(t) = t^\varepsilon$ *is* PEC, *if* $\varepsilon > 1$.

This is a direct consequence of the previous theorem and Sobolev's embedding theorem for dimension 2 (and 3).

Remark 3.1. The method that we employed to derive the $H^2(\Omega)$ convergence is based upon a well-known technique in finite element analysis, the essential ingredient of which is the inverse assumption (see [2, Chap. 4]). The characteristic steps in the context of the present example are given by Lemma 3.1 (inverse assumption), (3.5) for $i = 1, 2$ and (3.7), (3.8).

Remark 3.2. As in earlier results, assuming more smoothness of the initial data that are under consideration would allow one to weaken the singularity at $t = 0$ in the estimate of Theorem 3.1.

4. Discretization of the coefficients. If the coefficients in (1.2) are constant or if a priori information about them allows us to assume a certain shape of the coefficients, so that in fact only constants have to be identified, then $(\mathscr{P}^n)$ together with the

approximation (2.10) of the state is a finite dimensional problem. In case the coefficients lie in an infinite dimensional space, an additional discretization is required. We first generalize the definition of parameter estimation convergence to include *discretization* and identification of the coefficients. Let $Q^N \subset Q$ and consider the problems

$$(\mathcal{P}_V^N) \quad \underset{q \in Q^N}{\text{minimize}}\ J_i^N(q) \text{ subject to } u^N \text{ satisfying (1.4)}, \ i = 1, \cdots, 4.$$

DEFINITION 4.1. A sequence $(H^N, A^N(q), \mathcal{C}_i, Q^N)$, $i = 1, \cdots, 4$, is called a *variable coefficient parameter estimation convergent* (VAPEC) scheme if (a) there exist solutions $\bar{q}^N \in Q^N$ of $(\mathcal{P}_V^N)$, (b) there exists at least one convergent subsequence $\bar{q}^{N_k} \to q^*$ with $q^* \in \tilde{Q}$ a solution of $(\mathcal{P})$ and if (c) $u^{N_k}(t; \bar{q}^{N_k}) \to u(t; q^*)$ in $H^0(\Omega)$ for each $t \in [0, T]$ and $J_i^{N_k}(\bar{q}^N) \to J_i(q^*)$. Any other limit of a convergent subsequence must also solve $(\mathcal{P})$ and (c) must hold.

We shall make use of the following hypotheses:

(H6) There exists a sequence of finite dimensional compact sets $Q^N \subset Q$ and surjective maps $V^N: \tilde{Q}^N \to Q^N$ such that for any sequence $q^N \to q^0$ in $\tilde{Q}$, we have $|V^N q^N - q^N|_Q \to 0$.

(H6*) This is (H6) with $|V^N q^N - q^N|_Q \to 0$ replaced by

$$|V^N q^N - q^N|_{P \times H^1(\Omega)} \to 0.$$

The sequence in q^N can also be a constant sequence.

If in addition to (H6) we assume that (H1)-(H3) hold and that $H^N \subset W_0^{1,\infty}(\Omega) \cap C(\bar{\Omega})$, then there exist solutions $\bar{q}^N \in Q^N$ of $(\mathcal{P}_V^N)$ by Theorem 2.5. From (H6) we conclude that $V^N(\hat{q}^N) = \bar{q}^N$ for some $\hat{q}^N \in Q^N$. By (H3) one can extract a subsequence with $\hat{q}^{N_k} \to q^*$, $q^* \in \tilde{Q}$, and $|V(\hat{q}^{N_k}) - \hat{q}^{N_k}|_Q \to 0$. Consequently we have

$$\bar{q}^{N_k} \to q^* \quad \text{in } Q, \quad \bar{q}^{N_k} \in Q^{N_k}, \quad q^* \in \tilde{Q} \subset Q. \tag{4.1}$$

If instead of (H5), (H7) we assume (H5*), (H7*), then

$$\bar{q}^{N_k} \to q^* \quad \text{in } P \times H_0^1(\Omega), \quad \bar{q}^{N_k} \in Q^{N_k} \subset P \times H_0^1(\Omega), \quad q^* \in \tilde{Q} \subset P \times H_0^1(\Omega). \tag{4.2}$$

We are now prepared to study VAPEC of various schemes and observation operators analogous to our study of PEC in Theorems 2.6-2.8 and Theorem 3.1. Here we only present a result corresponding to Theorem 2.6.

THEOREM 4.1. *Let* (H1)-(H3), (H5, 2) *and* (H6) *hold. Then*

(a) $(H^N, A^N(q), \mathcal{C}_3, Q^N)$ *is* VAPEC,

(b) $(H^N, A^N(q), \mathcal{C}_4, Q^N)$ *with* $\mu(t) = t^\varepsilon$ *is* VAPEC *if* $\varepsilon > 1$,

(c) *if in addition* (H3*) *and* (H6*) *hold, then* $(H^N, A^N(q), \mathcal{C}_4, Q^N)$ *with* $\mu(t) = t^\varepsilon$, $\varepsilon > 0$, *is* VAPEC.

If Ω *is one-dimensional, then*

(d) $(H^N, A^N(q), \mathcal{C}_1, Q^N)$ *is* VAPEC,

(e) $(H^N, A^N(q), \mathcal{C}_2, Q^N)$ *with* $\mu(t) = t^\varepsilon$ *is* VAPEC *if* $\varepsilon > 2$,

(f) *if in addition* (H3*) *and* (H6*) *hold, then* $(H^N, A^N(q), \mathcal{C}_2, Q^N)$ *with* $\mu(t) = t^\varepsilon$, $\varepsilon > 1$, *is* VAPEC.

Proof. We only verify part (a). By definition of $\bar{q}^{N_k}$ it follows that $J^N(\bar{q}^N) \leqq J^N(q)$ for all $q \in Q^N$ and therefore (H6) implies

$$J^N(\bar{q}^N) \leqq J^N(V^N(q)) \quad \text{for all } q \in \tilde{Q}. \tag{4.3}$$

From Theorem 2.1(a) and (4.1) we have $J^N(\bar{q}^{N_k}) \to J(\bar{q})$. Next let $q \in \tilde{Q}$ be arbitrary.

By (H6) we have $V^N q \to q$ in Q and consequently by Theorem 2.1(e) we find $J^N(V^N(q)) \to J(q)$. Therefore $J(\bar{q}) \leqq J(q)$ for all $q \in \tilde{Q}$ and (a) holds. Similarly, (b), (c), (d), (e), and (f) follow from Theorem 2.1(a), (c), (b), (b), (d), respectively.

Example 4.1. Consider the case $\Omega = [0, 1]$, with (H1), (H2) holding and let $\tilde{Q} = B \cap Q$, where B is a closed bounded subset of $H^2(0, 1; \mathbb{R}^{n\times n}) \times H^1(0, 1; \mathbb{R}^n) \times H^1(0, 1; \mathbb{R}) \times H^1(0, 1; \mathbb{R})$. Denote by S_1^N the set of piecewise linear spline functions in Q with respect to the grid $t_i^N = i/N$, $i = 0, \cdots, N$ and let V^N stand for the spline interpolation operator from $\tilde{Q}$ to S_1^N. Clearly $\tilde{Q}$ is compact in $C^1(0, 1; \mathbb{R}^{n\times n}) \times C(0, 1; \mathbb{R}^n) \times C(0, 1; \mathbb{R}) \times C(0, 1; \mathbb{R})$ and in Q. Moreover $V^N\tilde{Q}$ is compact in $C(0, 1; \mathbb{R}^{n\times n}) \times C(0, 1; \mathbb{R}^n) \times C(0, 1; \mathbb{R}) \times C(0, 1; \mathbb{R})$ (and in Q). Finally, if $q^N \to q$ in $\tilde{Q}$, then

$$|V^N q^N - q^N|_Q \leqq \frac{C}{N} |q^N|_{\tilde{Q}},$$

so that (H6) is satisfied. Note, that $Q^N \not\subset \tilde{Q}$ in this example.

5. Conclusions. In this paper we study Galerkin approximations of the special optimization problem that arises in the study of parameter estimation problems associated with parabolic differential equations. We introduce the notion of parameter estimation convergence (PEC), which requires that the optimal parameters of the approximating problems converge to a solution of the original optimization problem and that the associated trajectories and values of the fit-to-data criteria converge. The novelty of this work as compared to [3]–[6] is that we treat the multidimensional case and that we allow for point observations, as opposed to distributed observations only. The convergence analysis is based upon the representation of the solution of the original as well as the approximating equations by contour integrals and, alternatively, upon the Trotter–Kato theorem from linear semigroup theory. The parabolic nature of the equation, see Lemma 2.1, 2.2 and 2.4, is used strongly. Once the desired parameter dependent convergence results are obtained in $H^0(\Omega)$ and $H^1(\Omega)$ (see Theorems 2.1–2.4). PEC is readily shown, if the fit-to-data criterion is continuous from $H^0(\Omega)$ or $H^1(\Omega)$ to R. In particular, point observations in dimension one are included by employing Sobolev's embedding theorem. To obtain point convergence in dimension higher than one, we make use of the fact that the contour integral method that we just mentioned, provides us with a rate of convergence. Utilization of this rate together with the inverse assumption which is common for the use of finite elements, allows us to demonstrate convergence (uniform in q) of the Galerkin scheme in a finer topology than $H^0(\Omega)$ or $H^1(\Omega)$, as for example $H^2(\Omega)$. Then again point observations can be handled by Sobolev's embedding theorem in dimensions two and three. The simultaneous approximation of the state of the equation and of an unknown variable coefficient was first discussed in [3] and Theorem 4.1 is a generalization of this result within our framework. Although several research groups have carried out computations for parameter estimation problems, see the survey articles ([16], [23] e.g.) and the work of Chavent ([8], and the references given there), a general study of the theoretical convergence questions has, to our knowledge, only been started in [3]–[6], and the present paper is a continuation of these efforts. For a recent survey article we also refer to M. P. Polis, *The distributed system parameter identification problem: a survey of recent results*, 3rd IFAC Symposium on Control of Distributed Parameter Systems, S.P. 45–S.P. 58, Toulouse, France, June 29–July 2, 1982.

Appendix A. Here we give the proofs of the lemmas and propositions of § 2.

Proof of Lemma 2.1. Let $u \in \operatorname{dom} A$. Then by (2.3) and (2.5), we have

$$\begin{aligned}|\operatorname{Im}(A(q)u, u)| &\leqq |(A(q)u, u)| \leqq C_1 |u|^2_{H^1(\Omega)} \\ &\leqq -\frac{C_1}{C_2} \operatorname{Re}(A(q)u, u) \\ &\leqq -\frac{2C_1}{C_2} \operatorname{Re}(A(q)u, u) - C_1 |u|^2_{H^1(\Omega)} \\ &= -\frac{2C_1}{C_2}\left(\operatorname{Re}(A(q)u, u) + \frac{C_2}{2}|u|^2_{H^1(\Omega)}\right).\end{aligned}$$

Since by (2.5) the term in the last parenthesis is nonpositive, this estimate implies (a) and (b). By the definition of $A(q)$ from $-B(q)$ we have (c). The theory of sectorial operators directly provides (d) and (e) (see [14, p. 280], and [13, p. 20]). For α, λ, and q in the sets specified in (f) we have for $x \in H^0(\Omega)$ and constants c and c_1

$$\begin{aligned}|(-A(q))^\alpha(\lambda - A(q))^{-1}x| &\leqq c|A(q)(\lambda - A(q))^{-1}|^\alpha |(\lambda - A(q))^{-1}x|^{1-\alpha} \\ &\leqq C_1 \left|\frac{1}{\lambda + C_2/2}\right|^{1-\alpha} |-x + \lambda(\lambda - A(q))^{-1}x|^\alpha\end{aligned}$$

(cf. [13, p. 19, 26]).

An elementary calculation shows that there exists a constant $\tilde{c}$ such that

$$\left|\frac{1}{\lambda + C_2/2}\right| \leqq \frac{\tilde{c}}{|\lambda| + 1} \quad \text{for all } \lambda \in S_L^C\left(\arctan\left(\frac{2C_1}{C_2}\right), 0\right).$$

Therefore $|(-A(q))^\alpha(\lambda - A(q))^{-1}x| \leqq C_1 \tilde{c}^{1-\alpha}(1/(|\lambda| + 1))^{1-\alpha}(1 + \tilde{c})^\alpha$, which implies (f). Finally, (g) is a direct consequence of the Hille-Yosida theorem ([22, p. 23]), and (h) follows from ([13, p. 26]).

Proof of Lemma 2.2. Although the technique of the proof is quite standard (see [11]), we present the details of this important technical lemma. Set $\mu(u) = (C_2/2)|u|^2_{H^1(\Omega)}|u|^{-2}$ and $\xi(u) = B(q)\langle u, u\rangle |u|^{-2}$ for $u \neq 0$. We have suppressed the q-dependence in μ and ξ. Lemma 2.1(b) implies that $\xi(u)$ belongs to $S_R(\theta_0, \mu(u))$ where $\theta_0 = \arctan(2C_1/C_2)$. In fact

$$\begin{aligned}|\operatorname{Im} \xi(u)| = \frac{|\operatorname{Im}(B(q)\langle u, u\rangle)|}{|u|^2} &\leqq \frac{2C_1}{C_2}\frac{1}{|u|^2}\left(\operatorname{Re}(B(q)\langle u, u\rangle) - \frac{C_2}{2}|u|^2_{H^1(\Omega)}\right) \\ &= \frac{2C_1}{C_2}\left[\frac{\operatorname{Re}(B(q)\langle u, u\rangle)}{|u|^2} - \frac{C_2}{2}\frac{|u|^2_{H^1(\Omega)}}{|u|^2}\right] \\ &= \frac{2C_1}{C_2}[\operatorname{Re} \xi(u) - \mu(u)].\end{aligned}$$

Consequently, $-\xi(u) \in S_L(\theta_0, -\mu(u))$. From geometric considerations one concludes that

$$\operatorname{dist}(\lambda, S_L(\theta_0, -\mu(u))) \geqq |\lambda| \sin(\theta_1 - \theta_0) + \mu(u) \sin \theta_0$$

for $\lambda \in S_L^C(\theta_1, 0)$. For $\lambda \in S_L(\theta_1, 0)$ and $u \in H_0^1(\Omega)$ it follows that

$$\begin{aligned}|\lambda|u|^2 + B(q)\langle u, u\rangle| &= |\lambda|u|^2 + \xi(u)|u|^2| = |u|^2|\lambda + \xi(u)| \\ &\geqq |u|^2 \operatorname{dist}(\lambda, S_L(\theta_0, -\mu(u))) \\ &\geqq |u|^2\left(|\lambda| \sin(\theta_1 - \theta_0) + \frac{C_2}{2}\frac{|u|^2_{H^1(\Omega)}}{|u|^2} \sin\theta_0\right) \\ &\geqq C^{-1}(|\lambda||u|^2 + |u|^2_{H^1(\Omega)})\end{aligned}$$

where $C^{-1} = \min(\sin(\theta_1 - \theta_0), (C_2/2)\sin\theta_0)$. This implies the claim.

Proof of Lemma 2.3. Let $\tilde{A}u = \Delta u - \gamma u$ on $H_0^1(\Omega) \cap H^2(\Omega)$ with γ sufficiently large so that (H1) and (H2) are satisfied for $\tilde{A}$. Then $\operatorname{dom}(-\tilde{A})^{1/2} = H_0^1(\Omega)$ and by [20], we have $\operatorname{dom}((-A(q))^{1/2}) = (\operatorname{dom}(A(q)), H^0(\Omega))_{1/2}$, where $\operatorname{dom}(A(q))$ is endowed with the graph norm. Here $(x, y)_{1/2}$ denotes an interpolation space as introduced e.g. in [7], [19]-[21]. By (2.6) the identity map J is a homeomorphism (uniform in q) between $H_0^1(\Omega) \cap H^2(\Omega)$ with its usual topology and $\operatorname{dom} A(q)$. Consequently, by the interpolation property, it follows that J is a homeomorphism (uniform in q) between $H_0^1(\Omega)$ and $\operatorname{dom}(-A(q))^{1/2}$. For a statement of the interpolation theorem see for example ([7, Thm. 3.2.23]). The interpolation space representation used in [7] is different from the one in [20]. That these two representations are in fact equivalent follows from ([7, Thms. 3.42, 3.52], [20]). These facts are summarized in [9]. Since

$$\tfrac{1}{2}C_2|u|^2 \leqq \operatorname{Re}((-A(q))^{1/2}u, u) \leqq |u||(-A(q))^{1/2}u|$$

for all $u \in H_0^1(\Omega)$ (cf. [15, p. 269]), the graph norm of $(-A(q))^{1/2}$ is equivalent uniformly in q to the norm given by $|(-A(q))^{1/2}u|_{H^0(\Omega)}$ and the lemma is proved.

Proof of Remark 2.1. We verify the estimate (2.8*) for $k = 1$ and 2. Let $f \in H_0^1(\Omega)$ and determine u so that $\lambda u - Au = f$. Throughout this argument we suppress the dependence on q. We have $u = (\lambda - A)^{-1}f = A^{-1}(\lambda u - f)$, and $\lambda u - f \in H_0^1(\Omega)$. From (2.8) it follows that $|(\lambda - A)^{-1}f|_{H^3(\Omega)} = |u|_{H^3(\Omega)} = |A^{-1}(\lambda u - f)|_{H^3(\Omega)} \leqq C|\lambda u - f|_{H^1(\Omega)} \leqq C|f|_{H^1(\Omega)} + C|\lambda(\lambda - A)^{-1}f|_{H^1(\Omega)}$ with C independent of λ. Therefore by Lemma 2.3 and the formula $A(\lambda - A)^{-1} = \lambda(\lambda - A)^{-1} - I$ we have $|(\lambda - A)^{-1}f|_{H^3(\Omega)} \leqq C|f|_{H^1(\Omega)}$. Next choose $f \in \operatorname{dom} A$, and again put $u = (\lambda - A)^{-1}f = A^{-1}(\lambda u - f) \in \operatorname{dom} A^2$. Then (2.8) implies

$$|(\lambda - A)^{-1}f|_{H^4(\Omega)} = |A^{-1}(\lambda u - f)|_{H^4(\Omega)} \leqq C|f|_{H^2(\Omega)} + C|\lambda(\lambda - A)^{-1}f|_{H^2(\Omega)},$$

with C independent of λ. From (2.6) and Lemma 2.1(f) we conclude that $|(\lambda - A)^{-1}f|_{H^4(\Omega)} \leqq C|f|_{H^2(\Omega)}$. The estimate for arbitrary k follows by an induction argument.

Proof of Lemma 2.4. From Lemma 2.2 and the definition of $A^N(q)$, we deduce that (a) holds and (b) is a direct consequence of the fact that the numerical range $\{(A^N u, u)_{H^0(\Omega)}: u \in H^N, |u|_{H^0(\Omega)} = 1\}$ and consequently $\sigma(A^N(q))$ is contained in $S_L(\arctan(2C_1/C_2), -C_2/2)$. Estimates (c) and (d) are direct consequences of (a), and the constant C is in fact the same as the one in Lemma 2.2. Finally, (e) follows from (2.4).

Proof of Proposition 2.1. Let u and v be given by $\lambda u - A(q^0)u = \varphi$ and $\lambda v - A(q^N)v = \varphi$. Certainly, v depends on N and $u - v = (\lambda - A(q^N))^{-1}(A(q^0)u - A(q^N)u)$. Lemma 2.1 implies that

$$|u - v|_{H^0(\Omega)} \leqq \frac{C}{|\lambda| + 1}|A(q^0)u - A(q^N)u|.$$

By (2.6) and Lemma 2.1(f) once again, we find

$$|A(q^0)u - A(q^N)u|_{H^0(\Omega)} \leqq C|q^N - q^0|_P|\varphi|_{H^0(\Omega)}$$

and (a) follows. By Corollary 2.1 we have $|u-v|_{H^1(\Omega)} \leqq (C/\sqrt{|\lambda|+1})|A(q^0)u - A(q^N)u|$, and we proceed as in (a). Similarly (c) is a consequence of (2.6) and Lemma 2.1(f). To verify (d) we again use Corollary 2.1 to get

$$|u-v|_{H^1(\Omega)} = |(\lambda - A(q^N))^{-1}(A(q^0)u - A(q^N)u)|_{H^1(\Omega)}$$
$$\leqq \frac{C}{\sqrt{|\lambda|+1}}|A(q^0)u - A(q^N)u|_{H^0(\Omega)}.$$

Consequently, by Lemma 2.1(f)

$$|u-v|_{H^1(\Omega)} \leqq \frac{C}{\sqrt{|\lambda|+1}}|A(q^0)(\lambda - A(q^0))^{-1}\varphi||q^N - q^0|_P$$
$$\leqq \frac{C}{|\lambda|+1}|q^N - q^0|_P |A(q)^{1/2}\varphi|_{H^0(\Omega)}.$$

Lemma 2.3 now implies the estimate in (d) and the proposition is proved.

Proof of Proposition 2.2.

(a) Let $\lambda \in S_L^C(\theta_1, 0)$ and put $\eta = (\lambda - A(q))^{-1}\varphi$, $\eta^N = (\lambda - A^N(q))^{-1}P^N\varphi$. Then

$$e_q^N(\lambda) = e^N = \eta - \eta^N \in H_0^1(\Omega).$$

For any $\psi^N \in H^N$ we find

$$(\lambda e^N, \psi^N) - (A(q)\eta - A^N(q)\eta^N, \psi^N) = 0$$

and therefore

$$(\lambda e^N, \psi^N) + B(q)\langle e^N, \psi^N\rangle = 0 \tag{A.1}$$

for all $\psi^N \in H^N$. Lemma 2.2, (2.3) and (A.1) imply that

$$\begin{aligned} |\lambda||e^N|^2 + |e^N|^2_{H^1(\Omega)} &\leqq C|\lambda|e^N|^2 + B(q)\langle e^N, e^N\rangle| \\ &= C|\lambda(e^N, \eta - \hat{\eta}^N) + B(q)\langle e^N, \eta - \hat{\eta}^N\rangle| \\ &\leqq C(|\lambda||e^N||\eta - \hat{\eta}^N| + |e^N|_{H^1(\Omega)}|\eta - \hat{\eta}^N|_{H^1(\Omega)} \end{aligned} \tag{A.2}$$

where $\hat{\eta}^N$ is chosen according to (H5, s). This is the essential inequality for the further estimates. Since

$$|\eta|_{H^2(\Omega)} = |(\lambda - A(q))^{-1}\varphi|_{H^2(\Omega)} \leqq C|\varphi|_{H^0(\Omega)},$$

by (2.7) and Remark 2.1, and

$$|\eta|_{H^1(\Omega)} = |(\lambda - A(q))^{-1}\varphi|_{H^1(\Omega)} \leqq \frac{C}{\sqrt{|\lambda|+1}}|\varphi|$$

by Corollary 2.1, we have

$$\begin{aligned} (1+|\lambda|)|e^N|^2 + |e^N|^2_{H^1(\Omega)} &\leqq \rho(N)\sqrt{|\lambda|+1}|e^N||\varphi| + \rho(N)|e^N|_{H^1(\Omega)}|\varphi| \\ &\leqq \rho(N)|\varphi|(\sqrt{1+|\lambda|}\|e^N| + |e^N|_{H^1(\Omega)}). \end{aligned}$$

Therefore, $\sqrt{1+|\lambda|}\|e^N| + |e^N|_{H^1(\Omega)} \leqq \rho(N)|\varphi|$ which implies (a).

(b) From (A.2) and (H5, s) it follows that

$$|\lambda||e^N|^2 + |e^N|^2_{H^1(\Omega)} \leqq \rho(N)(|\lambda||e^N||\eta|_{H^{s-1}(\Omega)} + |e^N|_{H^1(\Omega)}|\eta|_{H^s(\Omega)}). \tag{A.3}$$

Next we estimate $|\eta|_{H^{s-1}(\Omega)}$ and $|\eta|_{H^s(\Omega)}$. By (2.8*), since $\eta = (\lambda - A(q))^{-1}\varphi \in \operatorname{dom}(A^{s/2})$ we have $|\eta|_{H^2(\Omega)} \leqq C|\varphi|_{H^{s-2}(\Omega)}$. Similarly,

$$
\begin{aligned}
|\eta|_{H^{s-1}(\Omega)} &\leqq C(|A(q)\eta|_{H^{s-3}(\Omega)} + |\eta|_{H^1(\Omega)}) \\
&\leqq C(|(\lambda - A(q))^{-1}A^{(s/2)-1}\varphi|_{H^1(\Omega)} + \sum_{i=0}^{(s/2)-2} |(\lambda - A(q))^{-1}A^i(q)\varphi|_{H^1(\Omega)}) \\
&\leqq C\frac{1}{\sqrt{|\lambda|+1}}\left|\sum_{i=0}^{(s/2)-1} A^i\varphi\right|_{H^0(\Omega)} \\
&\leqq \frac{C}{\sqrt{|\lambda|+1}}|\varphi|_{H^{s-2}(\Omega)}.
\end{aligned}
$$

These estimates give

$$
\sqrt{1+|\lambda|}\|e^N| + |e^N|_{H^1(\Omega)} \leqq \rho(N)|\varphi|_{H^{s-2}(\Omega)}
$$

which proves (b).

(c) We derive estimates on η similar to those given in (b). Again (2.8*) implies $|\eta|_{H^s(\Omega)} \leqq C|\varphi|_{H^{s-2}(\Omega)}$. Further, for $s \geqq 3$, we get by (2.8*)

$$
\begin{aligned}
|\eta|_{H^{s-1}(\Omega)} &\leqq C(|A(q)\eta|_{H^{s-3}(\Omega)} + |\eta|_{H^1(\Omega)}) \\
&\leqq C\left(|A(q)^{(s-1)/2}\eta| + \sum_{i=0}^{((s-1)/2)-1} |A(q)^i\eta|_{H^1(\Omega)}\right) \\
&\leqq C\left(|A^{1/2}(q)(\lambda - A(q))^{-1}A^{1/2}(q)A^{(s-3)/2}(q)\varphi| \right. \\
&\qquad \left. + \sum_{i=0}^{((s-1)/2)-1} |(\lambda - A(q))^{-1}A^i(q)\varphi|_{H^1(\Omega)}\right) \\
&\leqq \frac{C}{\sqrt{|\lambda|+1}}\left(|(\lambda - A(q))^{-1}A^{1/2}(q)A^{(s-3)/2}(q)\varphi| + \sum_{i=0}^{((s-1)/2)-1} |A^i(q)\varphi|\right) \\
&\leqq \frac{C}{\sqrt{|\lambda|+1}}|\varphi|_{H^{s-2}(\Omega)},
\end{aligned}
$$

where we used Lemma 2.1(f), Lemma 2.3, and the fact that $A^{1/2}(q)$ and $A(q)$ commute. Together with (A.3) these estimates imply the claim.

Proof of Proposition 2.3. Using the fact that $2ab \leqq a^2 + b^2$ we find by (A.2) that

$$
\begin{aligned}
(1+|\lambda|)|e^N|^2 + |e^N|^2_{H^1(\Omega)} &\leqq C(|\lambda||\eta - \hat{\eta}^N|^2 + |\eta - \hat{\eta}^N|^2_{H^1(\Omega)}) \\
&\leqq C\rho(N)^2(|\lambda||\eta|^2_{H^{s-1}(\Omega)} + |\eta|^2_{H^s(\Omega)}).
\end{aligned} \tag{A.4}
$$

Let s be even. Then by (2.8*)

$$
|\eta|_{H^s(\Omega)} \leqq C\left(|A(q)^{s/2}\eta| + \sum_{i=0}^{(s/2)-2} |A^i(q)\eta|_{H^1(\Omega)}\right).
$$

We use (2.8*), Lemma 2.1(f) and Corollary 2.1 in the next estimate:

$$|\eta|_{H^s(\Omega)} \leqq C\left(|A^{1/2}(q)(\lambda - A(q))^{-1}A^{1/2}(q)A^{(s/2)-1}(q)\varphi| + \sum_{i=0}^{(s/2)-1} |(\lambda - A(q))^{-1}A^i\varphi|_{H^1(\Omega)}\right)$$

$$\leqq \frac{C}{\sqrt{|\lambda|+1}}\left(|A^{1/2}(q)A^{1/2-1}(q)\varphi| + \sum_{i=0}^{(s/2)-1} |A^i(q)\varphi|\right)$$

$$\leqq \frac{C}{\sqrt{|\lambda|+1}}|\varphi|_{H^{s-1}(\Omega)}.$$

Similarly by Lemma 2.3

$$|\eta|_{H^{s-1}(\Omega)} \leqq C\left(|(\lambda - A(q))^{-1}A(q)^{(s-2)/1}\varphi|_{H^1(\Omega)} + \sum_{i=0}^{(s/2)-2} |(\lambda - A(q))^{-1}A^i(q)\varphi|_{H^1(\Omega)}\right)$$

$$\leqq C\left(\sum_{i=0}^{(s/2)-1} |(\lambda - A(q))^{-1}A^{1/2}(q)A^i(q)\varphi|_{H^0(\Omega)}\right)$$

$$\leqq \frac{C}{|\lambda|+1}|\varphi|_{H^{s-1}(\Omega)}.$$

Using the estimates in (A.2) it follows that

$$(|\lambda|+1)|e^N|^2 + |e^N|^2_{H^1(\Omega)} \leqq \rho(N)^2\left(\frac{1}{|\lambda|+1}|\varphi|^2_{H^{s-1}(\Omega)} + \frac{1}{|\lambda|+1}|\varphi|^2_{H^{s-1}(\Omega)}\right)$$

$$\leqq \rho(N)^2\frac{1}{|\lambda|+1}|\varphi|^2_{H^{s-1}(\Omega)}.$$

This estimate implies the proposition for s even. The case in which s is odd follows from analogous estimates.

Proof of Lemma 2.5. We prove this lemma using interpolation space techniques as previously used in the proof of Lemma 2.3. Recall that $H_0^1(\Omega) = (H_0^1(\Omega) \cap H^2(\Omega), H^0(\Omega))_{1/2}$. Since $T(t; q)$ generates a C_0-semigroup in $H^0(\Omega)$, the restriction of $T(t; q)$ to dom A generates a C_0-semigroup, again denoted by $T(t; q)$ with $\|T(t; q)\|_{\text{dom}(A(q))} \leqq e^{-C_2 t}$, (see [31, Thm. 3.4.1]). Therefore $\|(\lambda - A(q))^{-n}\|_{H^0(\Omega)} \leqq 1$ and $\|(\lambda - A(q))^{-n}\|_{H^2(\Omega)} \leqq 1$ for $\text{Re}\,\lambda \geqq 0$ and $n = 1, 2, \cdots$. Now define $\hat{A}(q)$ by dom $\hat{A}(q) = \{\varphi \in \text{dom } A,\ A(q)\varphi \in H_0^1(\Omega))\}$ and $\hat{A}(q)\varphi = A(q)\varphi$ for $\varphi \in \text{dom } \hat{A}(q)$. By interpolation $\|(\lambda - A(q))^{-n}\|_{H_0^1(\Omega)} \leqq M$ uniformly in q and λ with $\text{Re}\,\lambda \geqq 0$ and $n = 1, 2, \cdots$. The Hille-Yosida theorem implies that $\hat{A}(q)$ generates a semigroup $\hat{T}(t; q)$ on $H_0^1(\Omega)$ which is easily seen to coincide with $T(t; q)$ on $H_0^1(\Omega)$. Thus (a) is verified.

To prove (b) let $\tilde{B}(q)$ be given by dom $\tilde{B}(q) = H_0^1(\Omega)$ with $\tilde{B}(q)\varphi = \sum_{i=1}^n b_i(x)D^i\varphi(x)$ and set $A_1(q) = A(q) - \tilde{B}(q)$, dom $A_1(q) = H_0^1(\Omega) \cap H^2(\Omega)$. Clearly $A_1(q)$ satisfies (H1) and (H2) and (2.11) with $A(q)$ replaced by $A_1(q)$. Note that for $u \in H^N$ we get

$$((-A_1(q))^{1/2}u, (-A_1(q))^{1/2}P^N A(q)u)$$

$$= ((-A_1(q))^{1/2}u, (-A_1(q))^{1/2}P^N A_1(q)u) + (-A_1(q)u, P^N\tilde{B}(q)u)$$

$$= -|A_1(q)u|^2 + \tfrac{1}{2}|A_1(q)u|^2 + \tfrac{1}{2}|\tilde{B}(q)u|^2$$

$$\leqq \tfrac{1}{2}|\tilde{B}(q)u|^2$$

$$\leqq C|A^{1/2}(q)u|^2$$

for an appropriately defined C independent of q. This implies that $\|T^N(t;q)\|_g \leqq e^{Ct}$ where $\|\cdot\|_g$ denotes the graph norm of $(-A^{1/2}(q))$. By (2.11) we have $\|T^N(t,q)\|_{H_0^1(\Omega)} \leqq e^{Ct}$ for some C independent of $q \in Q$ and N.

Appendix B. Here we give the proofs that were omitted in § 3.

Proof of Lemma 3.1. The proof follows almost directly from the Schmidt inequality [26, p. 7]. Let $s(r,\vartheta) = \sum_{i,j} \beta_{i,j} B_i(r) B_j(\vartheta)$. Then

$$\begin{aligned}\int_0^1 \int_0^1 |D_{rr} s(r,\vartheta)|^2 \, dr\, d\vartheta &= \int_0^1 \int_0^1 \left| \sum_{i,j} \beta_{i,j} D_{rr} B_i(r) B_j(\vartheta) \right|^2 dr\, d\vartheta \\ &= \int_0^1 \sum_{k=0}^{N} \int_{(k-1)/N}^{k/N} \left| \sum_{i,j} \beta_{i,j} D_{rr} B_i(r) B_j(\vartheta) \right|^2 dr\, d\vartheta \\ &\leqq cN^2 \int_0^1 \int_0^1 \left| \sum_{i,j} \beta_{i,j} D_r B_i(r) B_j(\vartheta) \right|^2 dr\, d\vartheta \\ &\leqq cN^2 |s|^2_{H^1(\Omega)}.\end{aligned}$$

This estimate together with analogous estimates for the other second order derivatives implies the claim.

Proof of Lemma 3.2. Let us start with two identities that will frequently be used in the following estimates. Let $\tilde{P}_0^N : H^0(0,1;\mathbb{R}) \to S_0^N = \{\psi : [0,1] \to \mathbb{R},\ \psi(0) = \psi(1) = 0,$ ψ a cubic spline with respect to the partition $t_i^N = i/N,\ i = 0, \cdots, N\}$, be the orthogonal projection, and put $(\tilde{P}_r^N \varphi)(\cdot, \vartheta) = \tilde{P}_0^N \varphi(\cdot, \vartheta)$. Similarly let $\tilde{P}_p^N : H^0(0,1;\mathbb{R}) \to S_p^N = \{\psi : [0,1] \to \mathbb{R},\ \psi(0) = \psi(1),\ \psi$ a cubic spline with respect to the partition $t_i^N = i/N,$ $i = 0, \cdots, N\}$ be the orthogonal projection and put $(\tilde{P}_\vartheta^N \varphi)(r,\cdot) = (\tilde{P}_p^N \varphi)(r,\cdot)$; here the subscript p stands for periodic. Then it is quite simple to show (compare [26, p. 83]) that

$$\tilde{P}^N \varphi = \tilde{P}_r^N \tilde{P}_\vartheta^N \varphi = \tilde{P}_\vartheta^N \tilde{P}_r^N \varphi. \tag{B.1}$$

Further,

$$D_r \tilde{P}_\vartheta^N \varphi = \tilde{P}_\vartheta^N D_r \varphi \quad \text{and} \quad D_\vartheta \tilde{P}_r^N \varphi = \tilde{P}_r^N D_\vartheta \varphi. \tag{B.2}$$

Let us verify (3.1) for $i = 0$ first. We have

$$|\varphi - \tilde{P}^N \varphi| = |\varphi - \tilde{P}_r^N \tilde{P}_\vartheta^N \varphi| \leqq |\varphi - \tilde{P}_r^N \varphi| + |\tilde{P}_r^N \varphi - \tilde{P}_r^N \tilde{P}_\vartheta^N \varphi|,$$

and these two terms are estimated separately:

$$\begin{aligned}|\varphi - \tilde{P}_r^N \varphi|^2 &= \int_0^1 \int_0^1 |\varphi(r,\vartheta) - \tilde{P}_r^N \varphi(r,\vartheta)|^2 \, dr\, d\vartheta \\ &\leqq \int_0^1 \int_0^1 |\varphi(r,\vartheta) - I_r^N \varphi(r,\vartheta)|^2 \, dr\, d\vartheta \\ &\leqq CN^{-6} |\varphi|^2_{H^3(R)},\end{aligned}$$

where we used [26, Thm. 6.9]. For the second term we find

$$\begin{aligned}|\tilde{P}_r^N \varphi - \tilde{P}_r^N \tilde{P}_\vartheta^N \varphi|^2 &= \int_0^1 \int_0^1 |\tilde{P}_r^N \varphi(r,\vartheta) - \tilde{P}_r^N \tilde{P}_\vartheta^N \varphi(r,\vartheta)|^2 \, dr\, d\vartheta \\ &\leqq \int_0^1 \int_0^1 |\varphi(r,\vartheta) - \tilde{P}_\vartheta^N \varphi(r,\vartheta)|^2 \, dr\, d\vartheta \\ &\leqq CN^{-6} |\varphi|^2_{H^3(R)},\end{aligned}$$

again by [26, Thm. 6.9]. These two estimates imply (3.1) for $i=0$. Next let $i=1$. Then by (B.1), (B.2) and the Schmidt inequality [26, p. 7] we have

$$\begin{aligned}|D_r(\varphi-\tilde{P}^N\varphi)| &\leqq |D_r(\varphi-\tilde{P}^N_\vartheta\varphi)|+|D_r\tilde{P}^N_\vartheta(\varphi-\tilde{P}^N_r\varphi)|\\ &\leqq |D_r\varphi-\tilde{P}^N_\vartheta D r\varphi|+CN|\varphi-\tilde{P}^N_r\varphi|.\end{aligned}$$

By [26, Thms. 4.5 and 6.9] we get

$$|D_r(\varphi-\tilde{P}^N\varphi)|\leqq CN^{-2}|D^2_\vartheta D_r\varphi|+CN^{-2}|D^3_r\varphi|.$$

This, together with an analogous estimate for $D_\vartheta(\varphi-\tilde{P}^N\varphi)$ implies (3.1) for $i=1$. We now give estimates for the second order derivatives, again using the Schmidt inequality, (B.1) and [26, Thm. 6.9]. Further

$$\begin{aligned}|D_{rr}(\varphi-\tilde{P}^N\varphi)| &\leqq |D_{rr}(\varphi-\tilde{P}^N_r\varphi)|+|D_{rr}\tilde{P}^N_r(\varphi-\tilde{P}^N_\vartheta\varphi)|\\ &\leqq |D_{rr}(\varphi-I^N_r\varphi)|+|D_{rr}(I^N_r\varphi-\tilde{P}^N_r\varphi)|+CN^2|\varphi-\tilde{P}^N_\vartheta\varphi|\\ &\leqq CN^{-1}|D^3_r\varphi|+CN^2|\varphi-I^N_r\varphi|+CN^{-1}|D^3_\vartheta\varphi|\\ &\leqq CN^{-1}(|D^3_r\varphi|+|D^3_\vartheta\varphi|).\end{aligned}$$

Similarly, we have

$$\begin{aligned}|D_{r\vartheta}(\varphi-\tilde{P}^N\varphi)| &\leqq |D_{r\vartheta}(\varphi-\tilde{P}^N_r\varphi)|+|D_{r\vartheta}\tilde{P}^N_r(\varphi-\tilde{P}^N_\vartheta\varphi)|\\ &\leqq |D_r(D_\vartheta\varphi-\tilde{P}^N_r D_\vartheta\varphi)|+N^2|\tilde{P}^N_r(\varphi-\tilde{P}^N_\vartheta\varphi)|\\ &\leqq |D_r(D_\vartheta\varphi-I^N_r D_\vartheta\varphi)|+|D_r(I^N_r D_\vartheta\varphi-\tilde{P}^N_r D_\vartheta\varphi)|+N^2|\varphi-\tilde{P}^N_\vartheta\varphi|\\ &\leqq CN^{-1}|D^2_r D_\vartheta\varphi|+2N|I^N_r D_\vartheta\varphi-D_\vartheta\varphi|+CN^{-3}|D^3_\vartheta\varphi|\\ &\leqq CN^{-1}(|D^2_r D_\vartheta\varphi|+|D^2_r D_\vartheta\varphi|+|D^3_\vartheta\varphi|).\end{aligned}$$

These estimates together with an analogous estimate for the second partial derivative with respect to ϑ imply (3.1) for $i=2$.

The proof of (b) is very similar to that of (a). Since $\varphi\in H^2(\mathbb{R})$, $\varphi(r,\cdot)\in H^2(0,1;\mathbb{R})$ for almost every r and $\varphi(\cdot,\vartheta)\in H^2(0,1;\mathbb{R})$ for almost every ϑ. Consequently the interpolation operators I^N_r and I^N_ϑ employed in the proof of (a) are well defined for $\varphi\in H^2(0,1;\mathbb{R})$ and (3.2) is verified just like (3.1) for $i=0$ only that [26, Thm. 4.5] is used instead of [26, Thm. 6.9].

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] I. BABUSKA AND A. K. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1976, pp. 3-359.

[3] H. T. BANKS AND P. L. DANIEL, *Estimation of variable coefficients in parabolic distributed systems*, preprint.

[4] H. T. BANKS AND P. KAREIVA, *Parameter estimation techniques for transport equations with application to population dispersal and tissue bulk flow models*, J. Math. Biology, 17 (1983), pp. 253-273.

[5] H. T. BANKS, J. M. CROWLEY AND K. KUNISCH, *Cubic spline approximation techniques for parameter estimation in distributed systems*, to appear in IEEE Trans. Auto Control.

[6] H. T. BANKS AND K. KUNISCH, *An approximation theory for nonlinear partial differential equations with applications to identification and control*, this Journal, 20 (1982), pp. 815-849.

[7] P. L. BUTZER AND H. BERENS, *Semi-Groups of Operators and Approximation*, Springer, New York, 1967.

[8] G. CHAVENT, *Identification of distributed parameter systems: about the output least square method, its implementation, and identifiability*, in Proc. of the 5th Symposium on Identifiability and System Parameter Estimation, Darmstadt, Pergamon Press, 1979.

[9] G. DiBLASIO, K. KUNISCH AND E. SINESTRARI, *L^2-regularity for parabolic partial integro-differential equations with delay in the highest order derivatives*, J. Math. Anal. Appl., to appear.
[10] H. O. FATTORINI, *The Abstract Cauchy Problem*, Springer, New York, 1983.
[11] H. FUJITA AND A. MIZUTANI, *On the finite element method for parabolic equations*, I; *approximation of holomorphic semigroups*, J. Math. Soc. Japan, 28 (1976), pp. 749-771.
[12] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, Springer, New York, 1977.
[13] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Springer, New York, 1981.
[14] T. KATO, *Perturbation Theory for Linear Operators*, Springer, New York, 1966.
[15] ———, *Fractional powers of dissipative operators*, J. Math. Soc. Japan, 13 (1961), pp. 246-274.
[16] C. S. KUBRUSLY, *Distributed parameter systems, a survey*, Int. J. Control, 26 (1977), pp. 509-535.
[17] K. KUNISCH, *Identification and estimation of parameters in abstract Cauchy problems*, Banach Center Publications, 14 (1983), pp. 279-300.
[18] O. A. LADYZHENSKAYA AND N. N. URAL'TSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
[19] J. L. LIONS, *Théorèmes de trace et d'interpolation* (I), Ann. Scuola Norm. Sup. Pisa, 13 (1959), pp. 389-403.
[20] ———, *Espaces d'interpolation et domaines de puissances fractionnaires d'opérateurs*, J. Math. Soc. Japan, 14 (1962), pp. 233-241.
[21] J. L. LIONS AND E. MAGENES, *Non Homogeneous-Boundary Value Problems and Applications*, Springer, New York, 1972.
[22] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Lecture Notes No. 10, Univ. Maryland, College Park, 1974.
[23] M. P. POLIS AND R. E. GOODSON, *Parameter identification in distributed systems: A synthesizing overview*, Proc. IEEE, 64 (1976), pp. 43-61.
[24] P. M. PRENTER, *Splines and Variational Methods*, John Wiley, New York, 1975.
[25] A. H. SCHATZ, V. THOMEE AND L. B. WAHLBIN, *Maximum norm stability and error estimates in parabolic finite element equations*, Comm. Pure Appl. Math., 33 (1980), pp. 265-304.
[26] M. H. SCHULTZ, *Spline Analysis*, Prentice Hall, Englewood Cliffs, NJ, 1973.
[27] L. L. SCHUMAKER, *Spline Functions: Basic Theory*, John Wiley, New York, 1981.
[28] C. SIVES AND L. SATO, *Computer estimation of parameters in a brain fluid transport equation: A cubic spline approximation method*, LCDS Report No. M-82-6, Brown Univ., Providence, RI, 1982.
[29] T. USHIJIMA, *On the uniform convergence for the lumped mass approximation of the heat equation*, J. Fac. Science, Tokyo, 24 (1977), pp. 477-489.
[30] ———, *Approximation theory for semi-groups of linear operators and its application to approximation of wave equations*, Japan J. Math., 1 (1975), pp. 185-224.
[31] J. A. WALKER, *Dynamical Systems and Evolution Equations, Theory and Applications*, Plenum Press, New York, 1971.
[32] H. T. BANKS, *personal communication*; see also H. T. Banks, P. L. Daniel and P. Kareiva, *Estimation techniques for transport equations*, LCDS Technical Report #83-23.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 6, November 1985

004

STRUCTURE AND STABILITY OF FINITE DIMENSIONAL APPROXIMATIONS FOR FUNCTIONAL DIFFERENTIAL EQUATIONS*

DIETMAR SALAMON†

Abstract. This paper deals with the structural and stability properties of the averaging approximation scheme for linear retarded functional differential equations. Both in the discrete- and in the continuous-time case the structure of the approximating systems is shown to be analogous to the structure of the underlying retarded equation. Moreover, it is shown that the approximating systems are exponentially stable in a uniform sense if the original system is asymptotically stable.

AMS (MOS) subject classifications. 34K99, 65L07

Key words. functional differential equations, averaging approximation, structural operators, uniform stability

1. Introduction. The object of this paper is to present some new results on the averaging approximation scheme for linear retarded functional differential equations (RFDE).

The averaging approximation scheme has been invented and studied by several Soviet authors in the early sixties (see e.g. Repin [18]; further references and a detailed review can be found in Banks-Burns [2]). A general convergence proof, a stability analysis and applications to optimal control problems have been presented for the first time by Banks-Burns [1], [2]. Related discrete-time approximations have been considered by Delfour [6], Reber [17], Rosen [19]. Recently, Gibson [9] has used the averaging scheme for approximating the solution of the algebraic Riccati equation associated with a retarded system. However, there remained one open problem in the convergence proof in [9] which has not yet been resolved. This is the question whether the approximating systems are uniformly exponentially stable for sufficiently large N if the underlying RFDE is stable. In [9] this has been stated as a conjecture without proof. We show in § 4.2 that this conjecture is in fact correct.

Another motivation for the present work comes from some recent developments in the theory of retarded systems in the product space framework. One of these is the introduction of so-called structural operators for the state space description of RFDE's which have made the linear theory much more elegant and efficient (see e.g. Bernier-Manitius [3], Manitius [14], Delfour-Manitius [7]). They have led to a number of new results in the control theory of RFDE's, namely on problems like completeness of eigenfunctions, controllability, observability, and the linear quadratic optimal control problem. Another important development was an interpretation of the adjoint semigroup in terms of the underlying RFDE. Extensions to neutral systems and further references can be found in Salamon [20].

The problem has not yet been considered whether analogous results can be developed for finite dimensional approximation of RFDE's, in particular the averaging approximation scheme. In this paper we fill this gap. It is shown that the approximating systems satisfy analogous duality relations as the RFDE and certain structural matrices are introduced which play an analogous role for the approximating systems as the

* Received by the editors October 18, 1983, and in revised form May 28, 1984. This research was sponsored by the U.S. Army under contract DAAG29-80-C-0041 and by the National Science Foundation under grant MCS-8210950. The material is based upon work supported by the Forschungsschwerpunkt "Dynamische Systeme", University of Bremen, West Germany.

† Mathematics Research Center, University of Wisconsin-Madison, Madison, Wisconsin 53706.

structural operators do for the RFDE. Moreover, it is shown that these matrices actually converge to the corresponding structural operators. These results have several important consequences. For example, they lead to a uniform convergence result for the resolvent operators and they are crucial for the proofs of the stability results in § 4.2.

In the preliminary § 2 we give a brief overview over some recent results in the theory of linear retarded systems in the product space framework and describe the averaging approximation scheme. Section 3 is devoted to the study of the structure of the approximating systems which is shown to be analogous to the structure of the underlying RFDE under several aspects. A number of convergence proofs is then given in § 4.1 and two stability results are proved in § 4.2. In the appendix (§ 6) we prove two functional analytic results which are frequently needed in § 4. In particular, we give a quantitative estimate for the equivalence of L^P-stability and exponential stability for strongly continuous semigroups.

2. Linear retarded systems and averaging approximation.

2.1. Linear retarded systems. We consider the linear retarded functional differential equation

$$\dot{x}(t) = Lx_t, \qquad t \geqq 0, \tag{2.1}$$

where $x(t) \in \mathbb{R}^n$ and x_t is defined by $x_t(\tau) = x(t+\tau)$, $-h \leqq \tau \leqq 0$, $h > 0$. Correspondingly L is a bounded linear functional from $\mathscr{C} = \mathscr{C}[-h, 0; \mathbb{R}^n]$ into $\mathbb{R}^n$ given by

$$L\phi = \int_{-h}^{0} d\eta(\tau)\phi(\tau), \qquad \phi \in \mathscr{C},$$

where $\eta(\tau)$ is an $n \times n$-matrix valued function of bounded variation. Without loss of generality we can assume that η is normalized which means that $\eta(\tau) = 0$ for $\tau \geqq 0$, $\eta(\tau) = \eta(-h)$ for $\tau \leqq -h$, and $\eta(\tau)$ is left continuous for $-h < \tau < 0$. At some places we will assume that L is given by

$$L\phi = \sum_{j=0}^{q} A_j\phi(-h_j) + \int_{-h}^{0} A_{01}(\tau)\phi(\tau)\, d\tau, \qquad \phi \in \mathscr{C}, \tag{2.2}$$

where $0 = h_0 < \cdots < h_q = h$ and $A_j \in \mathbb{R}^{n\times n}$, $j = 0, \cdots, q$, as well as $A_{01}(\cdot) \in L^2[-h, 0; \mathbb{R}^{n\times n}]$. In this case $\eta: \mathbb{R} \to \mathbb{R}^{n\times n}$ is clearly given by

$$\eta(\tau) = -A_0\chi_{(-\infty,0)}(\tau) - \sum_{j=1}^{q} A_j\chi_{(-\infty,-h_j]}(\tau) - \int_{\tau}^{0} A_{01}(\sigma)\, d\sigma, \qquad \tau \in \mathbb{R},$$

where χ_I denotes the characteristic function of the interval I.

It is well known that (2.1) admits a unique solution $x(\cdot) \in L^2_{\mathrm{loc}}[-h, \infty; \mathbb{R}^n] \cap W^{1,2}_{\mathrm{loc}}[0, \infty; \mathbb{R}^n]$ for every initial condition of the form

$$x(0) = \phi^0, \qquad x(\tau) = \phi^1(\tau), \qquad -h \leqq \tau < 0, \tag{2.3}$$

where $\phi = (\phi^0, \phi^1) \in \mathbb{R}^n \times L^2[-h, 0; \mathbb{R}^n] =: M^2$. This solution depends continuously on $\phi \in M^2$. The fundamental solution of (2.1) will be denoted by $X(t) \in \mathbb{R}^{n\times n}$, $t \geqq -h$, and corresponds to the initial condition $X(0) = I$, $X(\tau) = 0$, $-h \leqq \tau < 0$. It can also be characterized by the Volterra integral equation

$$X(t) = I - \int_0^t \eta(s-t)X(s)\, ds$$

and its Laplace transform is given by $\Delta(\lambda)^{-1}$ where $\Delta(\lambda) = \lambda I - L(e^{\lambda\cdot})$, $\lambda \in \mathbb{C}$, is the characteristic matrix of (2.1).

Proofs of these facts can be found e.g. in Hale [10] or Delfour-Manitius [7].

2.2. Semigroups and structural operators. In the theory of RFDE's, as well as other types of integral and functional differential equations, there are essentially two ways of introducing the state of the system which are actually dual to each other. The state of system (2.1) in the "classical" sense is the pair $z(t)=(x(t), x_t)\in M^2$ which completely describes the past history of the solution. Its evolution determines the strongly continuous semigroup $S(t)$ of bounded linear operators on M^2 defined by

$$S(t)\phi=(x(t), x_t)\in M^2, \qquad \phi\in M^2, \quad t\geqq 0,$$

where $x(t)$, $t\geqq -h$, is the unique solution of (2.1) and (2.3). The infinitesimal generator of $S(t)$ is given by

$$\text{dom } A=\{\phi\in M^2 | \phi^1\in W^{1,2},\ \phi^0=\phi^1(0)\},$$
$$A\phi=(L\phi^1, \dot{\phi}^1),$$

where $W^{1,2}$ denotes the Sobolev space $W^{1,2}[-h, 0, \mathbb{R}^n]$. In an analogous way we may introduce the semigroup $S_T(t)\in\mathscr{L}(M^2)$, $t\geqq 0$, with infinitesimal generator A_T corresponding to the transposed RFDE

$$\dot{x}(t)=L^T x_t, \qquad t\geqq 0. \tag{2.4}$$

The duality relation between (2.1) and (2.4) can be described by means of an alternative (dual) state concept which is due to Miller [15]. It can be motivated from the fact that the solution of the RFDE (2.1) ($t>0$) can be derived from the initial function ($t\leqq 0$) in two steps. First convert the initial function ϕ^1 into a forcing term of suitable length which determines the future behaviour of the solution. Secondly determine the solution which corresponds to this forcing term. The dual state concept is obtained by regarding this forcing term as the initial state of the system rather than the solution segment. To be more precise, we rewrite (2.1) as

$$\dot{x}(t)=\int_{-t}^{0} d\eta(\tau)x(t+\tau)+f^1(-t), \qquad x(0)=f^0, \tag{2.5}$$

where the pair $f=(f^0, f^1)\in M^2$ is given by

$$f^0=\phi^0, f^1(\sigma)=\int_{-h}^{\sigma} d\eta(\tau)\phi^1(\tau-\sigma), \qquad -h\leqq\sigma\leqq 0. \tag{2.6}$$

Now the initial state of (2.5) is given by $f\in M^2$. Correspondingly the state at time $t\geqq 0$ is the pair $w(t)=(x(t), x^t)\in M^2$ where $x^t\in L^2[-h, 0; \mathbb{R}^n]$ denotes the forcing term of the shifted equation (2.5) and is given by

$$x^t(\sigma)=\int_{\sigma-t}^{\sigma} d\eta(\tau)x(t+\tau-\sigma)+f^1(\sigma-t), \qquad -h\leqq\sigma\leqq 0. \tag{2.7}$$

The evolution of this state $(x(t), x^t)\in M^2$ is described by the semigroup $S_T^*(t)$ (see e.g. Bernier-Manitius [3] or Salamon [20]).

Summarizing this situation, we have to deal with the following four semigroups:

$$\begin{matrix} S(t) & S_T(t) \\ S_T^*(t) & S^*(t). \end{matrix}$$

The semigroups on the left correspond to the RFDE (2.1) and those on the right to the transposed RFDE (2.4). On each side the upper semigroup describes the respective

equation within the "classical" state concept (solution segments) and the one below within the dual state concept (forcing terms). The diagonal relations are actually given by functional analytic duality.

The relation between the two state concepts can be described by means of so-called structural operators. These have been introduced by Bernier-Manitius [3], Manitius [14], Delfour-Manitius [7] and have turned out to be a very elegant and efficient concept in the control theory of RFDE's. The operator $F \in \mathscr{L}(M^2)$ maps every $\phi \in M^2$ into the corresponding initial state

$$F\phi = f \in M^2$$

of (2.5) which is given by (2.6). The operator $G \in \mathscr{L}(M^2)$ maps every forcing term $f \in M^2$ into the corresponding solution segment

$$Gf = (x(h), x_h) \in M^2$$

of (2.5) at time h. Thus Gf can be explicitly described as

$$[Gf]^0 = [Gf]^1(0),$$

$$[Gf]^1(\tau) = X(h+\tau)f^0 + \int_0^h X(h+\tau-s)f^1(-s)\,ds, \qquad -h \leqq \tau \leqq 0.$$

Obviously, G is bijective as an operator from M^2 into dom A and its inverse is given by

$$[G^{-1}\phi]^0 = \phi^1(-h),$$

$$[G^{-1}\phi]^1(\sigma) = \dot{\phi}^1(-\sigma-h) - \int_\sigma^0 d\eta(\tau)\phi^1(\tau-\sigma-h), \qquad -h \leqq \sigma \leqq 0,$$

for $\phi \in \text{dom } A$. A remarkable fact is that the adjoint operators F^* and G^* play the same role for the transposed equation (2.4) as the operators F and G do for the original equation (2.1). Moreover, the following result has been proved by Manitius [14] and Delfour-Manitius [7].

THEOREM 2.1.

(i) $S(h) = GF$, $S_T^*(h) = FG$.

(ii) $FS(t) = S_T^*(t)F$, $S(t)G = GS_T^*(t)$, $t \geqq 0$.

(iii) *If* $\phi \in \text{dom } A$, *then* $F\phi \in \text{dom } A_T^*$ *and* $A_T^*F\phi = FA\phi$.

(iv) *If* $f \in \text{dom } A_T^*$, *then* $GA_T^*f = AGf$.

We close this section with a concrete representation of the resolvent operator. For this sake we introduce for any $\lambda \in \mathbb{C}$ the operators $E_\lambda : \mathbb{C}^n \to M^2$ and $T_\lambda : M^2 \to M^2$ by defining

$$[E_\lambda x]^0 = x, \qquad [E_\lambda x]^1(\tau) = e^{\lambda\tau}x, \qquad x \in \mathbb{C}^n,$$

$$[T_\lambda \phi]^0 = 0, \qquad [T_\lambda \phi]^1(\tau) = \int_\tau^0 e^{\lambda(\tau-\sigma)}\phi^1(\sigma)\,d\sigma, \qquad \phi \in M^2.$$

Then the following result has been proved in Manitius [13] and Delfour-Manitius [7].

PROPOSITION 2.2. *Let* $\det \Delta(\lambda) \neq 0$. *Then*

$$(\lambda I - A)^{-1} = E_\lambda \Delta(\lambda)^{-1} E_{\bar{\lambda}}^* F + T_\lambda,$$

$$(\lambda I - A_T^*)^{-1} = FE_\lambda \Delta(\lambda)^{-1} E_{\bar{\lambda}}^* + T_{\bar{\lambda}}^*.$$

2.3. Averaging approximation. In this section we briefly describe an approximation scheme for RFDE's which has been studied by Repin [18], Banks-Burns [1], [2],

Gibson [9] and many others. To this end we introduce for every $N \in \mathbb{N}$ the linear subspace $X^N \subset M^2$ defined by

$$X^N = \left\{ \phi \in M^2 \middle| \phi^1(\tau) = z_j \in \mathbb{R}^n, -\frac{j}{n}h \leqq \tau < -\frac{j-1}{N}h, \quad j = 1, \cdots, N \right\}$$

and denote the corresponding orthogonal projection by $p^N : M^2 \to X^N$. This subspace can be identified with $\mathbb{R}^{n(N+1)}$ by means of the embedding $\iota^N : \mathbb{R}^{n(N+1)} \to M^2$ which associates with every $z = (z_0^T, \cdots, z_N^T)^T \in \mathbb{R}^{n(N+1)}$ the pair

$$[\iota^N z]^0 = z_0,$$

$$[\iota^N z]^1(\tau) = z_j, \qquad -\frac{j}{N}h \leqq \tau < -\frac{j-1}{N}h, \qquad j = 1, \cdots, N.$$

On $\mathbb{R}^{n(N+1)}$ we will always consider the induced inner product

$$\langle z, w \rangle_N = z^T Q^N w, \qquad z, w \in \mathbb{R}^{n(N+1)},$$

where

$$Q^N = \begin{bmatrix} I & 0 & \cdots & 0 \\ 0 & \frac{h}{N} I & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{h}{N} I \end{bmatrix}. \tag{2.8}$$

The corresponding vector and matrix norms will be denoted by $\| \cdot \|_N$. The adjoint operator $\pi^N = (\iota^N)^* : M^2 \to \mathbb{R}^{n(N+1)}$ is then given by

$$[\pi^N \phi]_0 = \phi^0, \qquad [\pi^N \phi]_j = \frac{N}{h} \int_{-jh/N}^{-(j-1)h/N} \phi^1(\tau)\, d\tau, \qquad j = 1, \cdots, N.$$

Obviously, the operators ι^N and π^N satisfy

$$\pi^N \iota^N = \mathrm{id}, \qquad \iota^N \pi^N = p^N. \tag{2.9}$$

On $\mathbb{R}^{n(N+1)}$ we consider the differential equation

$$\dot{z}^N(t) = A^N z^N(t), \qquad t \geqq 0, \tag{2.10}$$

where

$$A^N = (Q^N)^{-1} H^N, \qquad H^N = \begin{bmatrix} A_0^N & A_1^N & \cdots & A_N^N \\ I & -I & & \\ & \ddots & \ddots & \\ & & I & -I \end{bmatrix}. \tag{2.11}$$

and

$$A_j^N = \lim_{\tau \uparrow -jh/N} \left[\eta\left(\tau + \frac{h}{N}\right) - \eta(\tau) \right], \qquad j = 0, \cdots, N. \tag{2.12}$$

In an analogous way we define the matrix $A_T^N = (Q^N)^{-1} H_T^N$ where the A_j^N are replaced by $(A_j^N)^T$ for $j = 0, 1, \cdots, N$. Then the adjoint matrix $(A_T^N)^*$ of A_T^N with respect to

the inner product $\langle \cdot, \cdot \rangle_N$ is given by

$$(A_T^N)^* = (Q^N)^{-1}(H_T^N)^T, \qquad (H_T^N)^T = \begin{bmatrix} A_0^N & I & & \\ A_1^N & -I & \ddots & \\ \vdots & & \ddots & I \\ A_N^N & & & -I \end{bmatrix}. \tag{2.13}$$

We also consider the differential equation

$$\dot{w}^N(t) = (A_T^N)^* w^N(t), \qquad t \geqq 0, \tag{2.14}$$

on $\mathbb{R}^{n(N+1)}$. The following theorem has been proved in Banks-Burns [2] and Gibson [9].

THEOREM 2.3. *Let $L: \mathscr{C} \to \mathbb{R}^n$ be given by* (2.2). *Then the following statements hold.*

(i) *For every $\phi \in M^2$ we have $\phi = \lim_{N\to\infty} p^N \phi$.*

(ii) *There exist constants $M \geqq 1$, $\omega \geqq 0$, such that*

$$\|e^{A^N t}\|_N \leqq M e^{\omega t}, \qquad \|e^{(A_T^N)^* t}\|_N \leqq M e^{\omega t}$$

for every $t \geqq 0$ and every $N \in \mathbb{N}$.

(iii) *For all $\phi \in M^2$, $f \in M^2$*

$$S(t)\phi = \lim_{N\to\infty} \iota^N e^{A^N t} \pi^N \phi, \qquad S_T^*(t) f = \lim_{N\to\infty} \iota^N e^{(A_T^N)^* t} \pi^N f$$

and the limits are uniform on every compact interval $[0, T]$.

Full discretization. A fairly general and extensive study of full discretization methods for RFDE's can be found in Reber [17] and Rosen [19]. Since the aim of this work is to explore the special structure of the averaging approximation scheme described above, we content ourselves with the consideration of a simple one step Euler approximation for the ODE (2.10) which has also been studied by Delfour [6] and Reber [17] for time varying systems.

Replacing the derivative in (2.10) by a difference quotient with step size h/N, we get the difference equation

$$z_{k+1}^N = \left(I + \frac{h}{N} A^N\right) z_k^N, \qquad k \geqq 0, \tag{2.15}$$

in $\mathbb{R}^{n(N+1)}$. Since

$$I + \frac{h}{N} A^N = \begin{bmatrix} I + \frac{h}{N} A_0^N & \frac{h}{N} A_1^N & \cdots & \frac{h}{N} A_N^N \\ I & 0 & & \\ & \ddots & \ddots & \\ & & I & 0 \end{bmatrix}, \tag{2.16}$$

the $n(N+1)$-dimensional first order difference equation (2.15) is equivalent to the n-dimensional $(N+1)$st order difference equation

$$\frac{N}{h}(x_{k+1}^N - x_k^N) = \sum_{j=0}^{N} A_j^N x_{k-j}^N, \qquad k \geqq 0, \tag{2.17}$$

by means of the identification

$$z_k^N = \begin{pmatrix} x_k^N \\ \vdots \\ x_{k-N}^N \end{pmatrix} \in \mathbb{R}^{n(N+1)}. \tag{2.18}$$

Equation (2.17) may be interpreted as a direct application of a 1-step difference approach to the RFDE (2.1) with x_k^N approximating $x(kh/N)$. Finally, note that this simplification of the difference equation (2.15) is only possible because of the coincidence of the step size h/N for the time-discretization with the mesh size of the spatial discretization in the subspace $X^N \subset M^2$.

3. The structure of the approximating systems. It is the goal of this section to analyse in detail the structure of the approximating systems (2.10), (2.14) and (2.15) respectively (2.17). It is shown that there is a strict analogy to the structure of the underlying RFDE (2.1) as it has been described in § 2.2. In particular, there are certain structural matrices F^N and G^N playing the same role for the approximating systems as the operators F and G do for the RFDE (2.1).

3.1. The structural matrices. Starting from (2.17), we observe that there is another way of transforming this $(N+1)$st order difference equation into an equivalent first order equation. For this sake let us rewrite (2.17) as

$$\begin{aligned} &\frac{N}{h}(x_{k+1}^N - x_k^N) = \sum_{j=0}^{k} A_j^N x_{k-j}^N + f_{k+1}^N, \qquad k \geqq 0, \\ &x_0^N = f_0^N, \end{aligned} \tag{3.1}$$

where $A_j^N := 0, f_j^N := 0$ for $j > N$ and

$$\begin{aligned} &f_0^N = x_0^N, \\ &f_k^N = \sum_{j=k}^{N} A_j^N x_{k-1-j}^N, \qquad k = 1, \cdots, N. \end{aligned} \tag{3.2}$$

The forcing term

$$f^N = \begin{pmatrix} f_0^N \\ \vdots \\ f_N^N \end{pmatrix} \in \mathbb{R}^{n(N+1)}$$

may be considered as the initial state of (3.1) since it contains all the information which is needed for determining the future behavior of its solution x_k^N, $k \geqq 0$. Correspondingly the state at instant $k \in \mathbb{N}$ is given by $w_k^N \in \mathbb{R}^{n(N+1)}$ where

$$\begin{aligned} &w_{k,0}^N = x_k^N, \\ &w_{k,l}^N = \sum_{j=l}^{k+l-1} A_j^N x_{k+l-1-j}^N + f_{k+l}^N, \qquad l = 1, \cdots, N. \end{aligned} \tag{3.3}$$

Then it is easy to see that w_k^N satisfies the first order difference equation

$$w_{k+1}^N = \left(I + \frac{h}{N}(A_T^N)^*\right) w_k^N, \qquad k \geqq 0, \tag{3.4}$$

since

$$I + \frac{h}{N}(A_T^N)^* = \begin{bmatrix} I + \frac{h}{N}A_0^N & \frac{h}{N}I & & \\ A_1^N & 0 & I & \ddots \\ \vdots & & \ddots & I \\ A_N^N & & & 0 \end{bmatrix}.$$

Note that (3.4) can be regarded as a one step Euler approximation for the ODE (2.14).

We conclude that there are two state concepts for the difference equation (2.17), namely (2.18) and (3.2-3), both of which lead to a first order difference equation in $\mathbb{R}^{n(N+1)}$, namely to (2.15) and (3.4). The relation between these two state concepts can be described by certain structural matrices F^N and G^N. Before defining these matrices, we introduce the concept of a fundamental solution for equation (3.1).

DEFINITION 3.1. The *fundamental matrix* of equation (3.1) is the sequence $X_k^N \in \mathbb{R}^{n\times n}$, $k \geqq 0$, defined by

$$\frac{N}{h}(X_{k+1}^N - X_k^N) = \sum_{j=0}^{k} A_j^N X_{k-j}^N, \qquad k\in\mathbb{N}, \quad X_0^N = I. \tag{3.5}$$

Remark 3.2. (i) By induction, it is easy to see that

$$\frac{N}{h}(X_{k+1}^N - X_k^N) = \sum_{j=0}^{k} X_{k-j}^N A_j^N, \qquad k\in\mathbb{N}. \tag{3.6}$$

(ii) The solution of (3.1) is given by

$$x_k^N = X_k^N f_0^N + \frac{h}{N}\sum_{j=0}^{k-1} X_j^N f_{k-j}^N, \qquad k\geqq 0. \tag{3.7}$$

Now we introduce the matrices

$$F^N = \begin{bmatrix} I & 0 & \cdots & \cdots & 0 \\ 0 & A_1^N & & & A_N^N \\ & & \ddots & & 0 \\ & & & \ddots & \vdots \\ 0 & A_N^N & 0 & \cdots & 0 \end{bmatrix} \tag{3.8}$$

and

$$G^N = K^N Q^N, \qquad K^N = \begin{bmatrix} X_N^N & \cdots & \cdots & X_0^N \\ \vdots & & & 0 \\ \vdots & & \ddots & \vdots \\ X_0^N & 0 & \cdots & 0 \end{bmatrix}. \tag{3.9}$$

Then it is easy to see that $f^N = F^N z_0^N$ if $z_0^N \in \mathbb{R}^{n(N+1)}$ is defined by (2.18) and $f^N \in \mathbb{R}^{n(N+1)}$ is the forcing term of (3.1) defined by (3.2). Moreover, if x_k^N, $k \geqq 0$, is the solution of (3.1) and $z_N^N \in \mathbb{R}^{n(N+1)}$ is defined by (2.18), then it follows from Remark 3.2 that $z_N^N = G^N f^N$. Making use of these facts, one can easily establish the following result which is strictly analogous to Theorem 2.1.

PROPOSITION 3.3.

(i) $$\left(I + \frac{h}{N}A^N\right)^N = G^N F^N, \qquad \left(I + \frac{h}{N}(A_T^N)^*\right)^N = F^N G^N.$$

(ii) $$F^N A^N = (A_T^N)^* F^N, \qquad A^N G^N = G^N (A_T^N)^*.$$

(iii) $$F^N e^{A^N t} = e^{(A_T^N)^* t} F^N, \qquad e^{A^N t} G^N = G^N e^{(A_T^N)^* t}, \quad t \geqq 0.$$

(iv)

$$(G^N)^{-1} = \begin{bmatrix} 0 & \cdots & 0 & I \\ \vdots & \iddots & \frac{N}{h}I & -\frac{N}{h}I \\ 0 & \iddots & \iddots & 0 \\ \frac{N}{h}I & -\frac{N}{h}I & 0 \cdots 0 & \end{bmatrix} - \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ & & & A_0^N \\ \vdots & & \iddots & \\ 0 & A_0^N & \cdots & A_{N-1}^N \end{bmatrix}.$$

Proof. Let $x_k^N \in \mathbb{R}^n$, $k \geqq -N$, be a solution of (2.17), let $z_k^N \in \mathbb{R}^{n(N+1)}$, $k \geqq 0$, be defined by (2.18) and $f^N \in \mathbb{R}^{n(N+1)}$ by (3.2). Then $f^N = F^N z_0^N$ and z_k^N satisfies (2.15) for $k \geqq 0$. Furthermore x_k^N, $k \geqq 0$, satisfies (3.1) and therefore $z_N^N = G^N f^N$. This implies

$$\left(I + \frac{h}{N}A^N\right)^N z_0^N = z_N^N = G^N f^N = G^N F^N z_0^N.$$

Thus we have proved the first equation in statement (i). Now let $w_k^N \in \mathbb{R}^{n(N+1)}$, $k \geqq 0$, be defined by (3.3). Then $w_0^N = f^N$ and it follows from (3.2) that $w_k^N = F^N z_k^N$ for every $k \geqq 0$. Since w_k^N satisfies (3.4), we conclude that

$$\left(I + \frac{h}{N}(A_T^N)^*\right) F^N z_0^N = \left(I + \frac{h}{N}(A_T^N)^*\right) w_0^N = w_1^N = F^N z_1^N = F^N \left(I + \frac{h}{N}A^N\right) z_0^N.$$

This proves the first equation in statement (ii).

In order to establish the second equations in (i) and (ii), let $f^N \in \mathbb{R}^{n(N+1)}$ be given, let $x_k^N \in \mathbb{R}^n$, $k \geqq 0$, be the unique solution of (3.1) and let $z_k^N \in \mathbb{R}^{n(N+1)}$, $k \geqq N$, and $w_k^N \in \mathbb{R}^{n(N+1)}$, $k \geqq 0$, be defined by (2.18) and (3.3), respectively. Then the sequence x_{k+l}^N, $l \geqq 0$, satisfies the equation

$$\frac{N}{h}(x_{k+l+1}^N - x_{k+l}^N) = \sum_{j=0}^{l} A_j^N x_{k+l-j}^N + w_{k,l+1}^N, \qquad l \geqq 0,$$

and hence it follows from Remark 3.2 (ii) that $z_{k+N}^N = G^N w_k^N$ for $k \geqq 0$. Furthermore $w_0^N = f^N$ and $w_k^N = F^N z_k^N$ for $k \geqq N$. Since w_k^N satisfies (3.4) for $k \geqq 0$ and z_k^N satisfies (2.15) for $k \geqq N$, we conclude that

$$\left(I + \frac{h}{N}(A_T^N)^*\right)^N f^N = w_N^N = F^N z_N^N = F^N G^N f^N$$

and

$$G^N \left(I + \frac{h}{N}(A_T^N)^*\right) f^N = G^N w_1^N = z_{N+1}^N = \left(I + \frac{h}{N}A^N\right) z_N^N = \left(I + \frac{h}{N}A^N\right) G^N f^N.$$

Thus we have proved the statements (i) and (ii). Statement (iii) is an immediate consequence of (ii).

Finally, let $(G^N)^{-1}$ be defined as in (iv) and let K^N be defined by (3.9). Then it follows from (3.5) that $(G^N)^{-1} K^N = (Q^N)^{-1}$ and hence $G^N = K^N Q^N$. This proves statement (iv). □

Proposition 3.3 shows that, for any solution $z^N(t)$ of (2.10), the function $w^N(t) = F^N z^N(t)$ satisfies (2.14) and, conversely, for any solution $w^N(t)$ of (2.14), the function $z^N(t) = G^N w^N(t)$ satisfies (2.10).

3.2. Spectral theory. In this section we give a brief overview over some spectral properties of A^N and $(A_T^N)^*$ which are analogous to well-known results in the theory of RFDEs. In particular, we will see that the rational complex $n \times n$-matrix valued function

$$\Delta^N(\lambda) = \lambda I - L^N(\lambda),\ L^N(\lambda) = \sum_{j=0}^{N} A_j^N \left(\frac{N}{N+\lambda h}\right)^j, \qquad \lambda \neq -\frac{N}{h}, \tag{3.10}$$

plays precisely the same role for the approximating systems as the characteristic matrix $\Delta(\lambda)$ does for the underlying RFDE (2.1). Moreover, we introduce the matrices

$$E_\lambda^N = \begin{bmatrix} I \\ \frac{N}{N+\lambda h} I \\ \vdots \\ \left(\frac{N}{N+\lambda h}\right)^N I \end{bmatrix} \in \mathbb{C}^{n(N+1)\times n} \tag{3.11}$$

and

$$T_\lambda^N = \frac{h}{N} \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & \frac{N}{N+\lambda h} I & 0 \cdots & 0 \\ \vdots & \vdots & \ddots \quad \ddots & \vdots \\ & & \ddots & 0 \\ 0 & \left(\frac{N}{N+\lambda h}\right)^N I & \cdots & \frac{N}{N+\lambda h} I \end{bmatrix} \in \mathbb{C}^{n(N+1)\times n(N+1)} \tag{3.12}$$

for $\lambda \in \mathbb{C}$, $\lambda \neq -N/h$.

LEMMA 3.4. *Let $\lambda \in \mathbb{C}$, $\lambda \neq -N/h$, and $z, w \in \mathbb{C}^{n(N+1)}$ be given. Then $(\lambda I - A^N)z = w$ if and only if*

$$z = E_\lambda^N z_0 + T_\lambda^N w \tag{3.13}$$

and

$$\Delta^N(\lambda) z_0 = (E_\lambda^N)^T Q^N F^N w. \tag{3.14}$$

Proof. Clearly $(\lambda I - A^N)z = w$ if and only if $(\lambda Q^N - H^N)z = Q^N w$ or equivalently

$$\lambda z_0 - \sum_{j=0}^{N} A_j^N z_j = w_0, \tag{3.15}$$

$$z_j = \frac{N}{N+\lambda h}\left[\frac{h}{N} w_j + z_{j-1}\right], \qquad j = 1, \cdots, N. \tag{3.16}$$

Equation (3.16) is equivalent to

$$z_j = \left(\frac{N}{N+\lambda h}\right)^j z_0 + \frac{h}{N} \sum_{\nu=1}^{j} \left(\frac{N}{N+\lambda h}\right)^\nu w_{j+1-\nu}, \qquad j = 1, \cdots, N,$$

and hence to (3.13). If this is satisfied, then (3.15) is equivalent to

$$\begin{aligned}\Delta^N(\lambda)z_0 &= \lambda z_0 - A_0^N z_0 - \sum_{j=1}^{N} A_j^N \left[z_j - \frac{h}{N} \sum_{\nu=1}^{j} \left(\frac{N}{N+\lambda h} \right)^{\nu} w_{j+1-\nu} \right] \\ &= w_0 + \frac{h}{N} \sum_{\nu=1}^{N} \left(\frac{N}{N+\lambda h} \right)^{\nu} \sum_{j=\nu}^{N} A_j^N w_{j+1-\nu} \\ &= (E_\lambda^N)^T Q^N F^N w. \end{aligned}$$

□

Note that the above lemma is strictly analogous to a well-known result in the theory of RFDEs (see e.g. Hale [10], Delfour-Manitius [7]). It has several important consequences which are summarized in the proposition below and can be proved straightforwardly. Statement (i) can be found in Banks-Burns [2]. Statement (iv) is the analogon of Proposition 2.2.

PROPOSITION 3.5.

(i) *Let* $\lambda \in \mathbb{C}$, $\lambda \neq -N/h$; *then* $\lambda \in \sigma(A^N)$ *if and only if* $\det \Delta^N(\lambda) = 0$.

(ii) $\lambda = -N/h \in \sigma(A^N)$ *if and only if* $\det A_N^N = 0$.

(iii) $\sigma((A_T^N)^*) = \sigma(A_T^N) = \sigma(A^N)$.

(iv) *If* $\lambda \neq -N/h$ *and* $\det \Delta^N(\lambda) \neq 0$, *then*

$$(\lambda I - A^N)^{-1} = E_\lambda^N \Delta^N(\lambda)^{-1} (E_\lambda^N)^T Q^N F^N + T_\lambda^N,$$

$$(\lambda I - (A_T^N)^*)^{-1} = F^N E_\lambda^N \Delta^N(\lambda)^{-1} (E_\lambda^N)^T Q^N + (T_\lambda^N)^T.$$

Remark 3.6. A solution $x(t)$ of the RFDE (2.1) is said to be *small* if it vanishes after some finite time T (Henry [11]). If $L: \mathscr{C} \to \mathbb{R}^n$ is given by (2.2) and if $A_{01}(\cdot) \equiv 0$, then there exist nonzero small solutions of (2.1) if and only if $\det A_q = 0$ (Manitius [14]). Now note that for sufficiently large N this means that $\det A_N^N = 0$ and hence $-N/h \in \sigma(A^N)$ (Proposition 3.5 (ii)). This indicates that the generalized eigenmodes of (2.10) respectively (2.14) corresponding to the eigenvalue $\lambda = -N/h$ play the role of the small solutions in the approximating systems. Moreover, note that the solutions of the difference equation (2.15) starting with generalized eigenvectors of A^N corresponding to $\lambda = -N/h$ are precisely those solutions which vanish after a finite time.

4. Convergence and stability. Having introduced a number of operators for the approximating systems which are analogous to well-known operators in the theory of RFDEs, we may pose the question, if—and in what sense—these operators converge. This problem will be considered in the next section.

4.1. Convergence. We begin with some preliminary facts.

Remark 4.1. (i) It is easy to see that the function $\eta^N: \mathbb{R} \to \mathbb{R}^{n \times n}$ defined by

$$\eta^N(\tau) = \lim_{\sigma \uparrow -kh/N} \eta(\sigma), \quad -\frac{k+1}{N} h < \tau \leqq -\frac{kh}{N}, \quad k \in \mathbb{Z},$$

satisfies the inequality

$$\int_{-h}^{0} |\eta(\tau) - \eta^N(\tau)| \, d\tau \leqq \frac{h}{N} \operatorname{VAR}(\eta). \tag{4.1}$$

(ii) For every $\lambda \in \mathbb{C}$, $\lambda \neq -N/h$, let us define the function $e_\lambda^N: [-h, 0] \to \mathbb{C}$ by

$$e_\lambda^N(\tau) = \left(\frac{N}{N+\lambda h} \right)^j, \qquad -\frac{j}{N} h \leqq \tau < -\frac{j-1}{N} h, \qquad j = 0, \cdots, N.$$

Then it is well known that the limit

$$\lim_{N\to\infty} \sup_{-h\leqq\tau\leqq 0} |e^{\lambda\tau} - e_\lambda^N(\tau)| = 0 \tag{4.2}$$

exists uniformly on bounded subsets of the complex plane.

The following convergence result for $\Delta^N(\lambda)$ has been shown by Banks and Burns [2]. For completeness, we present an alternative and simplified proof.

LEMMA 4.2. (i) $\Delta^N(\lambda)$ *converges to* $\Delta(\lambda)$ *uniformly on every bounded subset of the complex plane.*

(ii) *For every* $\alpha \geqq 0$ *there exists a constant* $c_\alpha > 0$ *such that* $|L^N(\lambda)| \leqq c_\alpha$ *for every* $N \in \mathbb{N}$ *with* $N > \alpha h$ *and every* $\lambda \in \mathbb{C}$ *with* $\operatorname{Re}\lambda \geqq -\alpha$.

Proof. Note that

$$L(e^{\lambda\cdot}) = \int_{-h}^{0} e^{\lambda\tau}\, d\eta(\tau) = -\eta(-h)\, e^{-\lambda h} - \lambda \int_{-h}^{0} \eta(\tau)\, e^{\lambda\tau}\, d\tau$$

and, by (3.10) and (2.12),

$$\begin{aligned} L^N(\lambda) &= \sum_{j=0}^{N} \left[\eta^N\left(-\frac{j-1}{N}h\right) - \eta^N\left(-\frac{j}{N}h\right)\right]\left(\frac{N}{N+\lambda h}\right)^j \\ &= -\eta^N(-h)\left(\frac{N}{N+\lambda h}\right)^N - \lambda\frac{h}{N}\sum_{j=1}^{N} \eta^N\left(-\frac{j-1}{N}h\right)\left(\frac{N}{N+\lambda h}\right)^j \\ &= -\eta(-h)\, e_\lambda^N(-h) - \lambda \int_{-h}^{0} \eta^N(\tau)\, e_\lambda^N(\tau)\, d\tau. \end{aligned}$$

Thus statement (i) follows immediately from Remark 4.1. Statement (ii) follows from (3.10) with $c_\alpha = \operatorname{VAR}(\eta) \sup\{(N/(N-\alpha h))^N | N > \alpha h\} < \infty$. □

For the next result we need the space

$$M^\infty = \mathbb{R}^n \times L^\infty[-h, 0; \mathbb{C}^n]$$

endowed with the norm $\|\phi\|_{M^\infty} = \max\{|\phi^0|, \|\phi^1\|_{L^\infty}\}$ for $\phi \in M^\infty$.

THEOREM 4.3. *The limits*

$$\lim_{N\to\infty} \|E_\lambda - \iota^N E_\lambda^N\|_{\mathscr{L}(\mathbb{C}^n, M^\infty)} = 0 = \lim_{N\to\infty} \|T_\lambda - \iota^N T_\lambda^N \pi^N\|_{\mathscr{L}(M^2, M^\infty)}$$

exist uniformly on bounded subsets of the complex plane.

Proof. The statement on E_λ is an immediate consequence of Remark 4.1 (ii), since $\iota^N E_\lambda^N x = (x, e_\lambda^N(\cdot)x) \in M^\infty$ for $x \in \mathbb{C}^n$, $N \in \mathbb{N}$ and $\lambda \in \mathbb{C}$, $\lambda \neq -N/h$.

In order to prove the second part of the theorem, let us first define $e(\lambda, \tau) \in M^2$ by

$$e(\lambda, \tau)^0 = 0, \qquad e(\lambda, \tau)^1(\sigma) = \begin{cases} 0, & -h \leqq \sigma < \tau, \\ e^{\bar{\lambda}(\tau-\sigma)}, & \tau \leqq \sigma \leqq 0, \end{cases}$$

for $\lambda \in \mathbb{C}$ and $-h \leqq \tau \leqq 0$. Then

$$[T_\lambda p^N \phi]^1(\tau) = \langle e(\lambda, \tau), p^N\phi\rangle = \langle p^N e(\lambda, \tau), \phi\rangle.$$

for all $\phi \in M^2$, $\lambda \in \mathbb{C}$, $N \in \mathbb{N}$ and $\tau \in [-h, 0]$. Moreover, the closure of the set $\{e(\lambda, \tau) | |\lambda| \leqq c, -h \leqq \tau \leqq 0\}$ in M^2 is compact and thus p^N converges uniformly on this set. Hence $T_\lambda p^N$ converges to T_λ in $\mathscr{L}(M^2, M^\infty)$ uniformly on bounded subsets of the complex plane.

Secondly, note that

$$[\iota^N T_\lambda^N z]^1(\tau) = \frac{h}{N}\sum_{\nu=1}^{j}\left(\frac{N}{N+\lambda h}\right)^\nu z_{j+1-\nu} = \int_{-jh/N}^{0} e_\lambda^N(\sigma)[\iota^N z]^1\left(-\frac{jh}{N} - \sigma\right) d\sigma$$

and hence

$$\left|[T_\lambda p^N\phi]^1(\tau)-[\iota^N T_\lambda^N \pi^N\phi]^1(\tau)\right|$$

$$\leqq\left|[T_\lambda p^N\phi]^1(\tau)-[T_\lambda p^N\phi]^1\left(-\frac{jh}{N}\right)\right|$$

$$+\int_{-jh/N}^{0}|e^{\lambda\sigma}-e_\lambda^N(\sigma)|\left|[p^N\phi]^1\left(-\frac{jh}{N}-\sigma\right)\right|d\sigma$$

for $-jh/N\leqq\tau<-(j-1)h/N$, $j=1,\cdots,N$. Thus the statement of the theorem follows from Remark 4.1 (ii) together with the fact that the set

$$\{[T_\lambda\phi]^1\,|\,|\lambda|\leqq c,\ \phi\in M^2,\ \|\phi\|<1\}$$

is equicontinuous for $c<\infty$. □

THEOREM 4.4. *For every* $\phi\in M^2$

$$F\phi=\lim_{N\to\infty}\iota^N F^N\pi^N\phi.$$

Proof. We prove this result in 3 steps. The first step is a formula for the operators $\pi^N F$ and $F^N\pi^N$.

Step 1. $[\pi^N F\phi]_0=[F^N\pi^N\phi]_0=\phi^0$ and for $j=1,\cdots,N$

$$[\pi^N F\phi]_j=-\frac{N}{h}\int_{-h}^{h/N}\eta\left(\tau-\frac{j}{N}h\right)[\phi^1(\tau)-\phi^1(\tau-h/N)]\,d\tau,$$

$$[F^N\pi^N\phi]_j=-\frac{N}{h}\int_{-h}^{h/N}\eta^N\left(\tau-\frac{j}{N}h\right)[\phi^1(\tau)-\phi^1(\tau-h/N)]\,d\tau.$$

Proof. Let us define $\phi^1(\tau):=0$ for $t\notin[-h,0]$. Then

$$[\pi^N F\phi]_j=\frac{N}{h}\int_{-jh/N}^{-(j-1)h/N}\int_{-h}^{0}d\eta(\tau)\phi^1(\tau-\sigma)\,d\sigma$$

$$=\frac{N}{h}\int_{-h}^{0}d\eta(\tau)\int_{\tau+(j-1)h/N}^{\tau+jh/N}\phi^1(\sigma)\,d\sigma$$

$$=-\frac{N}{h}\eta(-h)\int_{-h+(j-1)h/N}^{-h+jh/N}\phi^1(\sigma)\,d\sigma$$

$$-\frac{N}{h}\int_{-h}^{0}\eta(\tau)\left[\phi^1\left(\tau+\frac{j}{N}h\right)-\phi^1\left(\tau+\frac{j-1}{N}h\right)\right]d\tau$$

$$=-\frac{N}{h}\eta(-h)\int_{-h}^{-h+jh/N}\left[\phi^1(\tau)-\phi^1\left(\tau-\frac{h}{N}\right)\right]d\tau$$

$$-\frac{N}{h}\int_{-h+jh/N}^{h/N}\eta\left(\tau-\frac{j}{N}h\right)\left[\phi^1(\tau)-\phi^1\left(\tau-\frac{h}{N}\right)\right]d\tau$$

$$=-\frac{N}{h}\int_{-h}^{h/N}\eta\left(\tau-\frac{j}{N}h\right)\left[\phi^1(\tau)-\phi^1\left(\tau-\frac{h}{N}\right)\right]d\tau,$$

since $\eta(\tau)=\eta(-h)$ for $\tau\leqq -h$, and

$$
\begin{aligned}
[F^N\pi^N\phi]_j &= \sum_{\nu=j}^{N} A_\nu^N[\pi^N\phi]_{\nu-j+1} \\
&= \frac{N}{h}\sum_{\nu=j}^{N}\int_{-(\nu-j+1)h/N}^{-(\nu-j)h/N}\left[\eta^N\left(\tau-\frac{j-1}{N}h\right)-\eta^N\left(\tau-\frac{j}{N}h\right)\right]\phi^1(\tau)\,d\tau \\
&= \frac{N}{h}\int_{-h}^{0}\left[\eta^N\left(\tau-\frac{j-1}{N}h\right)-\eta^N\left(\tau-\frac{j}{N}h\right)\right]\phi^1(\tau)\,d\tau \\
&= -\frac{N}{h}\int_{-h}^{h/N}\eta^N\left(\tau-\frac{j}{N}h\right)\left[\phi^1(\tau)-\phi^1\left(\tau-\frac{h}{N}\right)\right]d\tau
\end{aligned}
$$

for $j=1,\cdots,N$. In the last but one equation we have used the fact that

$$
\eta^N\left(\tau-\frac{j-1}{N}h\right)=\eta^N\left(\tau-\frac{j}{N}h\right)\quad\text{for } \tau\leqq -\frac{N-j+1}{N}h.
$$

This proves step 1.

Step 2. $\|F^N\|_N\leqq\max\{1,\mathrm{VAR}(\eta)\}$, $\forall N\in\mathbb{N}$.

Proof. By the well-known convolution inequality, we have

$$
\begin{aligned}
\sum_{j=1}^{N}|[F^Nz]_j|^2 &= \sum_{j=1}^{N}\left|\sum_{\nu=j}^{N}A_\nu^N z_{\nu-j+1}\right|^2 \\
&\leqq \left[\sum_{\nu=1}^{N}|A_\nu^N|\right]^2\sum_{j=1}^{N}|z_j|^2\leqq[\mathrm{VAR}(\eta)]^2\sum_{j=1}^{N}|z_j|^2
\end{aligned}
$$

and hence

$$
\begin{aligned}
\|F^Nz\|_N^2 &= |z_0|^2+\frac{h}{N}\sum_{j=1}^{N}|[F^Nz]_j|^2 \\
&\leqq |z_0|^2+[\mathrm{VAR}(\eta)]^2\frac{h}{N}\sum_{j=1}^{N}|z_j|^2\leqq[\max\{1,\mathrm{VAR}(\eta)\}]^2\|z\|_N^2
\end{aligned}
$$

for $z\in\mathbb{R}^{n(N+1)}$ and $N\in\mathbb{N}$. This proves step 2.

Step 3. $F\phi=\lim_{N\to\infty}\iota^NF^N\pi^N\phi$, $\forall\phi\in M^2$.

Proof. Let us first assume that ϕ^1 is continuously differentiable and that $\phi^1(0)=0$. Moreover let us define $\phi^1(\tau)=0$ for $\tau>0$. Then it follows from step 1 and (4.1) that

$$
\begin{aligned}
&|[\pi^NF\phi]_j-[F^N\pi^N\phi]_j| \\
&\quad= \left|\int_{-h}^{h/N}\left[\eta^N\left(\tau-\frac{jh}{N}\right)-\eta\left(\tau-\frac{jh}{N}\right)\right]\frac{N}{h}[\phi^1(\tau)-\phi^1(\tau-h/N)]\,d\tau\right| \\
&\quad\leqq \int_{-h}^{0}|\eta^N(\tau)-\eta(\tau)|\,d\tau\cdot\sup_{-h+h/N\leqq\tau\leqq h/N}\frac{N}{h}\left|\int_{\tau-h/N}^{\tau}\dot\phi^1(\sigma)\,d\sigma\right| \\
&\quad\leqq \frac{h}{N}\mathrm{VAR}(\eta)\|\dot\phi^1\|_\infty
\end{aligned}
$$

and hence

$$\begin{aligned}\|F\phi - \iota^N F^N \pi^N \phi\| &\leqq \|F\phi - p^N F\phi\| + \|\pi^N F\phi - F^N \pi^N \phi\| \\ &\leqq \|F\phi - p^N F\phi\| + \frac{h}{N}\left[\sum_{j=1}^{N} |[\pi^N F\phi]_j - [F^N \pi^N \phi]_j|^2\right]^{1/2} \\ &\leqq \|F\phi - p^N F\phi\| + \frac{h^2}{N^{3/2}} \mathrm{VAR}(\eta)\|\dot{\phi}^1\|_\infty.\end{aligned}$$

For any $\phi \in M^2$ the statement follows from the Banach-Steinhaus theorem and step 2. This proves the theorem. □

Combining the above convergence results with the concrete representation of the resolvent operators given in Proposition 2.2 and Proposition 3.5, we obtain the following result. The proof is a straightforward application of Lemma 6.1 and will be omitted.

COROLLARY 4.5. *The limits*

$$\lim_{N\to\infty} \|(\lambda I - A)^{-1} - \iota^N (\lambda I - A^N)^{-1} \pi^N\|_{\mathcal{L}(M^2, M^\infty)} = 0,$$

$$\lim_{N\to\infty} \|(\lambda I - A_T^*)^{-1} - \iota^N (\lambda I - (A_T^N)^*)^{-1} \pi^N\|_{\mathcal{L}(M^2)} = 0$$

exist uniformly on those bounded subsets of the complex plane which are uniformly bounded away from the zeros of $\det \Delta(\lambda)$.

THEOREM 4.6.

$$\lim_{N\to\infty} \|G - \iota^N G^N \pi^N\|_{\mathcal{L}(M^2, M^\infty)} = 0.$$

Proof. We establish this result in three steps.

Step 1. Let $X_j^N, j \geqq 0$, be given by (3.5) and let us define

$$X^N(t) := X_j^N, \qquad \frac{j}{N}h \leqq t < \frac{j+1}{N}h, \qquad j = 0, 1, 2, \cdots.$$

Then $X^N(t)$ converges to $X(t)$ uniformly on every compact interval $[0, T]$.

Proof. For every $k \in \mathbb{N}$

$$\begin{aligned}X_k^N &= I + \sum_{\nu=0}^{k-1} [X_{\nu+1}^N - X_\nu^N] \\ &= I + \frac{h}{N} \sum_{\nu=0}^{k-1} \sum_{j=0}^{\nu} A_{\nu-j}^N X_j^N \\ &= I + \frac{h}{N} \sum_{j=0}^{k-1} \sum_{\nu=j}^{k-1} A_{\nu-j}^N X_j^N \\ &= I - \frac{h}{N} \sum_{j=0}^{k-1} \eta^N\left(\frac{j+1-k}{N}h\right) X_j^N \\ &= I - \int_0^{kh/N} \eta^N\left(s - \frac{kh}{N}\right) X^N(s)\, ds\end{aligned}$$

and hence for $kh/N \leqq t < (k+1)h/N$

$$X(t) - X^N(t) = X(t) - X\left(\frac{kh}{N}\right) + \int_0^{kh/N} \left[\eta^N\left(s - \frac{kh}{N}\right) - \eta\left(s - \frac{kh}{N}\right)\right] X(s)\, ds + \int_0^{kh/N} \eta^N\left(s - \frac{kh}{N}\right)[X^N(s) - X(s)]\, ds.$$

Thus the desired convergence result follows from Gronwall's lemma.

Step 2. Let $z \in \mathbb{R}^{n(N+1)}$ and $j \in \{1, \cdots, N\}$. Then

$$[\iota^N G^N z]^0 = X^N(h) z_0 + \int_0^h X^N(s)[\iota^N z]^1(s-h)\, ds,$$

$$[\iota^N G^N z]^1(\tau) = X^N(h+\tau) z_0 + \int_0^h X^N(s)[\iota^N z]^1\left(s - h + \frac{jh}{N}\right) ds,$$

$$-\frac{j}{N} h \leqq \tau < -\frac{j-1}{N} h.$$

Proof. If $-jh/N \leqq \tau < -(j-1)h/N$, then

$$[\iota^N G^N z]^1(\tau) = X^N_{N-j} z_0 + \frac{h}{N} \sum_{l=0}^{N-j-1} X^N_l z_{N-j-l}$$

$$= X^N(h+\tau) z_0 + \sum_{l=0}^{N-j-1} \int_{lh/N}^{(l+1)h/N} X^N(s)[\iota^N z]^1\left(s - h + \frac{jh}{N}\right) ds$$

$$= X^N(h+\tau) z_0 + \int_0^h X^N(s)[\iota^N z]^1\left(s - h + \frac{jh}{N}\right) ds.$$

In the case $j = 0$ this equation leads to the desired expression for $[\iota^N G^N x]^0$.

Step 3. $\lim_{N\to\infty} \| G - \iota^N G^N \pi^N \|_{(M^2, M^\infty)} = 0$.

Proof. First note that the functions $[Gf]^1 \in \mathscr{C}$, $f \in M^2$, $\|f\| \leqq 1$, are equicontinuous since the canonical embedding of $W^{1,2}$ into $\mathscr{C}$ is a compact operator.

Now let $z \in \mathbb{R}^{n(N+1)}$. Then, by step 2,

$$[G\iota^N z - \iota^N G^N z]^0 = [X(h) - X^N(h)] z_0 + \int_0^h [X(s) - X^N(s)][\iota^N z]^1(s-h)\, ds$$

and for $-jh/N \leqq \tau < -(j-1)h/N$, $j = 1, \cdots, N$,

$$[G\iota^N z - \iota^N G^N z]^1(\tau) = [G\iota^N z]^1(\tau) - [G\iota^N z]^1\left(-\frac{jh}{N}\right) + \left[X\left(h - \frac{jh}{N}\right) - X^N\left(h - \frac{jh}{N}\right)\right] z_0 + \int_0^h [X(s) - X^N(s)][\iota^N z]^1\left(s + \frac{jh}{N} - h\right) ds.$$

By step 1 and the equicontinuity mentioned above, this implies

$$\lim_{N\to\infty} \| G\iota^N - \iota^N G^N \|_{\mathscr{L}(\mathbb{R}^{n(N+1)}, M^\infty)} = 0.$$

Moreover, note that the operator $G: M^2 \to M^\infty$ is compact. So is the extended adjoint

operator $G^*: (M^\infty)^* \to M^2$. By Lemma 6.1, this implies

$$\lim_{N\to\infty} \|G - Gp^N\|_{\mathscr{L}(M^2,M^\infty)} = \lim_{N\to\infty} \|G^* - p^N G^*\|_{\mathscr{L}((M^\infty)^*,M^2)} = 0.$$

Hence the statement of the theorem follows from the inequality

$$\|G - \iota^N G^N \pi^N\|_{\mathscr{L}(M^2,M^\infty)} \leqq \|G - Gp^N\|_{\mathscr{L}(M^2,M^\infty)} + \|G\iota^N - \iota^N G^N\|_{\mathscr{L}(\mathbb{R}^{n(N+1)},M^\infty)}. \quad \square$$

Let $f \in M^2$ be given and let $x(t)$, $t \geqq 0$, be the corresponding solution of (2.5). Moreover, let $x^N(t)$, $t \geqq 0$, be defined by

$$x^N(t) = x_k^N, \qquad \frac{k}{N}h \leqq t < \frac{k+1}{N}h, \qquad k \geqq 0,$$

where x_k^N, $k \geqq 0$, is the unique solution of (3.1) corresponding to $f^N = \pi^N f \in \mathbb{R}^{n(N+1)}$. Then the previous theorem shows that

$$\lim_{N\to\infty} \sup_{[0,T]} |x(t) - x^N(t)| = 0$$

and moreover that this convergence is uniform for bounded $f \in M^2$. This has also been proved by Reber [17, Thm. 7.5] under the condition that $L: \mathscr{C} \to \mathbb{R}^n$ is given by (2.2).

Let us now introduce the operator families $S^N(t) \in \mathscr{L}(M^2)$, $S_T^{N*}(t) \in \mathscr{L}(M^2)$, $t \geqq 0$, by

$$S^N(t) = \iota^N \left[I + \frac{h}{N}A^N\right]^k \pi^N, \qquad S_T^{N*}(t) = \iota^N \left[I + \frac{h}{N}(A_T^N)^*\right]^k \pi^N, \tag{4.3}$$
$$\frac{k}{N}h \leqq t < \frac{k+1}{N}h, \qquad k = 0, 1, 2, \cdots.$$

Then the following result is a direct consequence of Theorem 4.4 and Theorem 4.6 together with the factorization results (Theorem 2.1 (i) and Proposition 3.3 (i)).

COROLLARY 4.7. (i) *For all* $\phi \in M^2$, $f \in M^2$

$$S(t)\phi = \lim_{N\to\infty} S^N(t)\phi, \qquad S_T^*(t)f = \lim_{N\to\infty} S_T^{N*}(\tau)f$$

and the convergence is uniform on every compact interval $[0, T]$.

(ii) *For every* $k \in \mathbb{N}$

$$\lim_{N\to\infty} \|S(kh) - S^N(kh)\|_{\mathscr{L}(M^2,M^\infty)} = \lim_{N\to\infty} \|S_T^*(kh) - S_T^{N*}(kh)\|_{\mathscr{L}(M^2)} = 0.$$

Proof. It only remains to note—for the proof of statement (ii)—that, by Lemma 6.1,

$$\lim_{N\to\infty} \|S(h) - G\iota^N F^N \pi^N\|_{\mathscr{L}(M^2,M^\infty)} = \lim_{N\to\infty} \|F^*G^* - \iota^N (F^N)^T \pi^N G^*\|_{\mathscr{L}((M^\infty)^*,M^2)} = 0. \quad \square$$

Statement (ii) of the above result is apparently new. The strong convergence of statement (i) has been stated without proof by Delfour [6]. The strong convergence of $S^N(t)$ has been shown by Reber [17] and Rosen [19].

4.2. Uniform stability. It is a simple consequence of Corollary 4.7 that the discrete time systems (2.15) and (3.4) are stable in a uniform sense if the underlying RFDE (2.1) is exponentially stable. More precisely, we have the following result.

THEOREM 4.8. *Let* $\omega < 0$ *and suppose that* $\det \Delta(\lambda) \neq 0$ *for every* $\lambda \in \mathbb{C}$ *with* $\operatorname{Re}\lambda \geqq \omega$. *Then there exist an* $N_0 \in \mathbb{N}$ *and a constant* $\gamma > 0$ *such that for every* $N \geqq N_0$

$$\left\| \left(I + \frac{h}{N}A^N\right)^k \right\|_N \leqq \gamma e^{\omega kh/N}.$$

Proof. It follows from a well-known result in semigroup theory that there exists a $k_0 \in \mathbb{N}$ such that $\|S(k_0 h)\|_{\mathcal{L}(M^2)} < e^{\omega k_0 h}$. By Corollary 4.7 (ii), this implies the existence of an $N_0 \in \mathbb{N}$ such that

$$\left\| \left(I + \frac{h}{N} A^N \right)^{k_0 N} \right\|_N < e^{\omega k_0 N}, \qquad N \geqq N_0.$$

Moreover, it follows from Corollary 4.7 (i), that

$$\gamma := e^{-\omega k_0 h} \sup \left\{ \left\| \left(I + \frac{h}{N} A^N \right)^l \right\|_N \,\middle|\, l = 0, \cdots, k_0 N - 1,\ N \in N \right\} < \infty.$$

We conclude that the following inequality holds for $N \geqq N_0$ and $k = \nu k_0 N + l$ with $\nu \in \mathbb{N}$ and $l \in \{0, \cdots, k_0 N - 1\}$

$$\left\| \left(I + \frac{h}{N} A^N \right)^k \right\|_N \leqq \left\| \left(I + \frac{h}{N} A^N \right)^l \right\|_N \left\| \left(I + \frac{h}{N} A^N \right)^{k_0 N} \right\|_N^{\nu}$$
$$\leqq \gamma\, e^{\omega k_0 h}\, e^{\omega \nu k_0 h} \leqq \gamma\, e^{\omega(\nu k_0 + l/N)h} = \gamma\, e^{\omega k h / N}. \qquad \square$$

It follows easily from Lemma 4.2 that the stability of the RFDE (2.1) also implies the stability of the approximating continuous-time systems (2.10) and (2.14) if N is sufficiently large (the precise arguments are given in the proof of Theorem 4.9 below). However, a uniform estimate in the spirit of Theorem 4.8 has not yet been proved in the literature on these approximation schemes. It has been stated as a conjecture by Gibson [9] and provides—in that paper—a crucial step in the convergence proof for the solutions of the algebraic Riccati equation. Repin [18] also claims the uniform stability of the approximating systems (2.10), however, his arguments are extremely unclear and it seems almost impossible to convert them into a rigorous proof. The following theorem closes this important gap in the approximation theory of RFDE's and may be considered as the main result of this paper.

THEOREM 4.9. *Let $L: \mathcal{C} \to \mathbb{R}^N$ be given by* (2.2) *and let the* RFDE (2.1) *be exponentially stable. Then the approximating systems* (2.10) *and* (2.14) *are uniformly exponentially stable for sufficiently large N. This means that there exists an $N_0 \in \mathbb{N}$ and constants $\varepsilon > 0$, $\gamma \geqq 1$ such that*

$$\|e^{A^N t}\|_N, \|e^{(A_T^N)^* t}\|_N \leqq \gamma\, e^{-\varepsilon t}$$

for every $t \geqq 0$ and every $N \geqq N_0$.

Proof. First note that the statement on $(A_T^N)^*$ follows from that on A^N. Secondly, it follows from Theorem 6.2 and the exponential estimates in Theorem 2.3. (ii) that it is enough to show that there exists an $N_0 \in \mathbb{N}$ and a constant $c > 0$ such that

$$\int_0^\infty \|e^{A^N t} z\|_N^2\, dt \leqq c^2 \|z\|_N^2$$

for every $z \in \mathbb{R}^{n(N+1)}$ and every $N \geqq N_0$. We will prove this in 5 steps.

Step 1. There exists an $N_0 \in \mathbb{N}$ such that $\det \Delta^N(\lambda) \neq 0$ for every $\lambda \in \mathbb{C}$ with $\operatorname{Re} \lambda \geqq 0$ and every $N \geqq N_0$.

Proof. By Lemma 4.2.(ii), the complex function $\det \Delta^N(\lambda)$ cannot have a zero in the closed right halfplane outside the disc of radius $\mathrm{VAR}(\eta)$ centered at the origin. Inside this disc the nonexistence of unstable eigenvalues of A^N follows from Lemma 4.2.(i) if N is sufficiently large.

Step 2. For $N \in \mathbb{N}$ let us introduce the matrix

$$a^N = \frac{N}{h}\begin{bmatrix} -1 & & & \\ 1 & \ddots & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

Then there exist constant $\varepsilon_0 > 0$, $\gamma_0 \geqq 1$ such that

$$|e^{a^N t}|_{N\times N} \leqq \gamma_0 e^{-\varepsilon_0 t} \quad \forall t \geqq 0, \quad \forall N \in \mathbb{N}.$$

Proof. First of all it is easy to see that $x^T a^N x \leqq 0$ for every $x \in \mathbb{R}^N$ and every $N \in N$. Hence it follows from a well-known result in semigroup theory that

$$|e^{a^N t}|_{N\times N} \leqq 1 \quad \forall t \geqq 0, \quad \forall N \in \mathbb{N}, \tag{4.4}$$

where $|\cdot|_{N\times N}$ denotes the operator norm on $\mathbb{R}^{N\times N}$ which corresponds to the Euclidean norm on $\mathbb{R}^N$. Moreover

$$e^{a^N t} = e^{-Nt/h}\begin{bmatrix} 1 & & & \\ \frac{Nt/h}{1!} & \ddots & & \\ \vdots & \ddots & & \\ \frac{(Nt/h)^{N-1}}{(N-1)!} & \cdots & \frac{Nt/h}{1!} & 1 \end{bmatrix}$$

$$= e^{-Nt/h}\sum_{k=0}^{N-1}\begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ \frac{(Nt/h)^k}{k!} & & & & \vdots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & \ddots & & & \vdots \\ 0 & \cdots & 0 & \frac{(Nt/h)^k}{k!} & 0 \end{bmatrix}$$

and hence

$$|e^{a^N t}|_{N\times N} \leqq \sum_{k=0}^{N-1}\frac{1}{k!}\left(\frac{Nt}{h}\right)^k e^{-Nt/h}$$

for every $t \geqq 0$ and every $N \in \mathbb{N}$. Since

$$\int_0^\infty t^k e^{-Nt/h}\,dt = k!(h/N)^{k+1},$$

this implies

$$\int_0^\infty |e^{a^N t}|_{N\times N}\,dt \leqq \sum_{k=0}^{N-1}\frac{h}{N} = h.$$

Together with (4.4) this estimate proves the statement of step 2 (Theorem 6.2, $p = 1$). More precisely, $\varepsilon_0 > 0$ may be chosen to be any constant less than $1/h$.

Step 3. For every $z \in \mathbb{R}^{n(N+1)}$ and every $N \in \mathbb{N}$

$$\int_{-\infty}^\infty \|T_{i\omega}^N z\|_N^2\,d\omega \leqq \frac{\pi\gamma_0^2}{\varepsilon_0}\|z\|_N^2.$$

Proof. First note that

$$(\lambda I_N - a^N)^{-1} = \frac{h}{N}\begin{bmatrix} \frac{N}{N+\lambda h} & & \\ \vdots & \ddots & \\ \left(\frac{N}{N+\lambda h}\right)^N & \cdots & \frac{N}{N+\lambda h} \end{bmatrix}$$

and hence

$$T_\lambda^N = \begin{bmatrix} 0 & 0 \\ 0 & (\lambda I_N - a^N)^{-1} \otimes I_n \end{bmatrix}. \tag{4.5}$$

Now let $z \in \mathbb{R}^{n(N+1)}$ be given. Then, by step 2, the function

$$\tilde{z}(t) = [e^{a^N t} \otimes I_n] \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix} \in \mathbb{R}^{nN}, \qquad t \geqq 0,$$

is square integrable on the interval $[0, \infty)$ and its Fourier transform

$$\hat{z}(i\omega) = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-i\omega t}\, \tilde{z}(t)\, dt = \frac{1}{\sqrt{2\pi}} [(i\omega I_N - a^N)^{-1} \otimes I_n] \tilde{z}(0)$$

satisfies $\|\hat{z}\|_{L^2[-\infty,\infty;\mathbb{C}^{nN}]} = \|\tilde{z}\|_{L^2[0,\infty;\mathbb{R}^{nN}]}$. Hence it follows from (4.5) and step 2 that

$$\begin{aligned} \int_{-\infty}^\infty \|T_{i\omega}^N z\|_N^2\, d\omega &= \frac{h}{N} \int_{-\infty}^\infty |[(i\omega I_N - a^N)^{-1} \otimes I_n] \tilde{z}(0)|_{\mathbb{C}^{nN}}^2\, d\omega \\ &= \frac{2\pi h}{N} \int_{-\infty}^\infty |\hat{z}(i\omega)|_{\mathbb{C}^{nN}}^2\, d\omega \\ &= \frac{2\pi h}{N} \int_0^\infty |\tilde{z}(t)|_{\mathbb{R}^{nN}}^2\, dt \\ &\leqq \frac{2\pi h}{N} \int_0^\infty \gamma_0^2\, e^{-2\varepsilon_0 t}\, dt\, |\tilde{z}(0)|_{\mathbb{R}^{nN}}^2 \\ &\leqq \frac{\pi \gamma_0^2}{\varepsilon_0} \|z\|_N^2. \end{aligned}$$

This proves step 3.

Step 4. There exists a constant $c > 0$ such that the following inequality holds for every $z \in \mathbb{R}^{n(N+1)}$ and every $N \geqq N_0$

$$\int_{-\infty}^\infty \|(i\omega I - A^N)^{-1} z\|_N^2\, d\omega \leqq 2\pi c^2 \|z\|_N^2.$$

Proof. Recall that

$$(i\omega I - A^N)^{-1} = E_{i\omega}^N \Delta^N(i\omega)^{-1} (E_{i\omega}^N)^T Q^N F^N + T_{i\omega}^N, \qquad \omega \in \mathbb{R},$$

(Proposition 3.5). By step 3, it remains to establish the desired inequality for the first term on the right-hand side of this equation. Moreover, it follows from Theorem 4.4 that the operators F^N are uniformly bounded and it is easy to see that the operators

E_λ^N and $(E_\lambda^N)^* = (E_\lambda^N)^T Q^N$ are uniformly bounded on the imaginary axis. Thus it remains to prove the desired estimate for the term $\Delta^N(i\omega)^{-1}$. But for $|\omega| > \mathrm{VAR}(\eta)$ it follows from Lemma 4.2 (ii) that

$$|\Delta^N(i\omega)^{-1}| = \left| \sum_{k=0}^{\infty} (i\omega)^{-k-1} L^N (i\omega)^k \right| \leqq \frac{1}{|\omega| - \mathrm{VAR}(\eta)}.$$

This inequality, together with Lemma 4.2 (i) shows that

$$\sup_{N \geqq N_0} \int_{-\infty}^{\infty} |\Delta^N(i\omega)^{-1}|^2 \, d\omega < \infty.$$

This proves step 4.

Step 5. For every $z \in \mathbb{R}^{n(N+1)}$ and every $N \geqq N_0$

$$\int_0^{\infty} \|e^{A^N t} z\|_N^2 \, dt \leqq c^2 \|z\|_N^2.$$

Proof. Let $z \in \mathbb{R}^{n(N+1)}$ and $z(t) = e^{A^N t} z$ for $t \geqq 0$ and $N \geqq N_0$. Then $z(t)$ is square integrable on $(0, \infty)$ and its Fourier transform is given by $\hat{z}(i\omega) = (2\pi)^{-1/2}(i\omega I - A^N)^{-1}$. By the Fourier-Plancherel theorem and step 4, we obtain

$$\int_0^{\infty} \|e^{A^N t} z\|_N^2 \, dt = (2\pi)^{-1} \int_{-\infty}^{\infty} \|(i\omega I - A^N)^{-1} z\|_N^2 \, d\omega \leqq c^2 \|z\|_N^2.$$

This proves step 5 and the statement of the theorem. □

Remark 4.10. The uniform exponential decay rate $-\varepsilon$ for the approximating systems (2.10), (2.14) which has been found in the proof of Theorem 4.9 is always larger than $-1/h$. The question remains open if one can find a uniform exponential bound for the approximating systems with the exponential decay rate $\omega_0 + \varepsilon$ where $\omega_0 = \sup\{\mathrm{Re}\,\lambda \,|\det \Delta(\lambda) = 0\}$ and $\varepsilon > 0$ can be chosen arbitrarily small. It is also an open problem if the operators $\iota^N e^{a^N t} \pi^N$ converge to the (compact) operators $S(t) \in \mathscr{L}(M^2)$ in the uniform operator topology if $t \geqq h$. If this could be shown, then the solution to the uniform stability problem mentioned above would be an immediate consequence.

5. Conclusions. The present paper studies in detail certain finite dimensional approximations for linear retarded systems, namely the averaging approximation scheme, both a continuous and a discrete time version as well as the relation between these two. It turns out that these finite dimensional approximations show—under several aspects—precisely the same structure as the underlying RFDE. In particular, the duality relations are of the same type and there are certain structural operators which play an important role for the description of the approximating systems and are analogous to those which have recently been introduced by Bernier-Manitius [3], Manitius [14], Delfour-Manitius [7] for the study of RFDE's. Moreover, it is shown that these operators actually converge to the corresponding operators in the theory of retarded systems. One of these convergence results, namely Theorem 4.6, is only a slight extension of a corresponding result by Reber [17, Thm. 7.5].

Based on this detailed analysis of the structure of the approximating systems, it is shown that both the discrete- and the continuous-time approximations are stable in a uniform sense if the underlying RFDE is asymptotically stable. Such a result is by no means obvious and not all approximation schemes have this property. For example, it is shown in Kappel-Salamon [12] that spline approximations for RFDE's can never have the property of uniform stability. Nevertheless, the uniform stability result provides

a crucial step in the convergence proof of Gibson [9] for the solutions of the algebraic Riccati equation. Moreover, the structural matrix F^N introduced in this paper allows a factorization of the approximate Riccati operator in precisely the same manner as it is shown in Kappel-Salamon [12] for the spline approximation scheme. Finally, it seems likely that the uniform stability results of this paper have some implications for the construction of finite dimensional compensators for RFDE's. This is a research problem for future investigations.

6. Appendix. In this section we formulate and prove two general functional analytic results which are needed frequently in § 4.

LEMMA 6.1. *Let A be an arbitrary set and let X, Y, Z be Banach spaces. Moreover, let $K^\alpha \in \mathcal{L}(X, Y)$, $T^\alpha \in \mathcal{L}(Y, Z)$, $T_k^\alpha \in \mathcal{L}(Y, Z)$, $\alpha \in A$, $k \in \mathbb{N}$, be bounded, linear operators with the properties*

(i) $\operatorname{cl}\{K^\alpha x | \alpha \in A, x \in X, \|x\| \leqq 1\} \subset Y$ *is compact,*

(ii) $T^\alpha y = \lim_{k\to\infty} T_k^\alpha y$ *for all $y \in Y$ uniformly in $\alpha \in A$.*

Then $T_k^\alpha K^\alpha$ tends to $T^\alpha K^\alpha \in \mathcal{L}(X, Z)$ in the uniform operator topology as k tends to infinity and this convergence is uniform in $\alpha \in A$.

Proof. Note that for every $\varepsilon > 0$ there exist finitely many $x_1, \cdots, x_n \in X$ and $\alpha_1, \cdots, \alpha_n \in A$ such that for every $\alpha \in A$ and every $x \in X$ with $\|x\| \leqq 1$ there is a $j \in \{1, \cdots, n\}$ such that $\|K^\alpha x - K^{\alpha_j} x_j\| \leqq \varepsilon$. Hence the desired uniform convergence result follows from the inequality

$$\|T_k^\alpha K^\alpha x - T^\alpha K^\alpha x\| \leqq [\|T_k^\alpha\| + \|T^\alpha\|]\|K^\alpha x - K^{\alpha_j} x_j\| + \|T_k^\alpha K^{\alpha_j} x_j - T^\alpha K^{\alpha_j} x_j\|. \qquad \square$$

The next result is a quantitative estimate for the equivalence of L^p-stability and exponential stability for strongly continuous semigroups. This equivalence has been proved—for the case $p = 2$—by several authors, see for example Datko [5], Curtain-Pritchard [4], Przyluski [16]. But none of these give the desired quantitative estimate which is essential for the proof of uniform stability in Theorem 4.9. Again in the case $p = 2$ such a quantitative estimate can be found without proof in Gibson [8]. We mention that some of the ideas in the proof of the theorem below are taken from Przyluski [16, Prop. 9] and Zabczyk [21, Thm. 5.1].

THEOREM 6.2. *Let $S(t)$, $t \geqq 0$, be a strongly continuous semigroup of bounded, linear operators on a Banach space X satisfying the exponential bound*

$$\|S(t)\|_{\mathcal{L}(X)} \leqq M e^{\omega t}, \qquad t \geqq 0, \tag{6.1}$$

for some constants $M \geqq 1$, $\omega \geqq 0$. Moreover, let $1 \leqq p < \infty$ and suppose that there exists a constant $c > 0$ such that

$$\int_0^\infty \|S(t)x\|^p \, dt \leqq c^p \|x\|^p, \qquad x \in X. \tag{6.2}$$

Then, for every

$$\alpha > -\frac{1}{pc^p M^p}, \tag{6.3}$$

there exists a $\gamma = \gamma(\alpha, \omega, M, c, p) \geqq 1$ such that

$$\|S(t)\|_{\mathcal{L}(X)} \leqq \gamma e^{\alpha t}, \qquad t \geqq 0. \tag{6.4}$$

Remark 6.3. If $-1/pc^pM^p < \alpha < 0$, then there exists a unique $T > 0$ satisfying

$$e^{-\alpha p T} = 1 + \frac{T}{c^p M^p e^{\omega p T}}$$

or equivalently

$$\alpha = \frac{1}{pT} \log \frac{c^p M^p e^{\omega p T}}{T + c^p M^p e^{\omega p T}}. \tag{6.5}$$

The proof of Theorem 6.2 shows that in this case $\gamma = \gamma(\alpha, \omega, M, c, p) \geqq 1$ can be chosen as

$$\gamma = \frac{[T + c^p M^p e^{\omega p T}]^{2/p}}{T^{1/p} c}. \tag{6.6}$$

Proof of Theorem 6.2. Let $T > 0$ be given and let us define

$$\varepsilon = \varepsilon(T) = \frac{T}{T + c^p M^p e^{\omega p T}} > 0. \tag{6.7}$$

Then it follows from (6.1) and (6.2) that

$$\begin{aligned}
\sum_{k=0}^{\infty} \|S(T)^k x\|^p &= \sum_{k=0}^{\infty} \frac{1}{T} \int_0^T \|S(kT)x\|^p \, dt \\
&\leqq \|x\|^p + \sum_{k=1}^{\infty} \frac{1}{T} \int_0^T \|S(T-t)\|^p \|S((k-1)T+t)x\|^p \, dt \\
&\leqq \|x\|^p + \frac{1}{T} \left[\sup_{[0,T]} \|S(t)\|^p \right] \int_0^{\infty} \|S(t)x\|^p \, dt \\
&\leqq \left[1 + \frac{c^p M^p e^{\omega p T}}{T} \right] \|x\|^p \\
&= \varepsilon^{-1} \|x\|^p,
\end{aligned}$$

for every $x \in X$. This implies that

$$\sum_{k=0}^{\infty} \|S(T)^{k+1} x\|^p = \sum_{k=0}^{\infty} \|S(T)^k x\|^p - \|x\|^p \leqq (1-\varepsilon) \sum_{k=0}^{\infty} \|S(T)^k x\|^p$$

and hence

$$\begin{aligned}
\|S(T)^m x\| &\leqq \left[\sum_{k=0}^{\infty} \|S(T)^{k+m} x\|^p \right]^{1/p} \\
&\leqq (1-\varepsilon)^{m/p} \left[\sum_{k=0}^{\infty} \|S(T)^k x\|^p \right]^{1/p} \\
&\leqq (1-\varepsilon)^{m/p} \varepsilon^{-1/p} \|x\|
\end{aligned}$$

for every $x \in X$ and every $m \in \mathbb{N}$. Now let $t = mT + \tau \geqq 0$ with $m \in \mathbb{N}$ and $0 \leqq \tau < T$. Then we conclude that

$$\begin{aligned}
\|S(t)x\| &\leqq \|S(\tau)\| \, \|S(T)^m x\| \\
&\leqq M e^{\omega T} \varepsilon^{-1/p} (1-\varepsilon)^{m/p} \|x\| \\
&\leqq M e^{\omega T} \varepsilon^{-1/p} \exp\left(\frac{m}{p} \log(1-\varepsilon) \right) \exp\left(\frac{\log(1-\varepsilon)}{p} \left(\frac{\tau}{T} - 1 \right) \right) \|x\| \\
&= M e^{\omega T} \varepsilon^{-1/p} (1-\varepsilon)^{-1/p} \exp\left(\frac{\log(1-\varepsilon)}{pT} (mT + \tau) \right) \|x\| \\
&= \gamma e^{\alpha t} \|x\|
\end{aligned}$$

where

$$\alpha = \alpha(T) = \frac{1}{pT} \log \frac{c^p M^p e^{\omega p T}}{T + c^p M^p e^{\omega p T}} < 0$$

and

$$\gamma = \gamma(T) = M e^{\omega T} \left[\frac{1}{\varepsilon(1-\varepsilon)} \right]^{1/p} = \frac{[T + c^p M^p e^{\omega p T}]^{2/p}}{T^{1/p} c}$$

(compare (6.5) and (6.6)). Thus the statement of the theorem follows from the fact that $\alpha(T)$ is strictly increasing for $T > 0$ and satisfies

$$\lim_{T \to 0} \alpha(T) = -\frac{1}{pc^p M^p}. \qquad \square$$

REFERENCES

[1] H. T. BANKS AND J. A. BURNS, *An abstract framework for approximate solutions to optimal control problems governed by hereditary system*, Proc. International Conference on Differential Equations, H. A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 10-25.

[2] ———, *Hereditary control problems: numerical methods based on averaging approximation*, this Journal, 16 (1978), pp. 169-208.

[3] C. BERNIER AND A. MANITIUS, *On semigroups in $\mathbb{R}^n \times L^p$ corresponding to differential equations with delays*, Canad. J. Math., 30 (1978), pp. 897-914.

[4] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Science 8, Springer-Verlag, Berlin, 1978.

[5] R. DATKO, *Uniform asymptotic stability of evolutionary processes in Banach space*, SIAM J. Math. Anal., 3 (1972), pp. 428-554.

[6] M. C. DELFOUR, *The linear quadratic optimal control problem for hereditary differential systems: theory and numerical solution*, Appl. Math. Optim., 3 (1977), pp. 101-162.

[7] M. C. DELFOUR AND A. MANITIUS, *The structural operator F and its role in the theory of retarded systems, Part* I, J. Math. Anal. Appl., 73 (1980), pp. 466-490; *Part* II, J. Math. Anal. Appl., 74 (1980), pp. 359-381.

[8] J. S. GIBSON, *The Riccati integral equations for optimal control problems in Hilbert space*, this Journal, 17 (1979), pp. 537-565.

[9] ———, *Linear quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, this Journal, 21 (1983), pp. 95-139.

[10] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[11] D. HENRY, *Small solutions of linear autonomous functional differential equations*, J. Differential Equations, 8 (1970), pp. 494-501.

[12] F. KAPPEL AND D. SALAMON, *Spline approximation for retarded systems and the Riccati equation*, MRC, Univ. of Wisconsin-Madison, TSR #2680, 1984.

[13] A. MANITIUS, *Controllability, observability, and stabilizability of retarded systems*, Proc. 1976 IEEE Conference on Decision and Control, IEEE Publications, New York, 1976, pp. 752-758.

[14] ———, *Completeness and F-completeness of eigenfunctions associated with retarded functional differential equations*, J. Differential Equations, 35 (1980), pp. 1-29.

[15] R. K. MILLER, *Linear Volterra integro-differential equations as semigroups*, Funkcial. Ekvac., 17 (1974), pp. 629-634.

[16] K. M. PRZYLUSKI, *The Lyapunov equation and the problem of stability for linear bounded discrete-time systems in Hilbert space*, Appl. Math. Optim., 6 (1980), pp. 97-112.

[17] D. D. REBER, *A finite difference technique for solving optimization problems governed by linear functional differential equations*, J. Differential Equations, 32 (1979), pp. 193-232.

[18] YU. M. REPIN, *On the approximate replacement of systems with lag by ordinary differential equations*, J. Appl. Math. Mech., 29 (1966), pp. 254-264.

[19] G. I. ROSEN, *A discrete approximation framework for hereditary systems*, J. Differential Equations, 40 (1981), pp. 377-449.

[20] D. SALAMON, *Control and Observation of Neutral Systems*, RNM 91, Pitman, London, 1984.

[21] J. ZABCZYK, *Remarks on the control of discrete time distributed parameter systems*, this Journal, 12 (1974), pp. 721-735.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 6, November 1985

005

AFFINE FEEDBACK CONTROLLABILITY OF CONSTANT COEFFICIENT DIFFERENTIAL EQUATIONS*

DAHLARD L. LUKES†

Abstract. This paper studies the constant coefficient (closed loop) differential equation $\dot{x} = (A+BK)x+Bv+f(t)$ in $[t_0, t_1]\times R^n$ which arises from substitution of the feedback controller $u = Kx+v$ into the (open loop) control equation $\dot{x} = Ax+Bu+f(t)$. The open loop system is assumed to be controllable. A classical result due to Kalman states that this assumption is equivalent to the condition that rank $[B, AB, \cdots, A^{n-1}B] = n$. Under mild restrictions on $f(t)$ we prove that for any states x^0, x^1 in R^n there exists a constant matrix K and a constant vector v for which the closed loop system has a response satisfying the boundary conditions $x(t_0) = x^0$ and $x(t_1) = x^1$. An equivalence between open loop and closed loop controllability is thereby established: Ultimately based on the implicit function theorem, the approach taken lays the groundwork for computing feedback controllers that do the steering.

Key words. control theory, ordinary differential equations, controllability, affine feedback controller, finite time

1. Introduction. In 1960 Kalman [2] introduced the concept that calls the differential control equation

$$\dot{x} = Ax+Bu \tag{1.1}$$

in $[t_0, t_1]\times R^n$ *controllable* if for any states x^0, x^1 in R^n there exists a (open loop) control function $u(\cdot):[t_0, t_1]\to R^m$ for which (1.1) has a response $x(\cdot)$ satisfying the end conditions $x(t_0) = x^0$ and $x(t_1) = x^1$. He showed that if the $n\times n$ A and $n\times m$ B are time-invariant matrices then this open loop controllability is equivalent to the condition

$$\text{rank}\,[B, AB, \cdots, A^{n-1}B] = n. \tag{1.2}$$

(The $n\times nm$ matrix in (1.2) is called the *controllability matrix* of (1.1).)

While it has an obvious significance in optimization problems for which boundary conditions are imposed on the end states, it turns out that the more subtle and most profound implications of open loop controllability appear in problems concerned with alteration of system dynamics and input-output characteristics by means of feedback control. (See [4] for a detailed treatment of the problems of pole placement, decoupling and reduction of control dimension using feedback.) These developments motivated the question studied in this paper: Is it possible, at the outset, to base the notion of controllability on feedback?

Before making a formal statement of the problem we should add that Kalman's definition applies to the more general (nonhomogeneous) (2.1) in which $f(t)$ is integrable on $[t_0, t_1]$. However, that equation can be reduced to one of the homogeneous type (1.1) without disturbing the controllability by a preliminary change of variable $y = x-\bar{x}(t)$ in which $\bar{x}(t)$ is any solution to the nonhomogeneous equation with the control function set to zero. It will become obvious that such a reduction is not possible for the problem treated in this article.

2. The feedback controllability problem and a preview of the results. This paper investigates the controllability of linear, constant coefficient, nonhomogeneous

* Received by the editors May 24, 1983, and in revised form August 16, 1984.
† Department of Applied Mathematics, University of Virginia, Charlottesville, Virginia 22901.

equations

$$\dot{x} = Ax + Bu + f(t), \tag{2.1}$$

working with feedback controllers defined by the equation

$$u = Kx + v \tag{2.2}$$

rather than directly with open loop control functions $u(t)$. The matrix K and the vector v are not allowed to be time dependent.

DEFINITION 2.1. An *admissible feedback controller* is an affine map of R^n into R^m determined by (2.2) when the real $m \times n$ matrix K and the real m-vector v are specified. The differential equation

$$\dot{x} = (A + BK)x + Bv + f(t) \tag{2.3}$$

that results from substitution of (2.2) into (2.1) to eliminate the control variable u is called the *corresponding closed loop system* of (2.1).

Recall that a vector function $x(\cdot)$ on an interval $[t_0, t_1]$ is called a *solution* to (2.3) if it is absolutely continuous and satisfies the differential equation at all points in $[t_0, t_1]$ with the possible exception of a subset of Lebesgue measure zero, i.e., almost everywhere in $[t_0, t_1]$. The forcing term f is assumed to be integrable. This assures the existence and uniqueness of a solution to the initial value problem for (2.3).

We investigate the problem of choosing K and v so that (2.3) has a solution satisfying the boundary conditions

$$x(t_0) = x^0, \qquad x(t_1) = x^1. \tag{2.4}$$

DEFINITION 2.2. For A, B and f fixed the differential control equation, (2.1), is called *feedback controllable on* $[t_0, t_1]$ if for any states x^0, x^1 in R^n there exists an admissible feedback controller for which the corresponding boundary-value problem (2.3)–(2.4) has a solution $x(\cdot)$ on $[t_0, t_1]$.

Obviously if (2.1) is feedback controllable on $[t_0, t_1]$, then it is open loop controllable on $[t_0, t_1]$ in the sense of Kalman. The main problem addressed in this paper is the question: Is an open loop controllable system feedback controllable? We are able to prove the following results.

THEOREM 2.1. *If the coefficients satisfy* (1.2) *and either n is equal to* 1, *n is even or f is of bounded variation on* $[t_0, t_1]$ *then* (2.1) *is feedback controllable on* $[t_0, t_1]$.

COROLLARY 2.1. *The homogeneous equation,* (1.1), *is closed loop controllable if and only if it is open loop controllable* (*on any and hence on all intervals*).

The results obtained in the article are stronger than what is stated in Theorem 2.1. For example, the restriction that f be of bounded variation can be relaxed to an attenuation rate condition on the spectral density function of f (see Theorem 4.3). More importantly, it is shown that if feedback controllability can fail at all, with f but integrable, then failure must occur for rather special f and for boundary states in a well isolated set in the complement of an open and dense subset of $R^n \times R^n$ for $n > 1$ odd. In particular, still assuming that A, B satisfy (1.2), the boundary control problem has a solution K, v for generic $(x^0, x^1, f) \in R^n \times R^n \times L_1^n(t_0, t_1)$. Whether the exceptional cases that appear in the proof are indicating a limit to the extent to which a controllable system can remain feedback controllable under severe external forcing or whether they are only artifacts of the author's method of proof remains an open question. We should add that the method of proof, ultimately based on the implicit function theorem, provides information that would be useful for computing the admissible controls which accomplish the steering.

The results are adequate to cover most applications. To support this claim we point out that from the change of variable $y = x - \int f\,d\sigma$ it becomes apparent that if (1.2) holds then, for integrable f, system (2.1) can be steered from any point (t_0, x^0) to any other point (t_1, x^1) by a controller of the type

$$u = K\left[x - \int_{t_0}^{t} f\,d\sigma\right] + v, \tag{2.5}$$

(K, v constant). The same remark applies for controllers of the type

$$u = K[x - \bar{x}(t)] + v \tag{2.6}$$

in which $\bar{x}(t)$ is as described in the Introduction. These conclusions follow from Theorem 2.1. (See Corollary 4.1 as well.)

Corollary 2.1 establishes the equivalence of the notions of open loop and feedback controllability.

3. Mathematical preliminaries. This section gathers together various preliminary results which contribute to the main proof and on occasion are of interest in their own right.

Let F be any field of characteristic zero and for n a positive integer denote an element of F^n by $k = (k_1, k_2, \cdots, k_n)$. Theorem 3.1 deals with a matrix equation having a coefficient of the form

$$A_n(k) = \begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & & 1 & 0 \\ 0 & 0 & & \cdots & & 0 & 1 \\ k_n & k_{n-1} & & \cdots & & k_2 & k_1 \end{bmatrix}. \tag{3.1}$$

THEOREM 3.1. *Select arbitrary $k \in F^n$ and let J be any $n \times n$ matrix over F. Then the matrix solutions N over F to the equation*

$$JN - NA_n(k) = 0 \tag{3.2}$$

are precisely those of the type

$$N = [p_{n-1}(J)a, p_{n-2}(J)a, \cdots, p_1(J)a, a], \tag{3.3}$$

$a \in \mathcal{N}$, in which $\mathcal{N}$ is the null space of $p_n(J)$ and p_r is the characteristic polynomial of the respective $r \times r$ submatrix $A_r(k)$, $(r = 1, 2, \cdots, n)$. Moreover,

$$\det(N) = \det[J^{n-1}a, J^{n-2}a, \cdots, Ja, a]. \tag{3.4}$$

Remark 3.1. One can easily verify that

$$p_r(x) = x^r - k_1x^{r-1} - k_2x^{r-2} - \cdots - k_{r-1}x - k_r. \tag{3.5}$$

Remark 3.2. An important special case of Theorem 2.1 is the one where the minimal polynomial of J divides p_n. This is precisely the case for which $\mathcal{N} = F^n$. Theorem 3.1 reduces the question of whether (3.2) has a nonsingular solution N over F to the question of whether J has a cyclic vector in $\mathcal{N}$.

Proof of Theorem 3.1. By writing out the $n \times n$ identity matrix I_n as $[e_1, e_2, \cdots, e_n]$ in terms of its columns it is apparent that

$$A_n(k)e_1 = k_n e_n, \tag{3.6}$$

$$A_n(k)e_r = e_{r-1} + k_{n-r+1}e_n \tag{3.7}$$

$(r = 2, 3, \cdots, n)$. Since J and N can be written, respectively, as $[Je_1, Je_2, \cdots, Je_n]$ and $[Ne_1, Ne_2, \cdots, Ne_n]$, by employing (3.6)-(3.7), the matrix equation (3.2) can be written out in terms of columns as the system of n-vector equations

$$JNe_r - Ne_{r-1} - k_{n-r+1}Ne_n = 0, \tag{3.8}$$

$(r = 1, 2, \cdots, n)$, in which e_0 is defined to be 0. The matrix form of system (3.8) is

$$\begin{bmatrix} J & 0 & 0 & \cdots & 0 & -k_nI \\ -I & J & 0 & \cdots & 0 & -k_{n-1}I \\ 0 & -I & J & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & & \ddots & J & -k_2I \\ 0 & 0 & 0 & \cdots & -I & (J - k_1I) \end{bmatrix} \begin{bmatrix} Ne_1 \\ Ne_2 \\ \vdots \\ \vdots \\ Ne_{n-1} \\ Ne_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{3.9}$$

in which I is the $n \times n$ identity matrix. By adding J times each row to the one above it, starting at the bottom and working upward to the top, the coefficient matrix in (3.9) is transformed by elementary row operations into a row equivalent matrix. From this procedure it is evident that (3.9) is equivalent to the equation

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & p_n(J) \\ -I & 0 & 0 & \cdots & 0 & p_{n-1}(J) \\ 0 & -I & 0 & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & & \ddots & 0 & p_2(J) \\ 0 & 0 & 0 & \cdots & -I & p_1(J) \end{bmatrix} \begin{bmatrix} Ne_1 \\ Ne_2 \\ \vdots \\ \vdots \\ Ne_{n-1} \\ Ne_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ 0 \end{bmatrix}. \tag{3.10}$$

Obviously a solution to (3.10) must have $Ne_n \in \mathcal{N}$ where $\mathcal{N}$ is the null space of $p_n(J)$. Hence for each $a \in \mathcal{N}$ equation (3.10) has the solution

$$Ne_r = p_{n-r}(J)a, \tag{3.11}$$

$(r = 1, 2, \cdots, n)$, with p_0 defined to be the identity, and there cannot be any other solutions. The reader can easily check the validity of (3.4) by keeping in mind the form of the p_r pointed out by Remark 3.1 and performing the appropriate column operations of the solution N given by (3.3). This concludes the proof of Theorem 3.1.

COROLLARY 3.1. *Assume that the matrices J and $A_n(k)$ have the same characteristic polynomial p_n. Then $\mathcal{N} = F^n$. If p_n splits into a product of linear factors over F,*

$$p_n(x) = (x - z_1)(x - z_2) \cdots (x - z_n), \tag{3.12}$$

then there exists a nonsingular solution N to (3.2) *if and only if for each eigenvalue z_r the associated eigenspace of J in F^n has dimension one. In particular this occurs if the eigenvalues $\{z_1, z_2, \cdots, z_n\}$ are distinct. If J is furthermore assumed to be in Jordan form, then each solution* (3.3) *to* (3.2) *has*

$$\det(N) = \prod_1^n a_r \prod_{j<s} (z_j - z_s). \tag{3.13}$$

Proof. If the characteristic polynomial p_n of $A_n(k)$ is also the characteristic polynomial of J, then $p_n(J)=0$ by the Cayley-Hamilton theorem. This says $\mathcal{N}=F^n$. For the remainder of the proof, since p_n is assumed to split over F, there is no loss in taking J to be in Jordan form. The standard theory of that form says that the one-dimensionality of the eigenspaces of J associated with its distinct eigenvalues is equivalent to J having but one elementary Jordan block for each distinct eigenvalue and that in turn is equivalent to J having a cyclic vector in F^n. Combined with the conclusion of Theorem 3.1, which says that formula (3.3) produces all solutions N to (3.2) and that their determinant is the value described in (3.4), the proof of Corollary 3.1 is complete except for verification of (3.13).

If the eigenvalues $\{z_1, z_2, \cdots, z_n\}$ are distinct, then the Jordan J is diagonal and we compute

$$[J^{n-1}a, J^{n-2}a, \cdots, Ja, a]=\begin{bmatrix} z_1^{n-1}a_1 & z_1^{n-2}a_1 & \cdots & z_1a_1 & a_1 \\ z_2^{n-1}a_2 & z_2^{n-2}a_2 & \cdots & z_2a_2 & a_2 \\ \vdots & \vdots & & \vdots & \vdots \\ z_n^{n-1}a_n & z_n^{n-2}a_n & \cdots & z_na_n & a_n \end{bmatrix}. \tag{3.14}$$

From (3.14) it follows readily that

$$\det[J^{n-1}a, J^{n-2}a, \cdots, Ja, a]=\prod_1^n a_r \prod_{j<s}(z_j-z_s). \tag{3.15}$$

Application of (3.4) of Theorem 3.1 to (3.15) finishes the proof.

Remark 3.3. All the assumptions of Corollary 3.1 are met for $F=C$, the complex field, if $J=\operatorname{diag}(z_1, z_2, \cdots, z_n)$ when we take the coordinates of k to be the symmetric functions of the variables $z_r \in C$

$$\begin{aligned} k_1(z_1, z_2, \cdots, z_n) &= z_1+z_2+\cdots+z_n, \\ k_2(z_1, z_2, \cdots, z_n) &= -(z_1z_2+z_1z_3+\cdots+z_1z_n+z_2z_3+z_2z_4+\cdots+z_{n-1}z_n), \\ &\vdots \\ k_n(z_1, z_2, \cdots, z_n) &= (-1)^{n+1}z_1z_2\cdots z_n. \end{aligned} \tag{3.16}$$

Recall the definition of the p_r given in Theorem 3.1. A simple inductive proof shows that

$$p_r(z_j)=-k_r(z_1, z_2, \cdots, z_{j-1}, 0, z_{j+1}, \cdots, z_n), \tag{3.17}$$

$(j=1, 2, \cdots, n)$, $(r=1, 2, \cdots, n)$. With the exception of possible multiplication by a minus sign, $p_r(z_j)$ can be computed by striking out z_j from the set of variables $\{z_1, z_2, \cdots, z_n\}$ and adding all terms obtained by taking products of the remaining variables r at a time. It will be helpful to think of it in this manner later in the paper where it plays a prominent role. The reader is hereby forewarned not to assign a simple functional interpretation to the convenient if unusual notation $p_r(z_j)$. (3.17) shows that $p_r(z_j)$ is independent of z_j.

The matrix $A_n(k)$ defined by (3.1) appears as a coefficient in the boundary-value problem on $[0, 1]\times R^n$,

$$\dot{x}=A_n(k)x+\gamma e_n+f(t), \tag{3.18}$$

$$x(0)=x^0, \qquad x(1)=x^1, \tag{3.19}$$

with the integrable function $f:[0, 1]\to R^n$ arbitrarily prescribed. (e_n denotes the last column of the $n\times n$ identity matrix.) Lemmas 3.1, 3.2 and 3.3 are related to the question

of whether $k \in R^n$ and $\gamma \in R$ can be chosen so that the boundary-value problem (3.18)-(3.19) has a solution.

Transform notation. For $f:[0,1] \to R^n$ integrable it will be convenient to work with the transformed functions of the variable $z \in C$,

$$F_r(z) = \int_0^1 e^{-\sigma z} f_r(\sigma)\, d\sigma, \tag{3.20}$$

$(r = 1, 2, \cdots, n)$. Since f_r is integrable, it follows from Fubini's theorem that F_r is an entire function of z. F_r can be viewed as the Laplace transform of the function f_r extended as zero on $[1, \infty)$. Throughout the paper F_0 denotes the transform of the function identically equal to 1 on $[0, 1]$.

We alert the reader to the fact that the system of nonlinear equations (3.21) that follow are highly coupled as are (3.27). (See Remark 3.3 and Equation (3.17).)

LEMMA 3.1. *Let $n = 2q$ be even. If there exist distinct $z_j \in C - R$ satisfying the equations*

$$\gamma F_0(z_j) = \sum_{r=1}^{n} [e^{-z_j} x_r^1 - x_r^0 - F_r(z_j)] p_{n-r}(z_j), \tag{3.21}$$

$(j = 1, 2, \cdots, n)$, *with* $\bar{z}_j = z_{q+j}$, $(j = 1, 2, \cdots, q)$, *then the boundary-value problem* (3.18)-(3.19) *has a solution for k determined by* (3.16).

Proof. Let z_j, $(j = 1, 2, \cdots, n)$, and $\gamma \in R$ satisfy the hypothesis of Lemma 3.1. Then equations (3.16) define a $k \in R^n$. The coordinates of k appear in the coefficients of the polynomials p_r according to (3.5). If we define the matrix $J = \operatorname{diag}(z_1, z_2, \cdots, z_n)$ and let $a \in R^n$ be the vector whose coordinates are all ones, then (3.21) can be rewritten as the n-vector equation,

$$\gamma F_0(J)a = e^{-J} \sum_{r=1}^{n} x_r^1 p_{n-r}(J)a - \sum_{r=1}^{n} x_r^0 p_{n-r}(J)a - \sum_{r=1}^{n} F_r(J) p_{n-r}(J)a. \tag{3.22}$$

Corollary 3.1 applies to J and $A_n(k)$ when the field is taken to be C and since the z_j's are distinct and $\prod_1^n a_r = 1$, it ensures that the matrix N defined by (3.3) is invertible and satisfies (3.2). Hence (3.22) can be rewritten in terms of N as

$$\gamma F_0(J) N e_n = e^{-J} \sum_{1}^{n} x_r^1 N e_r - \sum_{1}^{n} x_r^0 N e_r - \int_0^1 e^{-\sigma J} \sum_{1}^{n} f_r(\sigma) N e_r\, d\sigma, \tag{3.23}$$

which says that

$$\gamma \int_0^1 e^{-\sigma J} N e_n\, d\sigma = e^{-J} N x^1 - N x^0 - \int_0^1 e^{-\sigma J} N f(\sigma)\, d\sigma. \tag{3.24}$$

Multiplication of (3.24) by N^{-1} and application of (3.2) gives

$$\gamma \int_0^1 e^{-\sigma A_n(k)} e_n\, d\sigma = e^{-A_n(k)} x^1 - x^0 - \int_0^1 e^{-\sigma A_n(k)} f(\sigma)\, d\sigma \tag{3.25}$$

which is equivalent to saying that

$$x^1 = e^{A_n(k)} x^0 + \int_0^1 e^{(1-\sigma) A_n(k)} [\gamma e_n + f(\sigma)]\, d\sigma. \tag{3.26}$$

By the variation of parameters formula [4, p. 74] this equation simply says that $x(1) = x^1$ where $x(\cdot):[0,1] \to R^n$ is the unique solution of (3.18) satisfying the initial condition $x(0) = x^0$. This ends the proof of Lemma 3.1.

LEMMA 3.2. *Let $n=2q+1$ be odd. If there exist distinct $z_j \in C-R$ and $z_n=c\in R$ satisfying the equations*

$$F_0(z_j)\sum_{r=1}^{n}[e^{-c}x_r^1-x_r^0-F_r(c)]p_{n-r}(c)=F_0(c)\sum_{r=1}^{n}[e^{-z_j}x_r^1-x_r^0-F_r(z_j)]p_{n-r}(z_j), \tag{3.27}$$

$(j=1,2,\cdots,2q)$, with $\bar{z}_j=z_{q+j}$, $(j=1,2,\cdots,q)$, then the boundary-value problem (3.18)-(3.19) *has a solution where k is determined by* (3.16) *and $\gamma\in R$ is computed by the formula*

$$\gamma=F_0(c)^{-1}\sum_{r=1}^{n}[e^{-c}x_r^1-x_r^0-F_r(c)]p_{n-r}(c). \tag{3.28}$$

Proof. Let z_j, $(j=1,2,\cdots,n)$, satisfy the hypothesis of Lemma 3.2. Then (3.16) defines a $k\in R^n$. The coordinates of k appear in the coefficients of the polynomials p_r according to (3.5). Since these polynomials have real coefficients, it is clear that (3.28) defines a $\gamma\in R$. By rewriting (3.27) in terms of γ using (3.28), it is easily shown that (3.27)-(3.28) can be transformed into the equivalent system,

$$\gamma F_0(z_j)=\sum_{r=1}^{n}[e^{-z_j}x_r^1-x_r^0-F_r(z_j)]p_{n-r}(z_j), \tag{3.29}$$

$$\gamma F_0(c)=\sum_{r=1}^{n}[e^{-c}x_r^1-x_r^0-F_r(c)]p_{n-r}(c), \tag{3.30}$$

$(j=1,2,\cdots,2q)$, or more concisely, just

$$\gamma F_0(z_j)=\sum_{r=1}^{n}[e^{-z_j}x_r^1-x_r^0-F_r(z_j)]p_{n-r}(z_j), \tag{3.31}$$

$(j=1,2,\cdots,n)$. By defining $a\in R^n$ and $J=\operatorname{diag}(z_1,z_2,\cdots,z_n)$ as in the proof of Lemma 3.1, the proof of Lemma 3.2 can be completed by applying the same sequence of operations on (3.31) as was done in the proof of the preceding lemma. The details are not repeated here but are easily supplied. This proves Lemma 3.2.

In the next lemma, μ denotes Lebesgue measure and E' is the complement of E in $[0,1]$.

LEMMA 3.3. *Consider any x_1 and x_2 in R and assume that $f_1:[0,1]\to R$ is integrable. For some $c\in R$ there is a solution $w\in C$ to the equation*

$$e^w=\frac{x_1+F_1(c)}{e^{-c}x_1+x_2F_0(c)+F_1(c)} \tag{3.32}$$

if one of the following conditions holds:

(1) $x_1\neq 0$.

(2) $x_1=0$, $x_2\neq 0$ *and for some $E\subset[0,1]$ with $\mu(E)>0$, $f_1(\sigma)+x_2\neq 0$ and $f_1(\sigma)\neq 0$ for all $\sigma\in E$.*

(3) $x_1=0$, $x_2\neq 0$ *and for some $E\subset[0,1]$ with $0<\mu(E)<1$, $f_1(\sigma)+x_2=0$ for all $\sigma\in E$ and $f_1(\sigma)=0$ for a.e. $\sigma\in E'$.*

Proof. Let x_1 and x_2 in R and integrable $f_1:[0,1]\to R$ satisfy one of conditions (1)-(3) of Lemma 3.3. The problem is to prove that for some $c\in R$ the numerator and denominator of the right-hand side of (3.32) are not zero. To reach a contradiction suppose that there were no such c.

We first argue the impossibility of either of the equations

$$x_1+F_1(c)=0, \tag{3.33}$$

$$e^{-c}x_1 + x_2F_0(c) + F_1(c) = 0 \tag{3.34}$$

holding for all $c \in R$. Suppose that (3.33) held for all $c \in R$. By letting $c \to \infty$ it would follow from application of the Lebesgue dominated convergence theorem that $x_1 = 0$ and hence that

$$F_1(c) = \int_0^1 e^{-\sigma c} f_1(\sigma)\, d\sigma = 0 \tag{3.35}$$

for all $c \in R$. By writing the exponential function as a power series and applying Fubini's theorem we see (3.35) implies that

$$\sum_0^\infty \frac{(-c)^r}{r!} \int_0^1 \sigma^r f_1(\sigma)\, d\sigma = 0, \tag{3.36}$$

for all $c \in R$, $(r = 0, 1, 2, \cdots)$. This in turn supplies the equations

$$\int_0^1 \sigma^r f_1(\sigma)\, d\sigma = 0, \tag{3.37}$$

$(r = 0, 1, 2, \cdots)$. An integration by parts shows that

$$\int_0^1 \sigma^r \int_0^\sigma f_1(t)\, dt\, d\sigma = 0, \tag{3.38}$$

$(r = 0, 1, 2, \cdots)$. With f_1 integrable it is an easy matter to show that

$$g(\sigma) = \int_0^\sigma f_1(t)\, dt \tag{3.39}$$

defines an element of the Hilbert space $L_2(0, 1)$ and since the functions $\{1, \sigma, \sigma^2, \cdots\}$ constitute a basis for that Hilbert space it is concluded from (3.38) that $g(\sigma) = 0$ for a.e. $\sigma \in [0, 1]$. But this contradicts the assumption that one of the conditions (1)-(3) holds since it is inconsistent with each of them. A similar argument that starts by multiplying (3.34) by e^c and letting $c \to -\infty$ shows that the possibility for (3.34) holding for all $c \in R$ is likewise ruled out.

Could there be a $c^* \in R$ at which one of (3.33) or (3.34) holds while the other fails? If there were such a c^*, then by continuity the equation that fails at c^* would fail for all c in some open interval containing c^* and thus the equality holding at c^* must continue to hold on the mentioned open interval. (Otherwise we violate our assumption that for no $c \in R$ do both (3.33) and (3.34) fail.) Since the left-hand sides of (3.33) and (3.34) are entire functions, by the identity theorem of complex function theory we end up with one of (3.33)-(3.34) holding for all $c \in R$—a possibility already eliminated. Thus we are forced to conclude that for some $c \in R$ the right-hand side of (3.32) is defined and not zero. We can now take w to be any of its logarithms and the proof of Lemma 3.3 is finished.

The Riemann-Lebesgue theorem has the following consequences. (In Theorems 3.2 and 3.3 s denotes a real parameter.)

THEOREM 3.2. *For $f : [0, 1] \to R$ integrable and F its transform,*

$$F(w + s2\pi i) \to 0 \tag{3.40}$$

as $|s| \to \infty$, uniformly on compact subsets of C.

THEOREM 3.3. *Suppose that* $g \in L_1(0,1)$ *satisfies the inequality*

$$\left| \int_0^t e^{-\sigma\omega i} g(\sigma)\, d\sigma \right| \leqq \frac{h(t)}{\omega} \tag{3.41}$$

for a.e. $t \in [0,1]$ *and all* $\omega \geqq \Omega$ *for some* $\Omega \in R$ *and* $h \in L_1(0,1)$. *Then its finite Laplace transform* $G(z)$ *has*

$$\lim_{s\to\infty} e^s G(w+s+se^s i) = 0 \tag{3.42}$$

for all $w \in C$. *If* (3.41) *holds moreover at* $t=1$ *with* $h(1)$ *finite then* $g \in L_2(0,1)$. *In particular if* g *is of bounded variation on* $[0,1]$ *then* (3.41) *is satisfied for all* $t \in [0,1]$ *with* $h(t) = |g(t)| + |g(0)| + V$ *where* V *is the total variation of* g *on* $[0,1]$.

Proof. Assume the hypothesis. By an integration by parts we see that

$$|e^s G(w+s+se^s i)| \leqq |e^{-w} G(se^s i)| + |s+w| e^{s+|w|} \int_0^1 e^{-ts} \left| \int_0^t e^{-\sigma s e^s i} g(\sigma)\, d\sigma \right| dt. \tag{3.43}$$

Application of (3.41) to (3.43) and invocation of the Riemann-Lebesgue theorem and the Lebesgue dominated convergence theorem give (3.42).

Now further assume that (3.41) holds at $t=1$. By continuity there exists a bound b on the integral appearing in (3.41) for $t=1$ and $\omega \geqq \Omega$. By extending g to be zero outside $[0,1]$, applying the Plancherel theorem, (6, p. 187), as well as (3.41) with $t=1$, we see that

$$\begin{aligned}
\int_0^1 |g(\sigma)|^2\, d\sigma &= \int_{-\infty}^{\infty} |g(\sigma)|^2\, d\sigma \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} e^{-\sigma\omega i} g(\sigma)\, d\sigma \right|^2 d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \int_0^1 e^{-\sigma\omega i} g(\sigma)\, d\sigma \right|^2 d\omega \\
&= \frac{1}{2\pi} \int_{|\omega|\leqq\Omega} \left| \int_0^1 e^{-\sigma\omega i} g(\sigma)\, d\sigma \right|^2 d\omega + \frac{1}{2\pi} \int_{|\omega|\geqq\Omega} \frac{h^2(1)}{\omega^2}\, d\omega \\
&\leqq \frac{b^2\Omega}{\pi} + \frac{h^2(1)}{\pi\Omega} < \infty.
\end{aligned} \tag{3.44}$$

This proves that $g \in L_2(0,1)$. The conclusion concerning g of bounded variation follows readily by an integration by parts.

Remark 3.4. In dealing with the boundary-value problem based on Equations (4.1)-(4.2) in the next section of this paper there will be no loss of generality in assuming that $t_0 = 0$ and $t_1 = 1$. This normalization can be accomplished by means of the changes of variables and parameters

$$t = (1-\tau)t_0 + \tau t_1, \qquad 0 \leqq \tau \leqq 1, \tag{3.45}$$

$$x^*(\tau) = Dx(t), \tag{3.46}$$

$$f^*(\tau) = (t_1 - t_0) Df(t), \tag{3.47}$$

$$k^* = (t_1 - t_0) Dk, \tag{3.48}$$

$$\gamma^* = (t_1 - t_0)^n \gamma, \tag{3.49}$$

$$D = \operatorname{diag}[1, (t_1 - t_0), \cdots, (t_1 - t_0)^{n-1}]. \tag{3.50}$$

Remark 3.5. Equations (3.18)-(3.19), the normalized form described in Remark 3.4, can be simplified further by taking $x^1 = 0$ in the application forthcoming. To see this, introduce a new variable $x^* = x - x^1$, a new function $f^* = f + A_n(0)x^1$ and a new parameter $\gamma^* = k_n x_1^1 + k_{n-1}x_2^1 + \cdots + k_2 x_{n-1}^1 + k_1 x_n^1 + \gamma$. This transforms equations (3.18)-(3.19) into

$$\dot{x}^* = A_n(k)x^* + \gamma^* e_n + f^*(t), \tag{3.51}$$

$$x^*(0) = x^{*0}, \qquad x^*(1) = 0 \tag{3.52}$$

in which $x^{*0} = x^0 - x^1$. If for f^* and each $x^{*0} \in R^n$ the k and γ^* can be chosen so that (3.51)-(3.52) has a solution, then (3.18)-(3.19) will have a solution relative to k and the parameter $\gamma = \gamma^* - [k_n x_1^1 + k_{n-1}x_2^1 + \cdots + k_2 x_{n-1}^1 + k_1 x_n^1]$.

4. Feedback controllability in finite time. In its elemental form the main problem dealt with in this section concerns the existence of a solution to the boundary-value control problem on $[t_0, t_1] \times R^n$,

$$\dot{x} = A_n(k)x + \gamma e_n + f(t), \tag{4.1}$$

$$x(t_0) = x^0, \qquad x(t_1) = x^1. \tag{4.2}$$

THEOREM 4.1. *If n is even and f is integrable on $[t_0, t_1]$, then there exist $k \in R^n$ and $\gamma \in R$ such that* (4.1)-(4.2) *have a solution $x(\cdot)$.*

Proof. Let $n = 2q$. In view of Lemma 3.1 and Remarks 3.4 and 3.5 it is sufficient to prove the existence of a $\gamma \in R$ and distinct $z_j \in C - R$ satisfying

$$\gamma F_0(z_j) = -\sum_{r=1}^{n} [x_r + F_r(z_j)]p_{n-r}(z_j), \tag{4.3}$$

$(j = 1, 2, \cdots, n)$, with $\bar{z}_j = z_{q+j}$, $(j = 1, 2, \cdots, q)$. (We drop the superscript from x_r^0 to keep the notation simple.)

With this goal in mind consider the related equations

$$F_0(z_j) = -\varepsilon \sum_{r=1}^{n} [x_r + F_r(z_j)]p_{n-r}(z_j), \tag{4.4}$$

$(j = 1, 2, \cdots, n)$, that would result from dividing both sides of (4.3) by γ and renaming $1/\gamma$ as ε. Select any q integers s_j for which $1 \leqq s_1 < s_2 < \cdots < s_q$ and define

$$z_j^0 = 2\pi s_j i, \tag{4.5}$$

(in which $i^2 = -1$), and

$$z_{q+j}^0 = -2\pi s_j i, \tag{4.6}$$

$(j = 1, 2, \cdots, q)$. For this choice,

$$F_0(z_j^0) = 0, \tag{4.7}$$

$$F_0'(z_j^0) \neq 0, \tag{4.8}$$

$(j = 1, 2, \cdots, n)$. Therefore the implicit function theorem in C^n applies to (4.4), ensuring the existence of a unique solution $z(\varepsilon)$ defined on some open interval in R about $\varepsilon^0 = 0$ and having $z(0) = z^0$. The solution is in fact real analytic relative to ε. It will now be argued that

$$\bar{z}_j(\varepsilon) = z_{q+j}(\varepsilon), \tag{4.9}$$

$(j = 1, 2, \cdots, q)$, for all ε in the aforementioned interval.

The permutation of the variables $\{z_1, z_2, \cdots, z_n\}$ which is defined by the interchanges $z_j \leftrightarrow z_{q+j}$, $(j = 1, 2, \cdots, q)$, induces the interchanges $p_{n-r}(z_j) \leftrightarrow p_{n-r}(z_{q+j})$, $(j = 1, 2, \cdots, q)$, $(r = 1, 2, \cdots, n)$. (See Remark 3.3.) Thus the permutation leaves the collection of equations (4.4) invariant. Moreover it is apparent that for ε fixed the solution set of (4.4) is invariant relative to complex conjugation. Hence $(\bar{z}_{q+1}(\varepsilon), \bar{z}_{q+2}(\varepsilon), \cdots, \bar{z}_{2q}(\varepsilon), \bar{z}_1(\varepsilon), \bar{z}_2(\varepsilon), \cdots, \bar{z}_q(\varepsilon))$ also is a solution to (4.4) equaling z^0 *at* ε^0 due to (4.5) and (4.6). Consequently (4.9) follows from the uniqueness aspect of the implicit function theorem. We have established the existence of the required solution (in fact a family of such solutions), $\gamma = (1/\varepsilon)$, $z = z(\varepsilon)$ for all $\varepsilon \neq 0$ appropriately near zero. This completes the proof of Theorem 4.1.

Remark 4.1. For $n = 2q$ even, the proof of Theorem 4.1 actually shows the existence of a q-parameter family of functions $k(\cdot)$ (with integer parameters $s_1 < s_2 < \cdots < s_q$), analytic about the origin, for which (4.1)-(4.2) have a solution $x(\cdot)$ when one sets $k = k(1/\gamma)$ in those equations and keeps $|\gamma|$ large. The limit of $k(1/\gamma)$ as $\gamma \to \pm\infty$ can be computed readily in terms of the s_j's using the formulas

$$z_j^0 = \frac{2\pi s_j i}{t_1 - t_0}, \tag{4.10}$$

$$z_{q+j}^0 = \frac{-2\pi s_j i}{t_1 - t_0}, \tag{4.11}$$

$(j = 1, 2, \cdots, q)$, along with (3.16).

Remark 4.2. Certainly if the bias parameter γ were restricted to be zero, then for some x^0, x^1 and f the boundary-value problem (4.1)-(4.2) would fail to have a solution no matter how k were chosen. On the other hand, for $n = 2q$ even, if there are integers $s_1 < s_2 < \cdots < s_q$ for which

$$x_r^1 - x_r^0 - (t_1 - t_0) \int_0^1 e^{-\sigma 2\pi s_j i} f_r(t_0 + \sigma(t_1 - t_0))\, d\sigma = 0, \tag{4.12}$$

$(r = 1, 2, \cdots, n)$, then the boundary-value problem will have a solution for some $k \in R^n$ independent of how $\gamma \in R$ is chosen. This conclusion follows from Lemma 3.1.

COROLLARY 4.1. *Let n be odd with the real n-vector valued function f and scalar function f_0 integrable on $[t_0, t_1]$. Then for each x^0 and x^1 in R^n there exist $k \in R^n$, $\gamma \in R$ and $k_0 \in R$ such that the boundary-value problem*

$$\dot{x} = A_n(k)x + \left\{\gamma + k_0 \int_{t_0}^{t} [x_1(\sigma) + f_0(\sigma)]\, d\sigma\right\} e_n + f(t), \tag{4.13}$$

$$x(t_0) = x^0, \qquad x(t_1) = x^1 \tag{4.14}$$

has a solution $x(\cdot)$.

Proof. Observe that the given boundary-value problem embeds in the even $(n+1)$-dimensional problem on $[t_0, t_1] \times R^1 \times R^n$,

$$\dot{x}^* = A_{n+1}(k^*)x^* + \gamma e_{n+1} + f^*(t), \tag{4.15}$$

$$x^*(t_0) = x^{*0}, \qquad x^*(t_1) = x^{*1}, \tag{4.16}$$

where $f^* = f_0 \times f$, $k^* = k \times k_0$, $x^{*0} = 0 \times x^0$ and $x^{*1} = 0 \times x^1$. The existence of a $k^* = k \times k_0 \in R^n \times R$ and a $\gamma \in R$ such that Equations (4.15)-(4.16) have a solution $x^*(\cdot) = x_0(\cdot) \times x(\cdot)$ follows from direct application of Theorem 4.1. This provides $k \in R^n$, $k_0 \in R$ and $\gamma \in R$ for which Equations (4.13)-(4.14) have the solution $x(\cdot)$ and Corollary 4.1 is proved.

Remark 4.3. As pointed out in Remark 4.1 the $k^* = k(\cdot) \times k_0(\cdot)$ arising from the proof of Corollary 4.1 is an integral parameter family of functions analytic about the origin. Solutions to (4.13)-(4.14) are assured when one sets $k = k(1/\gamma)$ and $k_0 = k_0(1/\gamma)$ in those equations and keeps $|\gamma|$ large. It is worth noting that although $k(1/\gamma)$ and $k_0(1/\gamma)$ have limits as $|\gamma| \to \infty$, the limit of the latter is not zero.

The steering that Corollary 4.1 says is possible relies on the sum of an admissible feedback controller together with an integral feedback term. The next theorem examines the question: Can the integral feedback be eliminated?

THEOREM 4.2. *Let $n = 2q+1$ be odd and $f: [t_0, t_1] \to R^n$ be integrable. Then there exist $k \in R^n$ and $\gamma \in R$ for which* (4.1)-(4.2) *have a solution $x(\cdot)$ with the possible exception of the two special cases defined for $n \geqq 3$ by the conditions*

(1) $x_1^1 - x_1^0 = 0$, $x_2^1 - x_2^0 \neq 0$ *and* $f_1(t) + x_2^0 = 0$ *for a.e.* $t \in [t_0, t_1]$.

(2) $x_1^1 - x_1^0 = 0$, $x_2^1 - x_2^0 \neq 0$ *and* $f_1(t) + x_2^1 = 0$ *for a.e.* $t \in [t_0, t_1]$.

A solution is assured to exist in the special case as well if $f_2 \in L_2(t_0, t_1)$ and respectively

$$\lim_{s \to \infty} e^s F_2^{\pm}(w + s + se^s i) = 0 \tag{1'-2'}$$

for all $w \in C$ in which $F_2^{\pm}(z)$ are defined to be the respective transforms of $f_2(t_0 + \sigma(t_1 - t_0))$ and $f_2(t_1 - \sigma(t_1 - t_0))$.

Proof. The argument is based on application of Lemma 3.2. Let $n = 2q+1$ be odd and f be integrable. In view of Remarks 3.4 and 3.5 we can take $[t_0, t_1] = [0, 1]$ and it is sufficient to show the existence of distinct $z_j \in C - R$ and $c \in R$ satisfying the equations

$$F_0(z_j) \sum_{r=1}^{n} [x_r + F_r(c)] p_{n-r}(c) = F_0(c) \sum_{r=1}^{n} [x_r + F_r(z_j)] p_{n-r}(z_j), \tag{4.17}$$

$(j = 1, 2, \cdots, 2q)$, with $\bar{z}_j = z_{q+j}$, $(j = 1, 2, \cdots, q)$. Such a $z = (z_1, z_2, \cdots, z_{2q})$ and c determine the appropriate

$$\gamma = -F_0(c)^{-1} \sum_{r=1}^{n} [x_r + F_r(c)] p_{n-r}(c). \tag{4.18}$$

For the lowest dimension, $n = 1$, the $c \in R$ can be selected arbitrarily, Equations (4.17) are inactive, and γ is determined from (4.18). (Recall that we defined $p_0 = 1$.) Now consider any fixed $n = 2q+1 > 1$.

In the simplified form to which the problem has been reduced the conditions (1) and (2) in the hypothesis of Theorem 4.2 appear as:

(1) $x_1 = 0$, $x_2 \neq 0$ and $f_1(\sigma) + x_2 = 0$, a.e. $\sigma \in [0, 1]$.

(2) $x_1 = 0$, $x_2 \neq 0$ and $f_1(\sigma) = 0$, a.e. $\sigma \in [0, 1]$.

The proof partitions the problem into cases.

Case 1. This is the case where one of conditions (1)-(3) of Lemma 3.3 is satisfied by x_1, x_2 and f_1.

A good deal of the analysis is based on the more convenient form of (4.17),

$$e^{-z_j} = 1 - G_j(z, c), \tag{4.19}$$

$(j = 1, 2, \cdots, 2q)$, where by definition

$$G_j(z, c) = \frac{F_0(c) \sum_{r=1}^{n} [x_r + F_r(z_j)] z_j p_{n-r}(z_j)}{\sum_{r=1}^{n} [x_r + F_r(c)] p_{n-r}(c)}. \tag{4.20}$$

Equations (4.19) are obtained from (4.17) by multiplying by z_j, evaluating the integral defining $F_0(z_j)$ and solving for the free term e^{-z_j}. The constraint that $z_j \in C - R$, $(j = 1, 2, \cdots, 2q)$, ensures the equivalence of (4.17) with (4.19) as long as the denominator term in $G_j(z, c)$ is not zero.

In working with (4.20) it will be necessary to sort out those homogeneous terms in the z_r's of highest degree. This leads to rewriting

$$\sum_{r=1}^{n} [x_r + F_r(c)]p_{n-r}(c) = [x_1 + F_1(c)] \prod_{1}^{2q} z_r + \delta_n(z, c) \tag{4.21}$$

where

$$\delta_n(z, c) = \sum_{2}^{n} [x_r + F_r(c)]p_{n-r}(c) \tag{4.22}$$

and similarly

$$\begin{aligned} &F_0(c) \sum_{r=1}^{n} [x_r + F_r(z_j)]z_j p_{n-r}(z_j) \\ &= F_0(c)\left\{[(cx_1 - x_2) + cF_1(z_j) - F_2(z_j)] \prod_{1}^{2q} z_r + \delta_j(z, c)\right\} \end{aligned} \tag{4.23}$$

where

$$\begin{aligned} \delta_j(z, c) = [x_2 + F_2(z_j)]&\left[z_j p_{n-2}(z_j) + \prod_{1}^{2q} z_r \right] \\ &+ \sum_{3}^{n} [x_r + F_r(z_j)]z_j p_{n-r}(z_j), \end{aligned} \tag{4.24}$$

$(j = 1, 2, \cdots, 2q)$. Substitution of (4.21) and (4.23) into (4.20) gives

$$G_j(z, c) = F_0(c) \frac{[cx_1 - x_2 + cF_1(z_j) - F_2(z_j)] \prod_1^{2q} z_r + \delta_j(z, c)}{[x_1 + F_1(c) \prod_1^{2q} z_r + \delta_n(z, c)} \tag{4.25}$$

$(j = 1, 2, \cdots, 2q)$.

Temporarily deferring further direct analysis of (4.19), we switch our attention to the system

$$e^{-w_j} = 1 - G_j(w + s2\pi i\alpha, c), \tag{4.26}$$

$(j = 1, 2, \cdots, 2q)$, having dependent variables $w = (w_1, w_2, \cdots, w_{2q}) \in C^{2q}$ and $c \in R$. The vector $\alpha \in R^{2q}$ has its first q coordinates ones and the remaining q coordinates are minus ones. The $s \in R$ is treated as a parameter. Existence of a family of solutions w, c to (4.26) for all s with $|s|$ large will now be argued.

According to Lemma 3.3 there exists a $c_0 \in R$ and a $w_0 \in C$ satisfying the equation

$$e^{-w_0} = \frac{x_1 e^{-c_0} + x_2 F_0(c_0) + F_1(c_0)}{x_1 + F_1(c_0)}. \tag{4.27}$$

Choose any integers s_r, $(r = 1, 2, \cdots, q)$, satisfying $0 \leqq s_1 < s_2 < \cdots < s_q$ and define

$$w_j^0 = w_0 + 2\pi s_j i, \tag{4.28}$$

$$w_{q+j}^0 = \bar{w}_0 - 2\pi s_j i, \tag{4.29}$$

$(j = 1, 2, \cdots, q)$. To insure that the coordinates of $w^0 = (w_1^0, w_2^0, \cdots, w_{2q}^0)$ are distinct, w_0 is taken to be the solution of (4.27) with the smallest possible imaginary part. We shall show that the implicit function theorem applies to (4.26), producing a solution about $(w^0, c_0) \in C^{2q} \times R$ for $|s|$ large.

By recalling Remark 3.3 we carefully note that $z_j p_{n-2}(z_j) + \prod_1^{2q} z_r$ and the $z_j p_{n-r}(z_j)$, $(r = 3, 4, \cdots, n)$ of (4.24) as well as the $p_{n-r}(c)$, $(r = 2, 3, \cdots, n)$, of (4.22) are sums

of homogeneous terms in the z_r's of degree $2q-1$ or less. This observation combined with application of Theorem 3.2 to Equations (4.22) and (4.24) shows that

$$s^{-2q}\delta_j(w+s2\pi i\alpha, c)\to 0 \tag{4.30}$$

as $|s|\to\infty$ for all $(w, c)\in C^{2q}\times R$, $(j=1, 2, \cdots, n)$. It follows that

$$\lim_{|s|\to\infty} G_j(w+s2\pi i\alpha, c)=\frac{(cx_1-x_2)F_0(c)}{x_1+F_1(c)}, \tag{4.31}$$

$(j=1, 2, \cdots, 2q)$. Therefore it is possible to define a function $H_j(w, c, \varepsilon)$ which is continuous on a neighborhood of $(w^0, c_0, 0)\in C^{2q}\times R\times R$ for which

$$H_j(w, c, \varepsilon)=1-G_j\left(w+\frac{2\pi\alpha i}{\varepsilon}, c\right), \tag{4.32}$$

$(j=1, 2, \cdots, 2q)$, when $\varepsilon\neq 0$. Since the limit functions are continuous and are limits of functions analytic relative to (w, c) the convergence is uniform on compact sets and the limits are therefore analytic relative to those variables. In particular the H_j are once continuously differentiable relative to w about $(w^0, c_0, 0)$ with

$$H_j(w, c, 0)=1-\frac{(cx_1-x_2)F_0(c)}{x_1+F_1(c)}=\frac{e^{-c}x_1+x_2F_0(c)+F_1(c)}{x_1+F_1(c)} \tag{4.33}$$

and

$$\frac{\partial}{\partial w_r}H_j(w, c, \varepsilon)=0 \tag{4.34}$$

at $(w^0, c_0, 0)$ for $(j=1, 2, \cdots, 2q)$, $(r=1, 2, \cdots, 2q)$. Together, (4.27)-(4.29) imply that

$$e^{-w_j^0}=H_j(w^0, c_0, 0), \tag{4.35}$$

$(j=1, 2, \cdots, 2q)$. The implicit function theorem can now be applied to conclude that there exists a unique solution $\tilde{w}(\varepsilon)$ to the system

$$e^{-w_j}=H_j(w, c_0, \varepsilon), \tag{4.36}$$

$(j=1, 2, \cdots, 2q)$, defined and continuous for all ε in an open interval about the origin in R and satisfying $\tilde{w}(0)=w^0$. Due to the relationship (4.32) between H_j and G_j the conclusion translates into the statement saying that there exists a unique solution $w(s)=\tilde{w}(1/s)$ to (4.26) defined and continuous for all s with $|s|$ large and satisfying the condition that $w(s)\to w^0$ as $|s|\to\infty$.

Since the collection of equations (4.26) is invariant relative to the interchanges $w_j \leftrightarrow w_{q+j}$, $(j=1, 2, \cdots, q)$, and its solution set is invariant relative to complex conjugation it is apparent that $(\bar{w}_{q+1}, \bar{w}_{q+2}, \cdots, \bar{w}_{2q}, \bar{w}_1, \bar{w}_2, \cdots, \bar{w}_q)(s)$ is another solution to (4.26). However (4.28)-(4.29) imply that both solutions have the limiting value w^0 as $|s|\to\infty$ and the uniqueness leads to the conclusion that

$$\bar{w}_j(s)=w_{q+j}(s), \tag{4.37}$$

$(j=1, 2, \cdots, q)$, for $|s|$ large.

Returning our attention to (4.19), we now define

$$z(s)=w(s)+s2\pi i\alpha \tag{4.38}$$

with s restricted to integer values having $|s|$ large. For each such s we claim that $(z(s), c_0)$ provides the required solution to (4.19). Since $(w(s), c_0)$ satisfies (4.26), it

follows from (4.38) that

$$e^{-z_j(s)} = e^{-w_j(s)} = 1 - G_j(w(s) + s2\pi i\alpha, c_0) = 1 - G_j(z(s), c_0), \tag{4.39}$$

$(j = 1, 2, \cdots, 2q)$, as desired. Moreover (4.37) and (4.38) imply that

$$\bar{z}_j(s) = z_{q+j}(s), \tag{4.40}$$

$(j = 1, 2, \cdots, q)$. The requirement of Lemma 3.2 that the coordinates of $(z(s), c_0)$ be distinct is met since for $|s|$ large $w(s)$ is near w^0 whose coordinates, by their definition, are distinct. Moreover, it is clear from (4.38) that for $|s|$ large the condition that $z_j(s) \in C - R$, $(j = 1, 2, \cdots, 2q)$ is satisfied. Thus an appropriate solution to (4.17) has been shown to exist and the conclusion of Theorem 4.2 holds for Case 1.

Case 2. This is the case in which $x_1 = 0$, $x_2 \neq 0$ and $f_1(\sigma) = -x_2$ for a.e. $\sigma \in [0, 1]$.

Again the approach taken is based on proving the existence of an appropriate solution to (4.17). To obtain a more convenient form of the equations, we split off a term from the second sum in (4.17) and insert the assumed values of x_1 and f_1 to get

$$\begin{aligned} &F_0(z_j) \sum_{r=1}^{n} [x_r + F_r(c)] p_{n-r}(c) \\ &\quad = -x_2 F_0(c) F_0(z_j) p_{n-1}(z_j) + F_0(c) \sum_{r=2}^{n} [x_r + F_r(z_j)] p_{n-r}(z_j), \end{aligned} \tag{4.41}$$

$(j = 1, 2, \cdots, 2q)$. Multiplication of (4.41) by z_j, evaluation of $F_0(z_j)$ and solution of the resultant equation for the free term e^{-z_j} leads to the equivalent equations

$$e^{-z_j} = 1 - \frac{F_0(c) \sum_{r=2}^{n} [x_r + F_r(z_j)] z_j p_{n-r}(z_j)}{\sum_{r=1}^{n} [x_r + F_r(c)] p_{n-r}(c) + x_2 F_0(c) p_{n-1}(z_j)}, \tag{4.42}$$

$(j = 1, 2, \cdots, 2q)$. By writing the right-hand side of (4.42) over a common denominator, splitting off the terms containing x_2 in the resultant numerator and applying the relation

$$p_{n-1}(z_j) - z_j p_{n-2}(z_j) = c \sum_{r=1}^{2q} \prod_{m \neq r} z_m + \prod_{1}^{2q} z_r, \tag{4.43}$$

which can easily be derived, a long but elementary calculation leads to the equivalent form of (4.42),

$$e^{-z_j} = G_j(z, c), \tag{4.44}$$

$(j = 1, 2, \cdots, 2q)$, where this time

$$G_j(z, c) = \frac{h(z, c) + x_2 \sum_{r=1}^{2q} \prod_{m \neq r} z_m + \delta_j(z, c)}{h(z, c) + x_2 \prod_{r \neq j} z_r} \tag{4.45}$$

in which is defined

$$h(z, c) = F_0(c)^{-1} \sum_{1}^{n} [x_r + F_r(c)] c^{-1} p_{n-r}(c), \tag{4.46}$$

$$\begin{aligned} \delta_j(z, c) = {} & [x_2 + F_2(z_j)] c^{-1} \prod_{1}^{2q} z_r + F_2(z_j) \sum_{r \neq j} \prod_{m \neq r} z_m \\ & - \sum_{3}^{n} [x_r + F_r(z_j)] c^{-1} z_j p_{n-r}(z_j). \end{aligned} \tag{4.47}$$

As in Case 1 the main argument centers about an auxiliary system, here

$$e^{-w_j} = e^s G_j(w + s\beta + se^s i\alpha, c(s)), \tag{4.48}$$

$(j = 1, 2, \cdots, 2q)$. The symbols w, w_j and α denote the same kinds of objects as in the analysis of Case 1 and β is the $2q$-vector all of whose coordinates are ones. The parameter c is made dependent on s according to the equation

$$c(s) = -s^2 e^{2s}, \tag{4.49}$$

for $s > 0$. Existence of an appropriate family of solutions to (4.48) for all large s will now be argued.

Let w_0 be a solution to the equation

$$e^{-w_0} = -qi. \tag{4.50}$$

Choose any integers s_r, $(r = 1, 2, \cdots, q)$, satisfying $0 \leqq s_1 < s_2 < \cdots < s_q$ and define

$$w_j^0 = w_0 + 2\pi s_j i, \tag{4.51}$$

$$w_{q+j}^0 = \bar{w}_0 - 2\pi s_j i, \tag{4.52}$$

$(j = 1, 2, \cdots, q)$.

To insure that the coordinates of $w^0 = (w_1^0, w_2^0, \cdots, w_{2q}^0)$ are distinct, w_0 is taken to be the solution of (4.50) with the smallest positive imaginary part. We shall show that the dominant terms in (4.48) as $s \to \infty$ contributed by G_j arise from the two terms in (4.45) containing the products.

Keep in mind that $p_{n-r}(c)$ is a finite sum of homogeneous terms of degree $n - r$ in the coordinates of $w + s\beta + se^s i\alpha$. Then it is evident that for $r = 1$ the term

$$\frac{e^s F_0(c)^{-1}[x_r + F_r(c)]c^{-1} p_{n-r}(c)}{(se^s)^{2q-1}} \tag{4.53}$$

is convergent to zero along $c(s)$ as $s \to \infty$ since $x_1 = 0$ and $F_1(c) = -x_2 F_0(c)$. Moreover, with $f_2 \in L_2$, an application of Schwarz's inequality shows that for $r = 2$ the term given by (4.53) is also convergent to zero as $s \to \infty$. The computation

$$F_0(c)^{-1}[x_r + F_r(c)]c^{-1} = [e^c - 1]^{-1}\left[e^c x_r + \int_0^1 e^{(1-\sigma)c} f_r(\sigma)\, d\sigma\right] \tag{4.54}$$

shows that this factor is bounded along $c(s)$ on $0 \leqq s < \infty$ due to the assumption that $f_r \in L_1$ and thus (4.53) is convergent to zero as $s \to \infty$ for $r = 3, 4, \cdots, n$ as well. This analysis of (4.53) proves that

$$\lim_{s \to \infty} \frac{e^s h(w + s\beta + se^s i\alpha, c(s))}{(se^s)^{2q-1}} = 0 \tag{4.55}$$

for all $w \in C^{2q}$.

Next we want to show that

$$\lim_{s \to \infty} \frac{e^s \delta_j(w + s\beta + se^s i\alpha, c(s))}{(se^s)^{2q-1}} = 0, \tag{4.56}$$

$(j = 1, 2, \cdots, 2q)$, for all $w \in C^{2q}$.

First note that by the dominated convergence theorem,

$$F_r(w + s\beta + se^s i\alpha) \to 0 \tag{4.57}$$

as $s \to \infty$ for all $w \in C^{2q}$, $(r = 1, 2, \cdots, n)$. The dominant term as $s \to \infty$ for the sum

$$\sum_{r \neq j} \prod_{m \neq r} [w + s + se^s i\alpha]_m \tag{4.58}$$

clearly is

$$\sum_{r \neq j} \prod_{m \neq r} [se^s i\alpha]_m = (se^s i)^{2q-1} \chi(j) \tag{4.59}$$

in which

$$\chi(j) = \sum_{r \neq j} \prod_{m \neq r} \alpha_m \tag{4.60}$$

has the value $(-1)^{q-1}$ for $j \in \{1, 2, \cdots, q\}$ and the value $(-1)^q$ for $j \in \{q+1, q+2, \cdots, 2q\}$. It should now be evident from (4.47), (4.57) and condition (1′) that (4.56) is valid.

To evaluate the contribution of the dominant terms from G_j in (4.48) we first compute

$$\sum_{r=1}^{2q} \prod_{m \neq r} [se^s i\alpha]_m = (se^s i)^{2q-1} \sum_{r=1}^{2q} \prod_{m \neq r} \alpha_m = 0 \tag{4.61}$$

which shows that the dominant term as $s \to \infty$ for the sum

$$\sum_{r=1}^{2q} \prod_{m \neq r} [w + s + se^s i\alpha]_m \tag{4.62}$$

is

$$\begin{aligned} \sum_{r=1}^{2q} \sum_{k \neq r} s \prod_{m \neq k,r} [se^s i\alpha]_m &= s(se^s i)^{2q-2} \sum_{r=1} \sum_{k \neq r} \prod_{m \neq k,r} \alpha_m \\ &= s^{2q-1} e^{(2q-2)s} (-1)^{q-1} [2C_2^q (-1)^q + q^2 (-1)^{q-1}] \\ &= s^{2q-1} e^{(2q-2)s} (-1)^{q-1} q (-1)^{q-1} = q s^{2q-1} e^{(2q-2)s}. \end{aligned} \tag{4.63}$$

One can easily check that the dominant term of the product $\prod_{r \neq j} [w + s\beta + se^s i\alpha]_r$ is

$$\prod_{r \neq j} [se^s i\alpha]_r = (se^s i)^{2q-1} \chi(j). \tag{4.64}$$

From (4.45), (4.55), (4.56), (4.63) and (4.64) we see that

$$\lim_{s \to \infty} e^s G_j(w + s\beta + se^s i\alpha, c(s)) = \lim_{s \to \infty} \frac{e^s q s^{2q-1} e^{(2q-2)s}}{(se^s i)^{2q-1} \chi(j)} = \chi(j)(-1)^q qi, \tag{4.65}$$

$(j = 1, 2, \cdots, 2q)$ at w^0 and that the first limit in (4.65) defines an analytic function of w on a neighborhood of w^0. Applying this result, we obtain functions $H_j(w, \varepsilon)$ continuous on a neighborhood of $(w^0, 0) \in C^{2q} \times R$ for which

$$H_j(w, \varepsilon) = e^s G_j(w + s\beta + se^s i\alpha, c(s))\Big|_{s = \varepsilon^{-2}}, \tag{4.66}$$

$(j = 1, 2, \cdots, 2q)$, $\varepsilon \neq 0$. Moreover the H_j are analytic relative to w and in particular are once continuously differentiable relative to that variable about $(w^0, 0)$. The evaluation of the limit that occurred in (4.65) gives

$$H_j(w^0, 0) = -qi, \tag{4.67}$$

$$H_{q+j}(w^0, 0) = qi, \tag{4.68}$$

$(j = 1, 2, \cdots, q)$. By a careful examination of (4.45), (4.46) and (4.47), using the fact

that differentiation lowers the degrees of a homogeneous term in the z_r's, the condition (1') as well as previously encountered arguments, one can check that

$$\frac{\partial}{\partial w_r} H_j(w, \varepsilon) = 0 \tag{4.69}$$

at $(w^0, 0)$ for $(j = 1, 2, \cdots, 2q)$ and $(r = 1, 2, \cdots, 2q)$. As a consequence of (4.50), (4.51), (4.67) and (4.68) we see that

$$e^{-w_j^0} = H_j(w^0, 0), \tag{4.70}$$

$(j = 1, 2, \cdots, 2q)$. In light of (4.69) and (4.70) the implicit function theorem can now be applied to conclude that the system

$$e^{-w_j} = H_j(w, \varepsilon), \tag{4.71}$$

$(j = 1, 2, \cdots, 2q)$, has a unique complex $2q$-vector solution $\tilde{w}(\varepsilon)$ defined and continuous on an open interval about the origin in R and satisfying $\tilde{w}(0) = w^0$. Due to the relation (4.66) this conclusion translates into the statement that there exists a unique solution $w(s) = \tilde{w}(1/s)$ to (4.48) defined and continuous for all large s and having $w(s) \to w^0$ as $s \to \infty$. An argument similar to the one appearing in the treatment of Case 1 shows that

$$\bar{w}_j(s) = w_{q+j}(s), \tag{4.72}$$

$(j = 1, 2, \cdots, q)$, for all s large.

Returning our attention to (4.44) in which we set $c = c(s)$, we can now define

$$z(s) = w(s) + s\beta + se^s i\alpha \tag{4.73}$$

with $se^s/2\pi$ restricted to large integer values. For each such s we claim that $(z(s), c(s))$ provides the required solution to (4.44). Since $(w(s), c(s))$ satisfies (4.48), it follows with the aid of (4.73) that

$$e^{-z_j(s)} = e^{-s}e^{w_j(s)} = G_j(w(s) + s\beta + se^s i\alpha, c(s)) = G_j(z(s), c(s)), \tag{4.74}$$

$(j = 1, 2, \cdots, 2q)$, as desired. Moreover (4.72) and (4.73) give us

$$\bar{z}_j(s) = z_{q+j}(s), \tag{4.75}$$

$(j = 1, 2, \cdots, q)$. It is clear that the coordinates of $(z(s), c(s))$ are distinct for s large and that $z_j(s) \in C - R$, $(j = 1, 2, \cdots, 2q)$. Thus an appropriate solution to (4.17) has been shown to exist and the conclusion of Theorem 4.2 holds for Case 2.

Case 3. This is the case in which $x_1 = 0$, $x_2 \neq 0$ and $f_1(\sigma) = 0$ for a.e. $\sigma \in [0, 1]$.

Rather than apply Lemma 3.2 directly we transform the problem into one already covered in Case 2. The boundary-value problem under consideration is (3.18)-(3.19) in which $x^1 = 0$, $x_1^0 = 0$, $x_2^0 \neq 0$ and $f_1(t) = 0$ for a.e. $t \in [0, 1]$. Let Λ be the $n \times n$ diagonal matrix with $\Lambda_{jj} = (-1)^{j+1}$, $(j = 1, 2, \cdots, n)$. Consider the change of variable $x^*(t) = \Lambda[x(1-t) - x^0]$. By utilizing the relations $-\Lambda A_n(k)\Lambda^{-1} = A_n(\Lambda k)$ and $\Lambda e_n = e_n$ one can check that the proposed change of variable transforms the boundary-value problem under discussion into

$$\dot{x}^* = A_n(k^*)x^* + \gamma^* e_n + f^*(t), \tag{4.76}$$

$$x^*(0) = -\Lambda x^0, \qquad x^*(1) = 0 \tag{4.77}$$

in which $k^* = \Lambda k$, $f^*(t) = -\Lambda[A_n(0)x^0 + f(1-t)]$ and $\gamma^* = \gamma - [k_n x_1^0 + k_{n-1}x_2^0 + \cdots + k_1 x_n^0]$. Observe that $(-\Lambda x^0)_1 = -x_1^0 = -x_1 = 0$, $(-\Lambda x^0)_2 = x_2^0 =$

$x_2 \neq 0$ and $f_1^*(t) = -x_2^0 - f_1(t) = -x_2^0 = -x_2$ for a.e. $t \in [0, 1]$. Furthermore, $f_2^*(t) = x_3^0 - f_2(1-t)$. Since $f_2(t)$ is assumed to satisfy the condition (2'), it follows that $f_2(1-t)$ satisfies the condition imposed by (1') and because of Theorem 3.3 applied to the constant function x_3^0 it is evident that $f_2^*(t)$ satisfies the condition (1'). Thus the problem (4.76)-(4.77) has a solution k^*, γ^* since it has been verified to be of the type falling under Case 2 and obviously this in turn implies that each problem of the kind defined by Case 3 likewise is solvable. This concludes the proof for Case 3.

Case 4. This is the case where $x_1 = 0$, $x_2 = 0$ and $f_1(\sigma) = 0$ for a.e. $\sigma \in [0, 1]$.

Consider the problem with perturbed data $x_1 = \varepsilon_1$ and $x_2 = \varepsilon_1$ with the parameter ε_1 restricted to an open interval containing the origin. Choose any $c_0 \in R$ such that

$$e^{-c_0} + F_0(c_0) \neq 0. \tag{4.78}$$

(There are many such numbers c^0.) For $\varepsilon_1 \neq 0$ fixed the perturbed problem falls under the purview of Case 1 and moreover has the same limit equation (4.35),

$$\begin{aligned} e^{-w_j^0} = H_j(w^0, c_0, 0) &= \frac{e^{-c_0}x_1 + x_2F_0(c_0) + F_1(c_0)}{x_1 + F_1(c_0)} \\ &= \frac{e^{-c_0}\varepsilon_1 + \varepsilon_1 F_0(c_0)}{\varepsilon_1} = e^{-c_0} + F_0(c_0), \end{aligned} \tag{4.79}$$

$(j = 1, 2, \cdots, 2q)$, independent of $\varepsilon_1 \neq 0$. Therefore the w_0 and consequently w^0 that appears in the proof of Case 1 is independent of ε_1. In fact the function H_j defined by (4.32), now viewed as a function of $(w, c, \varepsilon, \varepsilon_1)$, can be extended by continuity so that it is defined and continuous about the point $(w^0, c_0, 0, 0)$ and again once continuously differentiable relative to w. The solution $w(s, \varepsilon_1)$ obtained from application of the implicit function theorem then has a limit, $\lim_{\varepsilon_1 \to 0} w(s, \varepsilon_1) = w(s, 0)$, which when substituted into (4.38) provides the required solution $(z(s), c_0)$, with $|s|$ large, to the unperturbed system, i.e., to the system (4.17) with $x_1 = x_2 = \varepsilon_1 = 0$. This concludes the proof of Theorem 4.2 for Case 4.

Case 5. In this final case $x_1 = x_2 = 0$ and for some measurable set $E \subset [0, 1]$ with $\mu(E) > 0$, $f_1(\sigma) \neq 0$ for all $\sigma \in E$.

For this case it is easy to check that by working with an appropriate subset of E there is no loss in assuming that $f_1(\sigma) + \varepsilon \neq 0$ for all $\sigma \in E$ and all ε with $|\varepsilon|$ small. Hence there exists a $c_0 \in R$ for which

$$F_1(c_0) \neq 0. \tag{4.80}$$

Now consider the problem with perturbed data $x_1 = 0$ and $x_2 = \varepsilon_1$. For fixed ε_1 with $|\varepsilon_1|$ small the perturbed problem is of the type considered under Case 1 as long as $\varepsilon_1 \neq 0$ since $f_1(\sigma) \neq 0$ and $f_1(\sigma) + x_2 \neq 0$ on E and $x_2 \neq 0$. The limit system (4.35) for the perturbed problem is

$$e^{-w_j^0} = H_j(w^0, c_0, 0) = \frac{e^{-c_0}x_1 + x_2F_0(c_0) + F_1(c_0)}{x_1 + F_1(c_0)} = \frac{\varepsilon_1 F_0(c_0) + F_1(c_0)}{F_1(c_0)}, \tag{4.81}$$

$(j = 1, 2, \cdots, 2q)$, and as a consequence of inequality (4.80), it follows that for $|\varepsilon_1|$ small the solution $w(s, \varepsilon_1)$ arising out of the analysis of Case 1 has a limit, $\lim_{\varepsilon_1 \to 0} \omega(s, \varepsilon_1) = w(s, 0)$, which as in the argument for Case 4 provides the required solution $(z(s), c_0)$, with $|s|$ large, to the unperturbed system. The details are omitted since they are very much a repetition of those for Case 4. This ends the analysis of Case 5.

Cases 1-5 are mutually exclusive and exhaustive. Theorem 4.2 is now established.

Remark 4.4. Theorem 3.3, when applied to the functions $g(\sigma)=f_2(t_0+\sigma(t_1-t_0))$ and $g(\sigma)=f_2(t_1-\sigma(t_1-t_0))$, provides sufficient conditions ensuring (1′) and (2′), respectively, of Theorem 4.2.

COROLLARY 4.2. *For integrable f, if $n=1$, n is even or the functions $g(\sigma)=f_2(t_0+\sigma(t_1-t_0))$ and $g(\sigma)=f_2(t_1-\sigma(t_1-t_0))$ are in $L_2(0,1)$ and each satisfy* (3.41) *or* (3.42) *then there exist $k\in R^n$ and $\gamma\in R$ such that Equations* (4.1)-(4.2) *have a solution* $x(\cdot)$.

Proof. The conclusions are direct consequences of Theorems 4.1, 4.2 and 3.3.

The stage is now set for dealing with the general linear equation.

THEOREM 4.3. *Equation* (2.1) *is feedback controllable on* $[t_0, t_1]$ *if its coefficients satisfy* (1.2) *and one of the following conditions is met*:

(1) $f\in L_1^n(t_0, t_1)$ *and* $n=1$ *or* n *is even.*

(2) $f\in L_2^n(t_0, t_1)$ *and for each of its coordinate functions f_r the corresponding functions $g(\sigma)=f_r(t_0+\sigma(t_1-t_0))$ and $g(\sigma)=f_r(t_1-\sigma(t_1-t_0))$ each satisfy* (3.41) *or* (3.42).

(3) $f\in BV[t_0, t_1]$.

Proof. A theorem due to Heymann [1] (or see [4, p. 279]) states that if the controllability matrix has full rank, then for b_i any nonzero column of B there exists a real matrix K_i such that the closed loop scalar control system

$$\dot{x}=(A+BK)x+vb_i+f(t) \tag{4.82}$$

corresponding to the controller

$$u=K_ix+ve_i \tag{4.83}$$

has its controllability matrix of full rank. Hence it is sufficient to prove Theorem 4.3 for the equation

$$\dot{x}=Ax+ub+f(t) \tag{4.84}$$

in which u is a scalar control variable and b is an n-vector. For a controllable scalar control equation, (4.84), there exists a real nonsingular linear change of state variables $x=Pz$ which transforms the system into the canonical form

$$\dot{z}=A_n(k_0)z+ue_n+P^{-1}f(t) \tag{4.85}$$

for an appropriate $k_0\in R^n$, [4, p. 276]. Therefore, since an f of any of the types occurring in conditions (1)-(3) of Theorem 4.3 remains of the same type under multiplication by P^{-1}, it follows that it is sufficient to prove Theorem 4.3 for the equation

$$\dot{x}=A_n(0)x+ue_n+f(t). \tag{4.86}$$

Since Corollary 4.2 applies to (4.86), the proof of Theorem 4.3 is finished.

Remark 4.5. The approach taken in this paper relied heavily on overpowering the complexity in (3.21) and (3.27) contributed by the forcing function f by exploiting the absence of any bounds on K and v. Perhaps nothing more than the limitation to this approach is showing through conditions (1′)-(2′) of Theorem 4.2. One can show that (1′)-(2′) fails for Weierstrass's continuous but nowhere differentiable function. (See [5, p. 148] for its definition.) The feedback controllability question is left open in odd dimension $n\geqq 3$ for such forcing functions for the special end states discussed by Theorem 4.2.

REFERENCES

[1] M. HEYMANN, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 748-749.

[2] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (1960), pp. 102–119.
[3] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.
[4] D. L. LUKES, *Differential Equations: Classical to Controlled*, Academic Press, New York, 1982.
[5] E. R. PHILLIPS, *An Introduction to Analysis and Integration Theory*, Intext Educational, Scranton, PA, 1971.
[6] W. RUDIN, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 6, November 1985

006

METRIC PROJECTIONS AND THE GRADIENT PROJECTION METHOD IN BANACH SPACES*

R. R. PHELPS†

Abstract. Let P_C denote the metric projection onto a closed convex subset C of a smooth, rotund and reflexive Banach space E. With an additional hypothesis on E, it is shown that P_C has directional derivatives in every direction at every point of C (a result which is well known for Hilbert space). This is used to prove that in such a reflexive space any cluster point for the gradient projection method (using Cauchy's steplength) is a constrained stationary point.

The gradient projection method is a well-known technique for constrained optimization of a real-valued C^1 function on Hilbert space. In this note it is shown that a useful tool for this method (the existence of directional derivatives for the metric projection) can be extended to a class of Banach spaces containing, for instance, the L^p spaces, $1<p<\infty$. As an application, it is easy to extend to such spaces a Hilbert space result of McCormick and Tapia [7], which shows that any cluster point for the gradient projection sequence is a constrained stationary point. We first establish some notation and recall several definitions and relevant facts. Details may be found in Day [2] or Diestel [3].

Throughout , C will denote a nonempty closed convex subset of a reflexive real Banach space E. The *metric projection* P_c (or simply P) of E onto C is defined for $x\in E$ by

$$\|x-Px\|=\inf\{\|x-y\|: y\in C\};$$

this will be single-valued if E is rotund (strictly convex). A *duality mapping* is any mapping $J: E\to E^*$ which satisfies, for each $x\in E$,

$$\|J(x)\|=\|x\| \quad \text{and} \quad \langle J(x), x\rangle=\|x\|^2.$$

By the Hahn–Banach theorem such a mapping always exists, and it will be uniquely determined if and only if E is smooth. Smoothness and the reflexivity of E imply that J is onto; it will be one-one if E if rotund. Thus, if E is smooth and rotund, then J is homogeneous and J^{-1} exists; the latter is characterized by

$$\langle u^*, J^{-1}u^*\rangle=\|u^*\|^2 \quad \text{and} \quad \|J^{-1}u^*\|=\|u^*\|, \qquad u^*\in E^*.$$

In Hilbert space, of course, J is just the canonical mapping which identifies E with E^*.

The space E is said to satisfy *property* (H) provided

(H) $$\|x_n-x\|\to 0 \quad \text{whenever } \|x_n\|\to\|x\| \text{ and } x_n\to x \text{ weakly.}$$

This well-known property of Hilbert space is satisfied by any uniformly rotund (or locally uniformly rotund) Banach space; hence it is satisfied by any L^p space with $1<p<\infty$. There are two useful consequences of property (H), both of which are straightforward to prove by exploiting the relative weak compactness of bounded subsets of a reflexive space and the weak lower semicontinuity of the norm.

LEMMA. 1. *If E is a reflexive, smooth and rotund Banach space which satisfies property* (H), *then both P_C and J^{-1} are continuous maps.*

* Received by the editors October 11, 1983, and in revised form August 9, 1984. This research was supported in part by a grant from the National Science Foundation.

† Department of Mathematics, University of Washington, Seattle, Washington, 98195.

The duality map J makes it possible to extend a well-known Hilbert space characterization of P_C to the present context. We refer to (*) below as the *defining inequality.*

LEMMA 2. *Suppose that E is a smooth, reflexive and rotund Banach space and that C is a nonempty closed and convex subset of E. Then for any $x \in E$ the point $z = P_C x$ is the unique element $z \in Z$ which satisfies*

$$(*) \qquad \langle J(x-z), u-z\rangle \leqq 0 \quad \textit{for all } u \in C.$$

Proof. Clearly, $z = Pz$ satisfies (*) if $x = Px \in C$, so suppose $x \in E \setminus C$. By the separation theorem, applied to C and the ball B of radius $\|x - Px\|$ centered at x, there exists $u^* \in E^*$, $\|u^*\| = 1$, such that

$$\sup\{\langle u^*, u\rangle : u \in C\} = \langle u^*, Px\rangle = \inf\{\langle u^*, u\rangle : u \in B\} = \langle u^*, x\rangle - \|x - Px\|.$$

Thus, $\langle u^*, x - Px\rangle = \|x - Px\|$ (so that $\|x - Px\| u^* = J(x - Px)$) and $\langle u^*, u - Px\rangle \leqq 0$ for all $u \in C$, as required. Suppose, now, that z is in C and satisfies (*). If $z = x$, then $z = Px$; otherwise, we have for $u \in C$,

$$0 \geqq \langle J(x-z), u-z\rangle = \langle J(x-z), x-z+u-x\rangle = \|x-z\|^2 + \langle J(x-z), u-x\rangle$$

so

$$\|x-z\|^2 \leqq \langle J(x-z), x-u\rangle \leqq \|x-z\| \cdot \|x-u\|.$$

and hence $\|x-z\| \leqq \|x-u\|$ for all $u \in C$, that is, $z = Px$.

We need one further definition: If $x \in C$, the *support cone* $S_C(x)$ to C at x is the closure of the convex cone $\cup\{\lambda(C-x) : \lambda > 0\}$. The set $S_C(x)$ (or simply $S(x)$) is clearly a closed convex cone with vertex 0 and is the smallest such cone S whose translate $x + S$ has vertex x and contains C. Its utility in the present situation appears in our main lemma, which shows that $P_{S(x)}$ is the directional derivative of P_C at points x of C. This is well-known in Hilbert space (see, for instance, McCormick-Tapia [7] and Zarantonello [10, p. 300]) and it is easily extended below to more general spaces. Mignot [8, Thm. 2.1] (see also Haraux [5]) has proved a result for Hilbert space (using a nonsymmetric inner product) which yields this lemma as a special case, but we do not see how to extend his result to the present context.

LEMMA 3. *Suppose that the reflexive Banach space E is smooth, rotund and has property* (H). *If C is a nonempty closed convex subset of E, then for each $x \in C$ and any $y \in E$ we have*

$$P_C(x+ty) = x + tP_S y + o(t), \qquad t > 0,$$

where P_S is the metric projection of E onto the cone $S_C(x)$. (Thus, $P_S y$ is the directional derivative of P_C at x in the direction y.)

Proof. There is no loss in generality in assuming that $x = 0$, so we want to prove that $t^{-1}P_C(ty) \to P_S y$ as $t \to 0^+$. If we apply the defining inequality for P_C to ty, $t > 0$, and use the positive homogeneity of the duality mapping J, we can divide through by t and conclude that

$$\langle J(y - t^{-1}P_C(ty)), v - t^{-1}P_C(ty)\rangle \leqq 0$$

whenever $v \in t^{-1}C$. By the defining inequality again, this implies that $P_{t^{-1}C}y = t^{-1}P_C(ty)$. Since $0 \in t^{-1}C$ we have

$$\|y - t^{-1}P_C(ty)\| \leqq \|y\|$$

so that $\|t^{-1}P_C(ty)\| \leqq 2\|y\|$ for all $t > 0$.

Now to prove convergence as $t \to 0^+$ it suffices to prove that $w_n = P_{t_n^{-1}C} y$ converges to $P_S y$ for each positive sequence $\{t_n\}$ decreasing to 0. By boundedness of such a sequence and reflexivity of E there exists a subsequence $t_n \to 0$ and $z \in E$ such that the corresponding w_n's converge weakly to z. Clearly, z is in the (weakly closed) cone $S_C(0) \equiv S$. Since $P_S y \in S$ and since S is the closure of $\cup t_n^{-1} C$, for each n we can choose $z_n \in t_n^{-1} C$ such that $\|P_S y - z_n\| \to 0$. Now, $w_n \in S$ and is the nearest point in $t_n^{-1} C$ to y, so

$$\|y - P_S y\| \leqq \|y - w_n\| \leqq \|y - z_n\| \leqq \|y - P_S y\| + \|P_S y - z_n\|$$

and hence $\|y - w_n\| \to \|y - P_S y\|$. Since $\|y - z\| \leqq \liminf \|y - w_n\| = \|y - P_S y\|$ and $z \in S$, we must have $z = P_S y$. Thus, $y - w_n \to y - P_S y$ weakly and therefore—by property (H)—in norm, which implies that $w_n \to P_S y$. By standard reasoning with subsequences, this argument shows that the original sequence $\{w_n\}$ converges to $P_S y$.

The Hilbert space version of Lemma 3 was applied by McCormick and Tapia [7] to prove a theorem about the gradient projection method for constrained optimization. The present version can easily be applied to obtain the same conclusion in somewhat more general spaces. We first describe the gradient projection method.

Suppose that f is a real valued continuously differentiable function on the Banach space E. Assume that both the inverse duality map J^{-1} and the metric projection P_C exist and are continuous. Following Golomb and Tapia [4] we define the *gradient* $\nabla f(x) \in E$ of f at $x \in E$ by the composition

$$\nabla f(x) = J^{-1}(f'(x)), \qquad x \in E$$

where $f'(x) \in E^*$ is the Fréchet derivative of f at x. The gradient projection method seeks to produce a minimizing sequence $\{x_n\}$ for f in C by the following inductive procedure:

Choose $x_1 \in C$. Having chosen $x_n \in C$, let

$$x_{n+1} = P_C[x_n - t_n \nabla f(x_n)] \tag{1}$$

where the steplength $t_n \geqq 0$ is chosen so as to minimize the function g_{x_n} defined by

$$g_{x_n}(t) = f(P_C[x_n - t\nabla f(x_n)]), \qquad t \geqq 0. \tag{2}$$

(We *assume* that such a minimum point exists.) Note that if we evaluate g_{x_n} at $t = t_n$ and $t = 0$, then

$$f(x_{n+1}) = g_{x_n}(t_n) \leqq g_{x_n}(0) = f(P_C[x_n]) = f(x_n)$$

so that the sequence $\{f(x_n)\}$ is nonincreasing. One cannot expect $\{x_n\}$ always to converge, but we would like to ascertain that any cluster point x^* of $\{x_n\}$ is, in some sense, a stationary point in C for f. It is too much to ask that $\nabla f(x^*) = 0$, but it is certainly appropriate to require that an application of the iteration (1) starting at x^* does not move to a different point, that is, $x^* = P_C[x^* - t\nabla f(x^*)]$, $t \geqq 0$. This is equivalent to the third assertion in the following proposition. The first assertion says that x^* is a solution to a variational inequality.

PROPOSITION 4. *For a point $x^* \in C$ each of the following four conditions implies the other three*:

(i) $\langle f'(x^*), x^* - y\rangle \leqq 0$ *for all y in C*,
(ii) $P_{S(x^*)}[-\nabla f(x^*)] = 0$,
(iii) $x^* = P_C(x^* - \nabla f(x^*))$,
(iv) $\langle f'(x^*), P_{S(x^*)}[-\nabla f(x^*)]\rangle \geqq 0$.

Proof. To see that (i) and (ii) are equivalent, we note that, by the defining inequality for $P_{S(x^*)}$, assertion (ii) is equivalent to

$$\langle J(-\nabla f(x^*)), u\rangle \leqq 0 \quad \text{for all } u \in S(x^*).$$

From the definitions of $\nabla f(x^*)$ and $S(x^*)$, this is equivalent to

$$\langle -f'(x^*), y - x^*\rangle \leqq 0 \quad \text{for all } y \in C,$$

which is the same as (i).

Similarly, if we restate (iii) in terms of the defining inequality for P_C, we obtain

$$\langle -f'(x^*), y - x^*\rangle = \langle J[x^* - \nabla f(x^*) - x^*], y - x^*\rangle \leqq 0 \quad \text{for all } y \in C,$$

which is again the same as (i).

Obviously, (ii) implies (iv). Suppose, then, that (iv) holds; we will deduce property (ii). Let $z = P_{S(x^*)}[-\nabla f(x^*)]$; knowing that $\langle f'(x^*), z\rangle \geqq 0$, we want to show that $z = 0$. For simplicity, let $w = -\nabla f(x^*)$. Since $0 \in S(x^*)$ and $z = P_{S(x^*)}w$, we have $\|w - z\| \leqq \|w\|$. By the definition of $\nabla f(x^*)$, $-f'(x^*) = Jw$ and hence $\|f'(x^*)\| = \|w\|$. Thus,

$$\|w\| \cdot \|w - z\| \geqq \langle -f'(x^*), w - z\rangle \geqq \langle -f'(x^*), w\rangle = \langle Jw, w\rangle = \|w\|^2.$$

It follows that either $w = 0$ (hence $z = 0$) or $\|w - z\| = \|w\|$. This latter equality says that 0 is a nearest point in $S(x^*)$ to w; by uniqueness of nearest points in rotund spaces, $z = 0$.

DEFINITION. We say that $x^* \in C$ is a *constrained stationary point* for f provided it satisfies any one of the properties (i)-(iv) above. In the fourth property we could have used equality in place of inequality, but the apparently weaker condition is useful in what follows. Note that if there are no constraints on f (that is, if $C = E$), then P_C is the identity map and the above conditions are equivalent to the usual definition: $\nabla f(x^*) = 0$.

Byrd and Tapia [1] have shown, in the *unconstrained* case, that cluster points of $\{x_n\}$ are stationary points for f, in arbitrary Banach spaces. (Actually, they utilize a more general steplength, which is discussed below.) The introduction of P_C forces adoption of the additional hypotheses in our final proposition.

PROPOSITION 5. *Suppose that E is a smooth, rotund and reflexive Banach space which satisfies property* (H). *If x^* is a cluster point of the sequence $\{x_n\}$ defined above, then x^* is a constrained stationary point for f.*

Proof. By using Lemma 3 and the chain rule, we can compute the right-hand derivative at $t = 0$ of

$$g_{x^*}(t) = f(P_C[x^* - t\nabla f(x^*)]).$$

Indeed, using the fact that $P_C x^* = x^*$, it is given by

$$\langle f'(x^*), P_{S(x^*)}[-\nabla f(x^*)]\rangle.$$

Thus, if x^* were not a constrained stationary point, this derivative would be negative and we could find $\tau > 0$ such that

$$f(P_C[x^* - \tau\nabla f(x^*)]) < f(P_C[x^*]) = f(x^*).$$

By continuity of ∇f and P_C and the fact that x is a cluster point of $\{x_n\}$, there would exist infinitely many n such that

$$f(P_C[x_n - \tau\nabla f(x_n)]) < f(x^*).$$

The minimum property used to define t_n would then imply that $f(x_{n+1}) < f(x^*)$ for

infinitely many n, contradicting the continuity of f and the fact that $\{f(x_n)\}$ is non-increasing.

The choice of steplength which was made in (1) is called *Cauchy's steplength.* It requires that the composition function g_{x_n} have a global minimum in $[0, \infty)$. A less stringent requirement is that t_n be the smallest stationary point of g_{x_n} in $[0, \infty)$; this is called *Curry's steplength.* For unconstained minimization, the result of Byrd and Tapia [1] referred to earlier is actually proved using this steplength. (See their paper for a thorough historical review of both methods.) In order to use Curry's steplength in the constrained case, one needs some sort of differentiability hypothesis on P_C at points *outside of C.* In general, this is a difficult question, even in Hilbert space. (See [6].) Some results in this direction are contained in [9].

REFERENCES

[1] R. H. BYRD AND R. A. TAPIA, *An extension of Curry's theorem to steepest descent in normed linear spaces*, Math. Programming, 9 (1975), pp. 247–254.

[2] M. M. DAY, *Normed Linear Spaces*, 3rd edition, Ergeb. der Math., Vol. 21, Springer-Verlag, Berlin-Heidelberg-New York, 1973.

[3] J. DIESTEL, *Geometry of Banach spaces—selected topics*, Lecture Notes in Mathematics 485, Springer-Verlag, Berlin-Heidelberg-New York, 1975.

[4] M. GOLOMB AND R. A. TAPIA, *The metric gradient in normed linear spaces*, Numer. Math., 20 (1972), pp. 115–124.

[5] A. HARAUX, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 20 (1977), pp. 615–631.

[6] J. B. HIRIART-URRUTY, *At what points is the projection mapping differentiable*? Amer. Math. Monthly, 89 (1982), pp. 456–460.

[7] G. P. MCCORMICK AND R. A. TAPIA, *The gradient projection method under mild differentiability conditions*, this Journal, 10 (1972), pp. 93–98.

[8] F. MIGNOT, *Contrôle dans les inéquations variationelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 130–185.

[9] R. R. PHELPS, *The gradient projection method using Curry's steplength*, this Journal, to appear.

[10] F. H. ZARANTONELLO, *Projections on convex sets in Hilbert space and spectral theory*, in Contributions to Nonlinear Functional Analysis, Publn. No. 27, Math. Res. Center, Univ. Misc., Academic Press, New York-London, 1971, pp. 237–424.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 23, No. 6, November 1985

007

ERRATA: ON STABILIZABILITY OF LINEAR SPECTRAL SYSTEMS VIA STATE BOUNDARY FEEDBACK*

RUTH F. CURTAIN†

The following corrections should be made to this article (with thanks to B. M. N. Clarke).

1. pp. 148ff. The state space should be $C \oplus Z$ instead of $R \oplus Z$.

2. p. 149. The first paragraph should read:

The biorthogonal system for $\tilde{\phi}_k$ is given by

$$\tilde{\psi}_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \tilde{\psi}_k = \begin{pmatrix} x_k \\ \psi_k \end{pmatrix}, \quad \text{where } x_k = -\langle \psi_k, h \rangle$$

and so

$$\tilde{b}_k = \bar{x}_k - b_k, \tag{3.12}$$

where the inner product is in $C \oplus Z$. We remark that if $q = 0$, then $h = 0$, $\tilde{b}_0 = 1$ and $\tilde{b}_k = -b_k$. We now state our main result.

* This Journal, 23 (1985), pp. 144–152.

† Rijksuniversiteit Groningen Mathematisch Instituut, Postbus 800, 9700 AV Groningen, the Netherlands.